SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES
http://saucis.sakarya.edu.tr/
e-ISSN: 2636-8129
Publisher: Sakarya University
Vol. 8, No. 1, 27-37, 2025
DOI: 10.35377/saucis...1525350

RESEARCH ARTICLE

Face Super Resolution Based on Identity Preserving V-Network

Ali Hüsameddin Ateş[1,2], Hüseyin Eski[1]

[1] Sakarya University, Department of Computer Engineering, Sakarya, Türkiye, ror.org/04ttnw109
[2] Sakarya University, Institute of Natural Sciences, Sakarya, Türkiye, ror.org/04ttnw109

Corresponding author:
Ali Hüsameddin Ateş,
Department of Computer Engineering, Sakarya University
aliates@sakarya.edu.tr

Article History:
Received: 31.07.2024
Revised: 27.02.2025
Accepted: 20.03.2025
Published Online: 27.03.2025

ABSTRACT
Numerous super-resolution methods have been developed to restore and upsample low-resolution and low-detail images to higher resolutions. Specifically, face super-resolution studies aim to restore various degradations in facial images while enhancing their resolution and preserving details. This study proposes the VNet architecture, which consists of a deep learning-based convolutional network for converting low-resolution and degraded facial images into high-quality and detailed images, and a pre-trained FaceNet model to preserve identity (biometric) information. The architecture leverages the advantages of the Encoder-Decoder structure bidirectionally to maintain details and recover lost information. In the initial stage, the Encoder module compresses the image representation, filtering out unnecessary information. The Decoder module then reconstructs the high-resolution and restored image from the compressed representation. The use of residual connections in this process helps minimize information loss while preserving details. The final stage utilizes the identity loss feedback from the FaceNet model to enhance the image without deviating from the original identity context. Tests conducted on various facial datasets demonstrate that VNet achieves high metric performance in both super-resolution and restoration tasks. The results indicate that the proposed architecture is effective in producing realistic and high-quality versions of low-resolution and degraded facial images.

Keywords: Face super resolution, Face restoration, Super resolution, Deep learning

1. Introduction

The transformation of low-resolution and low-detail images into high-resolution, detailed, and sharp images is studied under the title of super-resolution (SR) in the field of digital image processing. Various algorithms and methods are used to achieve super-resolution. Super-resolution is generally utilized in many areas such as improving image quality in photographs or videos, enhancing the resolution of medical imaging devices like magnetic resonance imaging (MRI) and computed tomography (CT), and making satellite images more detailed and clearer. The main goal in super-resolution is to improve visual quality and increase the resolution of the image while preserving details. On the other hand, super-resolution can be used as an auxiliary tool in tasks such as object detection and segmentation. Additionally, when looking specifically at facial images, super-resolution techniques are employed to enhance the effectiveness of security cameras, especially under challenging conditions such as low light, noise, and blur, by increasing the resolution of low-resolution images to facilitate face detection [1].

While there are many interpolation-based and similar super-resolution techniques in image processing, studies have shown that deep learning-based SR techniques perform higher than traditional methods. Especially GAN and CNN-based approaches can successfully enhance details in low-resolution images, recovering fine details lost in low-resolution facial images. Therefore, deep learning-based SR techniques are more widely used compared to classical methods. However, training and applying deep learning models require high computational power and large datasets. Additionally, the performance of SR algorithms is directly related to the quality and diversity of the datasets used. Therefore, creating large and diverse datasets containing facial images captured under different conditions and training SR techniques on these datasets is crucial for developing more generalizable and high-performance SR models [2].

Super-resolution techniques used for image enhancement gained a new deep learning-based perspective with Dong et al. [3] using convolutional neural networks in super-resolution. SRCNN proposed a three-layer structure. The first layer extracts patches from the low-resolution image and maps each patch to a high-dimensional space, the second layer performs a non-linear mapping of these high-dimensional representations from low-resolution patch space to high-resolution patch space,

Cite as: A. H. Ateş and H. Eski, "Face super resolution based on identity preserving V-Network," *Sakarya University Journal of Computer and Information Sciences,* vol. 8, no. 1, pp. 27-37, 2025, doi: 10.35377/saucis...1525350

and the third layer combines high-resolution patches to produce the high-resolution image. To reduce the computational complexity of SRCNN, Dong et al. also proposed FSRCNN, which uses deconvolution for upsampling.

Kim et al. proposed the Very Deep Super Resolution – VDSR [4] network by further deepening the networks suggested by Dong. The proposed network is based on VGGNet used for ImageNet classification and has 20 weight layers. Feature extraction was performed using 64 filters (3x3) in each layer. Additionally, a residual learning connection was established between the first and last layer to improve the learning rate, which decreases as the network deepens.

To address the vanishing gradient problem, which occurs due to the deepening of networks leading to the activation functions approaching zero and the network being untrainable, He et al. proposed Residual Networks (ResNet) [5]. Ledig et al. adapted it to super-resolution as SRResNet (Super Resolution Residual Network), achieving superior performance compared to previous techniques due to feedback connections that feed the deepening network [6].

Lim et al. removed the normalization layers in the SRResNet network to reduce unnecessary memory and computation costs during training, increased the model's capacity using a deeper structure, and achieved higher PSNR and SSIM with more residual blocks [7].

With the proposal of Dense Blocks (DenseNet) [8] by Huang et al., Tong et al. proposed the SRDenseNet network to enhance information flow, backpropagate gradients more effectively, and leverage the advantages of residual learning, achieving high clarity while preserving details and edges [9].

Upon the proposal of Generative Adversarial Networks (GAN) by Goodfellow et al. [10], Lim et al. proposed the SRGAN (Super-Resolution using a Generative Adversarial Network) network along with SRResNet, achieving higher accuracy outputs by using generative networks as an alternative to classical linear networks [6].

Wang et al. built on the SRGAN network to propose ESRGAN, removing the normalization layer in the residual block and using Residual-in-Residual Dense Blocks, achieving higher visual quality with the Relativistic Average GAN, which learns whether one image is more realistic compared to another [11].

General super-resolution networks showed low performance due to the high structural complexity of facial images, leading to specialized studies for facial images. Zhou et al. proposed a bi-channel convolutional neural network (BCCNN) for face super-resolution (FSR). To obtain super-resolution output, one channel extracts features related to face regions from the input image, while the other appropriately combines the extracted features with the low-resolution input image, achieving higher success on facial images with Gaussian and motion blur compared to classical CNN architectures [12].

Yu et al. proposed the UR-DGN (Ultra-Resolution by Discriminative Generative Networks) network, the first GAN-based network for face image restoration. Here, the discriminator network learns significant components of the human face, while the generator network combines these components with the input image. A pixel-based ℓ_2 loss term was used in the generator model, and the discriminator network's feedback was used to make the upsampled face images more similar to the real ones, enabling upsampling of very small inputs up to eight times [13].

In classical convolutional networks, each channel is treated equally during feature extraction, despite some channels carrying more features than others. Zhang et al. achieved more efficient results in general super-resolution by using channel-focused feature extraction (Channel Attention) that assigns different importance levels (weights) to each channel [14].

Zhao et al. proposed the SAAN (Semantic Attention Adaptation Network for Face Super-Resolution) network, using a semantic channel attention feature extraction method to protect important regions such as eyes, nose, and mouth in face images and achieve more realistic and detailed restoration. This channel-focused mechanism assigns different weights to different regions of the input image, emphasizing important facial features, achieving more efficient results in face image super-resolution [15].

Lu et al. proposed the FSAAN (Face Hallucination via Split-Attention in Split-Attention Network) network, using split-attention feature extraction by splitting the input channels, focusing on both facial texture details and structural information, achieving structurally richer faces [16].

With the adaptation of the Transformer architecture used in natural language processing to image processing as Vision Transformers (ViT) by Dosovitskiy et al. [17], an alternative to networks that process images using convolution in the classical CNN architecture emerged. Subsequently, many super-resolution networks using Transformer architectures were proposed.

Wang et al. leveraged the strengths of both CNNs and Transformer structures to enhance face image reconstruction, proposing the TANet (A new Paradigm for Global Face Super-resolution via Transformer-CNN Aggregation Network) architecture based on CNN and Vision Transformer (ViT). The CNN part is used for restoring fine details and local features of the face, while the Transformer part is used to maintain the consistency of the overall face structure with the Global Attention mechanism, achieving higher fidelity and naturalness in reconstructed face images [18].

Gao et al. proposed the CTCNet (A CNN-Transformer Cooperation Network for Face Image Super-Resolution) architecture, a hybrid CNN-Transformer network. CTCNet, with a U-Net-like structure, uses a combination of Facial Structure Attention

Unit and Transformer blocks on the encoder side to capture both local facial textures and general facial structural information. In the intermediate transition, the Feature Refinement Module focuses on the essential facial structure among extracted features, and the Multi-scale Feature Fusion Units module on the decoder side combines and up-samples the extracted local and global features, resulting in high-resolution images. Thanks to its hierarchical U-Net-like architecture with interconnections and advanced modular structure, CTCNet achieved higher results compared to many architectures in various benchmarks [19].

In studies on restoring or upsampling face images with super-resolution, the quality of the obtained image, high metric performance, and visual enhancements are often prioritized, while identity fidelity, the extent to which the output image resembles the original in terms of identity, is considered secondary. In this study, a model is proposed that harmonizes high metric performance and visual enhancements without compromising the inherent similarity of the generated output image to the original. To achieve this, in this study, VNet architecture is proposed that uses encoder-decoder based convolutional neural networks to obtain high resolution restored images from low resolution corrupted images. In the architecture, the image is restored and upsampled with V-shaped blocks similar to U-Net, while the biometric verification of the person or fidelity is performed with the FaceNet model used, preserving personal characteristics.

2. Material and Methods

2.1. VNet Architecture

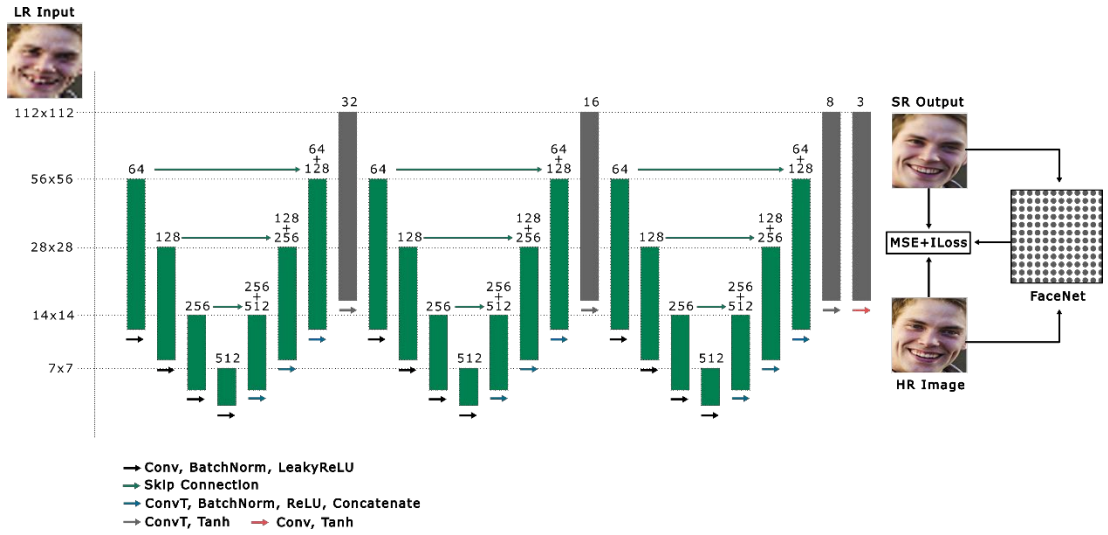

Figure 1. Proposed VNet Architecture

In Figure 1, the resolutions on the left indicate the output resolution of each convolution layer, the green and gray bars indicate the convolution layers, the numbers above the bars indicate the number of convolution filters, the jump links between the layers indicated by the green arrows indicate the merging process between the layers, the green bars indicate the encoder-decoder layers, the gray bars indicate the intermediate transition layers between the decoder and encoder, and the colored arrows indicate the operations performed in the layers.

Convolutional Neural Networks from deep learning architectures are used for the super-resolution of face images in the VNet network. The network, designed with an encoder-decoder architecture similar to U-Net, connects layers with residual learning shortcuts proposed in the ResNet architecture. Residual Learning aims to address the vanishing gradient problem by preventing the gradient function from converging to zero as the network deepens and layers increase.

In the VNet network, feature extraction from the image is performed through V-shaped green layers, and the extracted features are combined while maintaining the input resolution of the image through gray intermediate layers. The network, which has a total of 33 million trainable parameters, processes input images through a data preprocessing step. The network operates by feeding pairs of original (HR) and distorted (LR) images, where the original image is subjected to a distortion function to obtain an LR image. The distortion function reduces a 112x112 pixel image to a size of 28x28 pixels, then enlarges it back to its original size to produce the LR image.

No prior-based approach is used in architecture. The network takes a holistic approach to input images, learning the distortion function between HR-LR image pairs through various layers and connections, and attempts to predict the original image from the degraded image by learning the inverse of this function.

The overall structure of the network is shown in Figure 1, where the output resolution of the image is indicated after each layer in the left chart. The image resolution, which continuously halves after each layer, is restored to the input size with intermediate convolutions having 32, 16, and 8 filters, and finally, a super-resolved RGB image (SR output) is obtained with a 3-filter convolution.

The VNet network has a fixed input size, receiving inputs of 112x112 pixels. The input image progresses through the green blocks specified by the number of filters from left to right. The V-shaped green blocks consist of 4 encoder layers shown with black arrows, 3 decoder layers shown with blue arrows, and finally, 1 transposed convolution layer.

The input layer consists of 64 convolution filters, and in subsequent layers, the number of filters doubles in the encoder layers, reaching up to 512 filters as the network deepens. Then, in the decoder layers, the number of filters is halved, making the network shallower. All convolution filters used in the network are 4x4 in size.

In the encoder layers, convolution operations use a stride of 2 and same padding. This ensures that the image resolution is halved smoothly after each layer. Filter weights are initialized with a random starter value between 0 and 0.02. Batch normalization is used for optimization after convolution, and the activation function used is LeakyReLU.

In the decoder layers, the transposed convolutions also use a stride of 2, same padding, and random starter values between 0 and 0.02. Batch normalization is used for optimization, and the activation function is ReLU. The transposed convolution and same padding used ensure that the halved image resolution is smoothly doubled from layer to layer. The 512 filters used in the deepest part of the network are halved through the decoder layers; each layer's filter weights are combined with the filter weights (embeddings) of the corresponding layer with the same resolution as the image resolution in that layer. The merging process (green arrows in Figure 1) combines encoder and decoder features, refining feature maps and retaining important details from the original image.

In the gray layers with 32-16-8 filters following the decoder layers, the image resolution is restored to the input resolution through transposed convolutions, while feature maps are transferred between V-shaped blocks. The transposed convolution process uses a stride of 2, same padding, and a Tanh activation function.

In the final layer, the feature maps obtained from restoring and super-resolving the low-resolution and detailed input image through the network are combined in RGB channels with a 4x4 filter convolution and Tanh activation function, resulting in a high-resolution and detailed image.

As the loss function of the network, both Mean Squared Error (MSE) and Identity Loss (ILoss) functions are used together, as suggested by Khazaie et al. [20]. To obtain the Identity Loss, the super-resolved image produced by the network is paired with the original image and fed into a face recognition network, FaceNet [21]. The pre-trained FaceNet model [22], which was trained using the VGGFace2 dataset, is a successful network for obtaining filtered data (embeddings) related to human faces. The MSE function aims to numerically approach the original image by examining the pixel-based numerical difference between the original and predicted images. On the other hand, Identity Loss evaluates how similar the predicted image is to the original image of the person, providing feedback through similarity errors. This approach ensures that the output image not only has high visual quality and detail but also maintains similarity to the original person.

2.2. Dataset

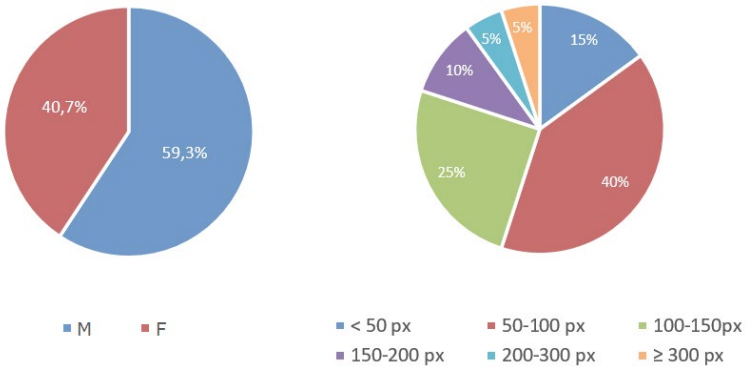

Figure 2. VGGFace2 Dataset Distribution [23]

The VGGFace2 dataset announced by Cao et al. [23] was used for training the proposed VNet network. This dataset contains 3.31 million face images of 9,131 individuals, obtained from Google Images, covering a wide range of poses, lighting conditions, ethnic backgrounds, races, ages, and professions (actors, athletes, politicians, etc.). With an average of 362 images per person, the dataset also includes facial bounding boxes, and the images vary in resolution and are loosely cropped.

In the data pre-preparation phase, for computational and time gain purposes, approximately 100 images were randomly selected from the images of each person from the dataset (9131 people in total), making a total of 863,732 images. The face

images in the selected images were tightly cropped with a face detection tool, and the images with different resolutions were resized to 112x112 pixels.

2.3. Metrics

Several evaluation metrics are used to measure the similarity between images in terms of both quantity and quality [24]. These metrics, commonly used in the field of super-resolution, provide numerical information on how well the distorted image has been restored compared to the original image by establishing a similarity relationship using different methods and approaches between the two versions of the image. In this study, MAE (Mean Absolute Error), MSE (Mean Squared Error), PSNR (Peak Signal-To-Noise Ratio) metrics are used as pixel-based metrics and SSIM (Structural Similarity Index Measure) and LPIPS (Learned Perceptual Image Patch Similarity) metrics are used as perceptual metrics.

MAE and MSE focus on the numerical difference between the pixels of two images. As the images become more similar and the numerical difference between them decreases, these error values decrease. PSNR measures the error rate logarithmically with respect to the maximum pixel value between the two images, again using MSE. A high PSNR value indicates a high similarity between the two images. These pixel-based metrics focus on pixel-wise differences and do not provide information about perceptual similarity. These metrics range from a minimum value of 0 (perfect agreement) to a maximum value of ∞ (maximum error), with similarity approaching a maximum as the value approaches zero.

The SSIM structural similarity index measures perceptual similarity by comparing the brightness, contrast, and structural information between two images. The SSIM value ranges from 0 to 1. The SSIM value approaches 1 as the similarity between two images increases and 0 as the similarity decreases. LPIPS is a perceptual metric that measures image similarity by comparing features extracted from deep learning models and provides results closer to human visual perception. Although there is no exact range of values for the LPIPS metric, in practice, values close to 0 indicate a high similarity between two images, while values close to 1 or higher indicate that the similarity of these images is very low [25].

Table 1. VNet Model Hyperparameters

Hyperparameter	**Value**
Input Size	112x112
Input/Output Channels	3
Convolution Filter Size	4x4
Convolution Strides	2
Convolution Padding	Same
Use Bias	False
Apply Dropout	False
Activation Functions	Encoder Layers: LeakyReLU Decoder Layers: ReLU Gray Layers: Tanh
Loss Function	MSE + Identity Loss
Optimizer Function	Adam

3. Experiments

For training, the hardware used includes an Intel Core i9-7960X CPU and an Nvidia RTX 4090 24GB GPU, while the software includes Python 3.11.9 and TensorFlow 2.15.0. The VNet network was trained for 49 epochs using 863,732 images selected from the VGGFace2 dataset with the parameters shown in the table below (as seen in Table 1).

The number of images in the dataset, input size and epochs are kept minimum considering the hardware constraints, and more sample, input size and epochs will give better results. The error function of the network is calculated by the sum of MSE and Identity Loss errors and both pixel and perceptual based error control are performed. More efficient results can be obtained by using different weights while summing these two error functions. Since RGB images have 3 channels, the number of input and output channels must be 3. 4×4 filters keep the number of parameters and computational cost low compared to larger filters such as 5×5 or 7×7, while providing a wider detection area compared to 3×3 filters. Striding 2 enables learning hierarchical features by changing the resolution step by step in the encoder-decoder structure by downsampling and upsampling, halving the image size in each convolution and doubling it in deconvolution. In addition, equal padding is used

to prevent the change in spatial size. Since Batch normalization is used as the normalization layer, no additional Bias usage was required. Dropout wasn't used to ensure that the data set was sufficiently diverse and to prevent unwanted visual artifacts. In the encoder layers, LeakyReLU prevents the vanishing gradient problem by keeping the negative region active with a small slope, while in the decoder layers ReLU encourages positive outputs and provides smooth outputs in image production. Using Tanh in gray layers allows the pixel values of the images to be pulled into a normalized range (-1,1). Thanks to the adaptive learning rates, momentum utilization and fast convergence features of the Adam optimization function, it enables the complex structure of the model and detailed parameter updates to be managed effectively and the fine details in the image to be learned more reliably.

The outputs of the Mean Squared Error (MSE), Identity Loss, and total loss functions used during training are shown in the graphs for the training and validation phases (as seen in Figure 3). The closely aligned training and validation loss outputs indicate the stability of the training process and the network.

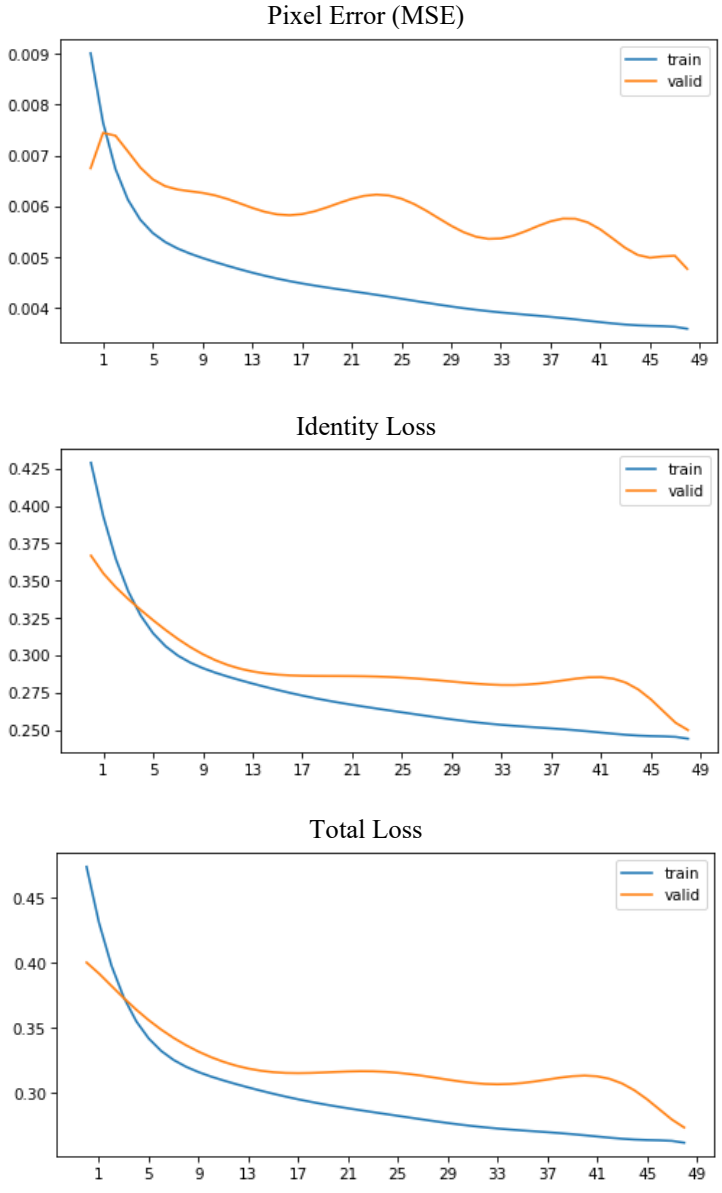

Figure 3. Train and Validation Losses

3.1. Comparative Results

After the training process, various evaluation metrics were used to quantitatively measure the difference between the original and super-resolved images during the testing phase, in order to assess the model's performance on the test data. The metrics used include MAE (Mean Absolute Error), MSE (Mean Squared Error), PSNR (Peak Signal-To-Noise Ratio), SSIM (Structural Similarity Index Measure), and LPIPS (Learned Perceptual Image Patch Similarity) [24].

For the test datasets, VGGFace2 [23], CelebA [26], UTKFace [27], FEIFace [28], and MultiPie [29] datasets were used, each consisting of face images obtained through different methods and purposes, with 2,000 images randomly selected from each. The evaluation metrics for the images obtained by enlarging 28x28 pixel resolution images by a factor of 4 are shown in Table 2.

VGGFace2 provides a large number of samples and people, a wide age range, demographic diversity and pose richness, while CelebA stands out in learning facial attributes by providing 40 different feature labels for each face. UTKFace has a relatively small dataset with images rich in age, gender and ethnicity and focuses on age in general, while FEIFace is useful in making sense of facial components thanks to its small scale but high-quality images. Multi-PIE, on the other hand, can prevent performance degradation against illumination and pose changes thanks to images of individuals taken from different angles, in various poses and lighting conditions.

Table 2. Test Datasets ×4 Test Scores

Test Dataset	MAE↓	MSE↓	PSNR↑	SSIM↑	LPIPS↓
VGGFace2	.02812	.00169	28.730	0.8806	0.0937
CelebA	.02757	.00172	28.271	0.8891	0.0835
UTKFace	**.02650**	**.00163**	**28.982**	**0.8999**	**0.0626**
FEIFace	.02717	.00165	28.769	0.8980	0.0747
MultiPie	.03984	.00362	25.884	0.8800	0.1131

When we look at the results obtained, the UTKFace dataset has the highest metric performance, while the other test datasets have similar values. The UTKFace dataset shows the highest metric performance due to its high-quality images that are standardized and well-aligned in terms of angle and exposure, while the MultiPie dataset shows lower metric performance due to its richness in different angles, lighting and exposure. On the other hand, the low LPIPS score (0.0626) obtained on the UTKFace dataset reflects superior perceptual similarity, consistent with high PSNR and SSIM values. The fact that the results obtained on the test datasets are close to each other is important in obtaining a balanced and generalizable model.

The images from the VGGFace2, CelebA, UTKFace, MultiPie, and FEIFace datasets, including their originals (HR), distorted versions reduced to 28x28 pixel resolution (LR), and the outputs obtained by upscaling the distorted images by a factor of 4 (SR), are shown sequentially in Figure 4.

When looking at the visual test outputs, it is seen that VNet shows a sampling performance suitable for the original image in the tests based on increasing the image size to 4 times super resolution, in the restoration of eyes, mouth, nose and other facial components that are not present in the distorted image; in the repair and upsampling of blurry, noisy and distorted textural regions. While doing this, as seen in detail in Figure 5, the low-resolution image passes through the filters in the encoder and decoder layers of the model. Each filter reconsiders the image with its learned weights. While the initial layers focus on low-level features such as edges and corners, emphasizing the contrasting regions of the image, simple lines and prominent edges; when the image reaches the mid-level layers, textures and facial features are addressed. When the network reaches the high-level layers, the output of the filters becomes more abstract as the network gets deeper. The filter outputs (as seen in Figure 5) may appear blurry and complex while being simple at the beginning; because the network has now learned high-level features that can distinguish and restore the entire face or certain key regions. This deep learning allows the model to create a rich feature representation. As a result, the model updates its weights to reproduce the face based on this information. Each filter capturing a feature of a specific part of the image, and the image given as input to the model is processed one by one by these filters to restore the image.

4. Conclusion

In this study, VNet architecture is proposed, which aims to protect biometric identity information in the restoration of facial images with super resolution. The identity preservation capability of the model is supported by the FaceNet-based identity loss function. In this way, critical facial components such as eyes, nose and mouth in images generated from low-resolution inputs are restored consistent with the biometric features of the original person. The model trained with the VGGFace2 dataset was tested on both VGGFace2 images not used in training and various datasets such as CelebA, UTKFace, FEIFace and MultiPie in 4-fold super-resolution tests. The obtained MAE, MSE, PSNR, SSIM and LPIPS metric scores revealed that the model exhibited stable and generalizable performance on different datasets. The reason for the relatively low SSIM value is that this metric considers brightness and contrast similarity as well as structural similarity when comparing images. Although the VNet model provides sufficient improvements structurally, it needs to be trained and optimized further in terms of brightness and contrast.

The UTKFace dataset achieved the highest metric score thanks to its standard and well-aligned high-quality images in terms of angular and exposure. On the other hand, relatively low metric performances were obtained in the tests conducted on the MultiPie dataset. The reason for this is that the images in the other test dataset were mostly taken from the front, while the MultiPie dataset consists of face images taken from many angles and with different exposures. In order to prevent such handicaps, the datasets used in training should be diversified, rich in terms of angular and exposure, and the number of samples should be as high as possible. Again, the integration of advanced loss functions to improve brightness and contrast restoration, attention mechanisms to improve global and local feature extraction, and the Transformer architecture are among the important improvements to present a more advanced and efficient architecture.

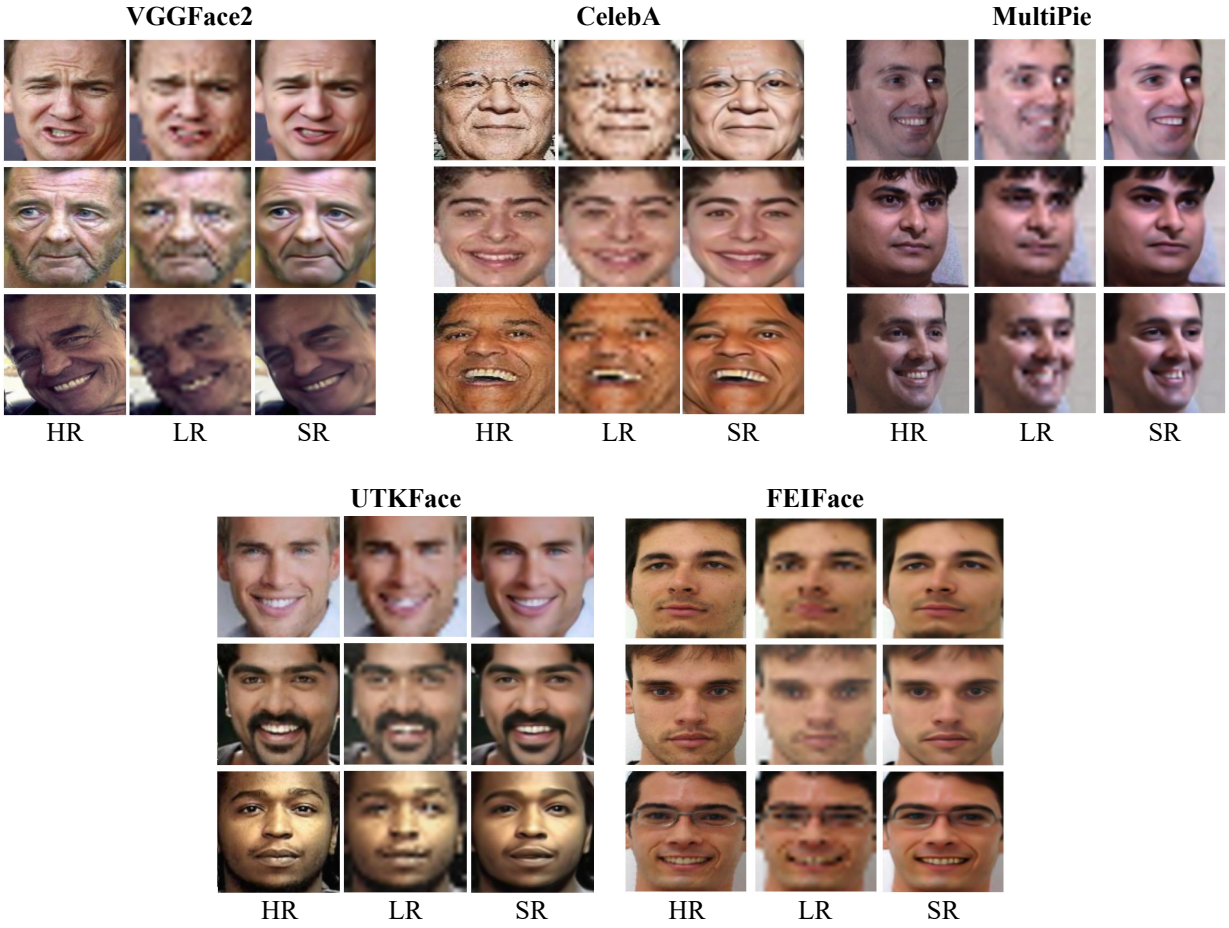

Figure 4. Test Datasets ×4 Outputs

This study offers a methodology that can be referenced especially in areas where identity verification is critical, such as security systems and forensic analysis. However, more advanced architectures are needed for images under difficult conditions. Expanding the capacity of the model with increased data diversity and hybrid architecture is of critical importance for the integration of the model to real-world conditions.

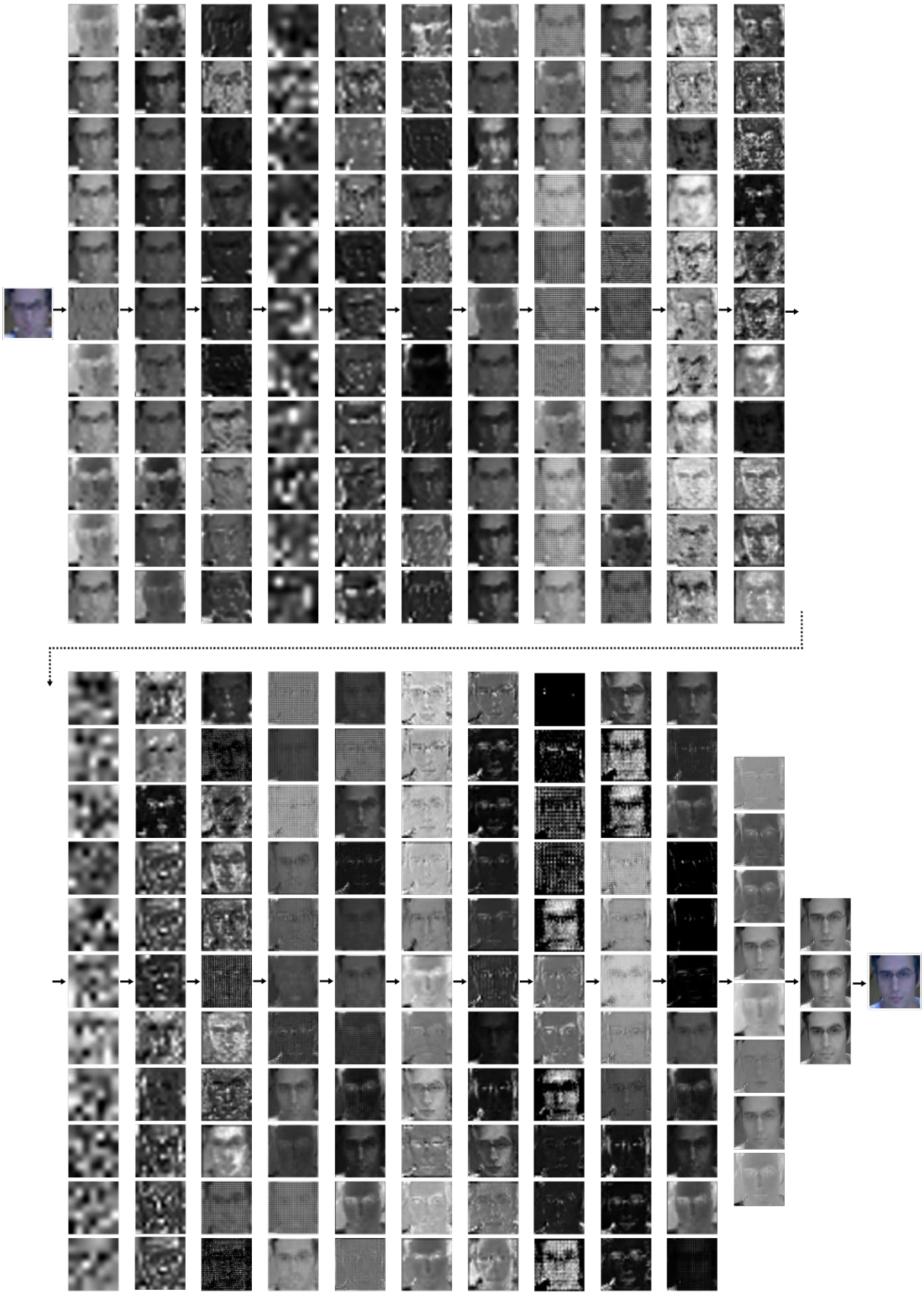

Figure 5. Sample Outputs of Model Layers (Layer Dimensions Shortened)

References

[1] N. Singh, S. S. Rathore, and S. Kumar, "Towards a super-resolution based approach for improved face recognition in low resolution environment," *Multimed Tools Appl*, vol. 81, no. 27, pp. 38887–38919, Nov. 2022, doi: 10.1007/S11042-022-13160-Z/FIGURES/16.

[2] J. Jiang, C. Wang, X. Liu, and J. Ma, "Deep Learning-based Face Super-Resolution: A Survey," *ACM Comput Surv*, vol. 55, no. 1, Jan. 2021, doi: 10.1145/3485132.

[3] C. Dong, C. C. Loy, K. He, and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," *IEEE Trans Pattern Anal Mach Intell*, vol. 38, no. 2, pp. 295–307, Dec. 2014, doi: 10.1109/TPAMI.2015.2439281.

[4] J. Kim, J. K. Lee, and K. M. Lee, "Accurate Image Super-Resolution Using Very Deep Convolutional Networks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 1646–1654, Nov. 2015, doi: 10.1109/CVPR.2016.182.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, Dec. 2015, doi: 10.1109/CVPR.2016.90.

[6] C. Ledig *et al.*, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 105–114, Sep. 2016, doi: 10.1109/CVPR.2017.19.

[7] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2017-July, pp. 1132–1140, Jul. 2017, doi: 10.1109/CVPRW.2017.151.

[8] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 2261–2269, Aug. 2016, doi: 10.1109/CVPR.2017.243.

[9] T. Tong, G. Li, X. Liu, and Q. Gao, "Image Super-Resolution Using Dense Skip Connections".

[10] I. J. Goodfellow *et al.*, "Generative Adversarial Nets", Accessed: May 07, 2024. [Online]. Available: http://www.github.com/goodfeli/adversarial

[11] X. Wang *et al.*, "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11133 LNCS, pp. 63–79, Sep. 2018, doi: 10.1007/978-3-030-11021-5_5.

[12] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Learning face hallucination in the wild," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, in AAAI'15. AAAI Press, 2015, pp. 3871–3877.

[13] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9909 LNCS, pp. 318–333, 2016, doi: 10.1007/978-3-319-46454-1_20/TABLES/1.

[14] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image Super-Resolution Using Very Deep Residual Channel Attention Networks," 2018.

[15] T. Zhao and C. Zhang, "SAAN: Semantic Attention Adaptation Network for Face Super-Resolution," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020, pp. 1–6. doi: 10.1109/ICME46284.2020.9102926.

[16] T. Lu *et al.*, "Face Hallucination via Split-Attention in Split-Attention Network," in *Proceedings of the 29th ACM International Conference on Multimedia*, in MM '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 5501–5509. doi: 10.1145/3474085.3475682.

[17] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR 2021 - 9th International Conference on Learning Representations*, Oct. 2020, Accessed: Jul. 10, 2024. [Online]. Available: https://arxiv.org/abs/2010.11929v2

[18] Y. Wang *et al.*, "TANet: A new Paradigm for Global Face Super-resolution via Transformer-CNN Aggregation Network," Sep. 2021, Accessed: Jul. 10, 2024. [Online]. Available: https://arxiv.org/abs/2109.08174v1

[19] G. Gao, Z. Xu, J. Li, J. Yang, T. Zeng, and G.-J. Qi, "CTCNet: A CNN-Transformer Cooperation Network for Face Image Super-Resolution," *IEEE Transactions on Image Processing*, vol. 32, pp. 1978–1991, Apr. 2022, doi: 10.1109/TIP.2023.3261747.

[20] V. R. Khazaie, N. Bayat, and Y. Mohsenzadeh, "Multi Scale Identity-Preserving Image-to-Image Translation Network for Low-Resolution Face Recognition," *Proceedings of the Canadian Conference on Artificial Intelligence*, Oct. 2020, doi: 10.21428/594757db.66367c17.

[21] "davidsandberg/facenet: Face recognition using Tensorflow." Accessed: Jul. 15, 2024. [Online]. Available: https://github.com/davidsandberg/facenet?tab=MIT-1-ov-file#readme

[22] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, pp. 815–823, Mar. 2015, doi: 10.1109/cvpr.2015.7298682.

[23] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *International Conference on Automatic Face and Gesture Recognition*, 2018.

[24] T. Wang *et al.*, "A Survey of Deep Face Restoration: Denoise, Super-Resolution, Deblur, Artifact Removal," Nov.

2022, Accessed: May 08, 2024. [Online]. Available: https://arxiv.org/abs/2211.02831v1
[25] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 586–595, Jan. 2018, doi: 10.1109/CVPR.2018.00068.
[26] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," *CoRR*, vol. abs/1411.7766, 2014, [Online]. Available: http://arxiv.org/abs/1411.7766
[27] S. Y. Zhang Zhifei and H. Qi, "Age Progression/Regression by Conditional Adversarial Autoencoder," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
[28] C. E. Thomaz and G. A. Giraldi, "A new ranking method for principal components analysis and its application to face image analysis," *Image Vis Comput*, vol. 28, no. 6, pp. 902–913, Jun. 2010, doi: 10.1016/J.IMAVIS.2009.11.005.
[29] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis Comput*, vol. 28, no. 5, pp. 807–813, May 2010, doi: 10.1016/J.IMAVIS.2009.08.002.

Conflict of Interest Notice
Authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical Approval and Informed Consent
It is declared that during the preparation process of this study, scientific and ethical principles were followed, and all the studies benefited from are stated in the bibliography.

Availability of data and material
Not applicable / or link

Plagiarism Statement
This article has been scanned by iThenticate ™.