

RESEARCH ARTICLE

Enhancing Face Image Quality: Strategic Patch Selection With Deep Reinforcement Learning and Super-Resolution Boost via RRDB

EMRE ALTINKAYA^{1,2} AND BURHAN BARAKLI³, (Member, IEEE)

¹Department of Hybrid and Electric Vehicle Technology, Vocational School, Bilecik Şeyh Edebali University, 11100 Bilecik, Türkiye

²Department of Electrical-Electronics Engineering, Institute of Natural Sciences, Sakarya University, Serdivan, 54050 Sakarya, Türkiye

³Department of Electrical-Electronics Engineering, Faculty of Engineering, Sakarya University, Serdivan, 54050 Sakarya, Türkiye

Corresponding author: Burhan Baraklı (barakli@sakarya.edu.tr)

This work was supported by the Computing Resources funded by the National Center for High Performance Computing of Turkey (UHeM) under Grant 1011402021.

ABSTRACT Facial super-resolution (FSR) is a critical research area whose goal is to improve visual quality by converting low-resolution facial images to high resolution ones. Research in FSR has come a long way thanks to advances in deep learning technologies. However, there is still a need to develop effective methods for revealing facial details and preserving the overall appearance. For this purpose, a new approach called Deep Reinforcement Learning Based Super Resolution of Face Regions (DRL-SRFR) is proposed. It is based on Deep Reinforcement Learning (DRL) and Deep Residual Dense Block (RRDB) architectures. In the DRL part of the method, new regions that need attention are identified at each step using the repeated visual attention methodology. The details in different parts of the face image are iteratively improved to produce more natural and high-quality face images. In addition, with the stochastic action-taking process, the decision-making process is made flexible by focusing on important facial regions. The focused region is improved with the RRDB structure using dense connections and residual learning. Experiments and ablation studies show that the developed model provides a significant advantage over existing methods in improving local details and preserving appearance integrity.

INDEX TERMS Facial super-resolution, deep reinforcement learning, super resolution of face regions, deep residual dense block.

I. INTRODUCTION

Facial super-resolution (FSR) methodology has made remarkable progress in improving low-quality facial images in recent years. It has also opened up new possibilities in the field of visual data processing and analysis, forming an important research area along with other super-resolution (SR) tasks. The main goal of this methodology is to improve the quality and usability of visual data analysis by enhancing low-resolution facial images from various real-world scenarios to get high-resolution, detailed, and clear images. Facial hallucination (FH), also known as FSR, aims to improve human perception as well as computer-based face recognition

The associate editor coordinating the review of this manuscript and approving it for publication was Janmenjoy Nayak¹.

and analysis systems. It enhances the accuracy of face recognition systems, especially in the field of security, by making facial features more prominent in images taken remotely or in poor lighting conditions. Its importance has been increasing in biometrics, surveillance, digital entertainment, and gaming industries. FSR-derived facial images are also effective in converting eyewitness sketches into photographs to facilitate the identification of criminals. It also plays an important role in various face-related subtasks, such as face alignment, face segmentation, and face detection [1], [2], [3], [4].

Baker and Kanade first defined the FSR problem and presented a Bayesian-based statistical method for extracting high-frequency components in images [5]. FSR algorithms can be classified into four main classes: interpolation, statistical, reconstruction, and learning-based methods [6].

Traditional interpolation methods estimate values between available pixels. The quality of interpolation-based methods is low because estimation processes performed in spatial space lead to the loss and blurring of high-frequency details [5]. Methods based on modeling statistical relationships between high-resolution and low-resolution images are referred to as statistically based methods, and leading examples are given in [7], [8], [9], [10], [11], and [12]. In some of these methods, the variance principal component is used in different spatial transformations, while in others it is used to extract local and global relationships between neighborhoods or to model the statistical distribution of pixels. Statistical inference is used to enhance the emphasis of high-frequency components while preserving facial contour and the overall appearance of the image. Due to fixed parameters or limited modeling capabilities, statistical methods are not flexible enough particularly in cases where facial structures and variations are complex, there are difficulties in capturing details. In addition, the use of fixed values of the method parameters may result in blurred or unrealistic regions in the resulting images, as they inadequately reflect some features of the face. More flexible techniques have been developed compared to statistical methods. The first is the reconstruction technique, which tries to make the blurred and subsampled version of the target image close to the low-resolution image. The methods aim to focus on sharp details by edge blurring and gradient profiling [13], [14], [15] or to preserve sharp details by removing noise in low-frequency regions [16], [17]. There are methods that combine motion detection from multiple images or combine images using some estimation methods [18], [19], [20], [21]. Due to the focus on sharp details in the reconstruction technique, the solution space is limited. Furthermore, the performance degrades rapidly when the scale factor increases.

Recent research in the field of SR has focused on learning-based methods [22], [23], [24], [25], [26]. These methods use low-resolution images and their corresponding high-resolution counterparts as training data. By using machine learning (ML) and deep learning (DL) techniques, different features of the image are analyzed, and image details are highlighted. ML methods that operate with image features instead of spatial operations are also called feature-based methods. In images obtained using ML, there are limits in focusing on high-frequency details. DL models have been proposed to overcome these problems and achieve more effective results [27].

SR studies based on DL are divided into three categories: local-based methods, global-based methods, and local-global-based methods. In the local-based category, the image or a block of the image is given as input to the deep network model [28], [29], [30], [31], [32], [33], [34], [35]. The details in each region are learned and missing details are revealed. SR methods applied by dividing an image into smaller blocks cause global information loss, blurring and artifacts at the edges of the blocks. Global-based approaches

have been developed to preserve the main structure of the image. Zhou et al. and Huang et al. presented an approach to preserve the global structure of the face by processing the entire face with convolution neural networks (CNN) [36], [37]. In addition to CNN methods, there are methods that include additional face information and maps prepared before training [38], [39]. Additionally, methods that combine spatial and frequency information have been proposed to maintain both frequency details and the overall structure [40], [41]. The shortcomings of local- and global-based methods have been overcome by local-global-based methods that make better use of distance and inter-layer relationships [42], [43], [44], [45], [46], [47].

Recently, deep generative models have emerged. They are preferred for many challenging tasks such as style transfer, data inpainting, image generation, and super-resolution due to their capability to learn and generate realistic data samples (e.g., images, audio, text). The most prominent capability of the methods is to learn the distribution in the data and generate new similar samples through a two-stage process [48]. Examples of methods are GAN [38], [42], [45], [49], Normalizing Flows [NF] [50], [51], Autoregressive Models [52], [53] and Diffusion-based studies [48], [54], which produce high-quality images in the super-resolution domain. Each type of deep generative model has certain advantages but also some weaknesses. High computational cost and slow generation speed of autoregressive models, sub-optimal sample quality of NFs, and limitations such as optimization instability and mode collapse of GANs are the most important weaknesses of deep generative models. On the other hand diffusion models require high computational costs and long training periods due to their iterative nature. Moreover, the process of uncertainty reduction can sometimes lead to artificial effects affecting the natural appearance of facial images. Despite their limitations, diffusion-based methods are prominent in solving problems that require detail and accuracy and are expected to be more widely used in SR applications [48]. Although deep generative networks can produce high-quality and realistic images that appeal to human perception, there is uncertainty about whether their outputs fulfill the authenticity of the target image [55], [56]. Furthermore, despite their perceptual performance, diffusion methods do not perform as expected in super-resolution comparison metrics [57].

In the field of DL, one of the recent advances in computer vision problems is attention mechanisms [58]. During an event or phenomenon, the human brain processes large amounts of data efficiently by extracting meaningful information and focusing on relevant information. Inspired by this feature of the human brain, attention mechanisms have been proposed. The idea of attention in DL aims to focus on the most relevant data. Attention-based studies can be categorized into two classes: soft attention and hard attention. In soft attention, input data is treated globally, and more attention is given to the most important pieces of data while less important ones are ignored. The mechanism mostly uses

differentiable smooth transition functions such as Softmax and sigmoid. This way, during the training period, the model continuously and efficiently improves itself. On the other hand, it has a high amount of input data and computational complexity [30], [31], [32], [58]. Unlike soft attention, hard attention focuses more intensively on the most important parts of the data and thus aims to extract more salient features. However, instead of differentiable functions, variational methods or reinforcement learning (RL) methods are used in hard attention [58], [59], [60], [61]. In RL, more effective and specific learning is achieved, but there can be difficulties in the optimization and training processes.

In this study, we present a deep reinforcement learning (DRL)-based local-global category method that focuses on the face's local details, preserves the face's original form, and utilizes the recurrent attention mechanism. The RL paradigm is known as the ability to find ideal behaviors, especially in complex and uncertain environments. A decision maker interacting with its environment tries to find the best behavioral strategy using its actions. As the decision maker interacts with its environment, it tries to correct its behavior based on feedback (rewards) and creates policies to improve model performance as time progresses. The DRL model in our study detects important facial regions using the visual attention mechanism with a recurrent policy [62], [63], [64], [65]. The proposed method is a FSR method that aims to fully utilize hierarchical features and uses a stochastic process to improve the detected region.

GAN, CNN and diffusion-based methods often use deterministic approaches, which can limit the exploration of the relevant models. The attention mechanism realized using RL is dynamically learned to select and improve the most appropriate region (patch) at each step. Instead of a fixed strategy, a balance between exploration and exploitation for different situations is used to determine the patches to be improved. The knowledge and global contextual clues from the regions optimized in past steps become a guide for future optimizations. Furthermore, training is performed to make optimal decisions by considering the long-term consequences of the decisions made at each step. This results in significant improvements in the final image quality. CNN and GAN-based methods are usually based on a loss function (e.g. l_1 or l_2 -norm), which is not directly related to human perception. On the other hand, DRL optimizes the reward function according to specific metrics such as PSNR and SSIM.

Contributions of the study can be listed as follows:

- A DRL method based on the attention mechanism is proposed by utilizing stochastic processes in the selection of actions. In contrast to traditional methods, actions are generated using a stochastic function that follows a normal distribution instead of being obtained directly from the layers of a DL model. Stochastic actions add flexibility to the model's decision-making process and increase its learning capacity. Moreover, the attention mechanism allows the model to focus on important features, leading to a more efficient learning process.

- Although patch-based improvements are performed at the local level, the chosen reward functions and the way in which they are implemented successfully preserve the overall integrity of the model. Thus, by significantly supporting the preservation of the global content while improving the local details, the proposed methodology enables superior results in terms of overall quality and consistency.
- To realize the patch-based image enhancement process, a DL model based on the Deep Residue Dense Block (RRDB) architecture has been developed. In particular, special improvements have been made to the RRDB blocks in the input and output layers of the model and additional residual links have been added. These improvements allow the model to reconstruct local details more efficiently while maintaining overall image integrity.
- By adding a strategic parameter to the reward functions in the DRL part, accuracy of the action selection process is improved. The proposed parameter extension allows the reward mechanism to evaluate the long-term effects of actions more effectively and thus improve the agent's decision-making.
- A DRL model is used to manage the patch selection processes, while a separate FSR model is integrated to perform improvements to the selected patches. The two independent models are merged into a single model to enable seamless and efficient management of parameter updates and gradient flow. The integrated approach allows the model to optimize both patch selection actions and perform patch improvements efficiently.

II. RELATED WORK

The field of DL-based image SR has been significantly improved by the Super Resolution Convolutional Neural Network (SRCNN) and its derived variations [66]. SRCNN has limited learning capacity and low performance because it does not have a deep network structure. Therefore, it cannot adequately identify local and global relationships and features. In recent years, deeper networks and residual structures have been preferred in SRCNN-based studies [67], [68], [69], [70], [71], [72]. The Very Deep Super-Resolution (VDSR) model extends the basic principles of SRCNN by increasing the depth of the network [70]. Fast Super-Resolution Convolutional Neural Network (FSRCNN) is characterized by fast computations and low memory requirements with an efficient architecture including convolution, scaling and pixel augmentation steps [68]. Enhanced Deep Super-Resolution (EDSR) adopts the basic ideas of SRCNN and stands out with a larger and deeper network structure [67], while Laplacian Pyramid Super-Resolution Network (LapSRN) provides a multi-scale SR approach using Laplace pyramid [71]. MemNet, a memory-oriented network structure, aims to achieve memory optimization and high-quality results [72].

Residual Dense Block (RDB) and Residual in Residual Dense Block (RRDB) approaches have contributed

significantly to the field of FSR [43], [44], [46], [47], [73], [74], [75], [76], [77]. In the case of facial images, RRDB is designed to provide more effective learning by establishing deep connections within the network. The deep network connection structures allow more consistent and meaningful extraction of facial features. In particular, each RDB block accumulates information from different regions of the face to prevent information loss in facial details during the learning process. RRDB represents an evolved form of RDB and aims to represent complex facial features more effectively by adding an additional residual link network within each RDB block. The additional residual connections within the block are more focused on the outputs of the previous RDB blocks, fostering deeper and more detailed learning, especially in facial anatomy. In this context, the deep connections and residual networks offered by RDB and RRDB in facial image SR tasks represent a significant advance by enabling more effective learning and better feature extraction in facial details.

GAN methods, whose main starting point is to generate images and videos, have also been used for image SR applications. It has been observed that the deeper the generating network, the better the output [1], [78], [79], [80], [81]. Although GAN-based methods provide an output rich in detail, sometimes regions with artificial output may occur. In addition, they can be difficult to train due to their unstable nature and complex structure [79]. In the first GAN and SR-based study, SRResNet and SRGAN methods were presented together [42]. In SRResNet, the solution is provided with a deep network and residual blocks, while in SRGAN, the generating network in the GAN method works like SRResNet. In contrast, the supervisory network tries to increase the accuracy of the outputs. The loss functions of the SRGAN model were enriched with the Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN) method, and more detailed outputs were obtained [45]. Generative networks have been proposed to extract more features from the image [38], [78], [80].

GAN-based studies have led to the development of new deep generative models. For example, an autoregressive model was developed that emphasizes textural details by splitting the image into patches and assumes that long-term pixel dependencies are not needed [82]. In another work, the SR-Flow model was introduced, which tries to learn the conditional probability distribution of high-resolution images from low-resolution images and performs inverse transformation [83]. Recently, the diffusion-based SR3 model has been proposed, which realizes super-resolution by transforming a standard normal distribution into an empirical data distribution through successive refinement steps [48]. The model produces diverse and photo-realistic high-resolution images using a fixed number of inference steps independent of the output resolution.

In the methods above, SR images are generated through pixel-level improvements, leading to a fixed receptive field problem in some CNN-based FSR approaches. To address

this, the SFMNet method incorporates frequency information [40]. It preserves global information through amplitude and textural information through phase in the Fourier transform of the image.

With the increasing amount of data and the emergence of complex tasks, it is imperative that DL models focus on the important aspects of the data. For this reason, soft and hard attention methods aim to use the low-resolution input image by emphasizing the important details while ignoring the irrelevant parts of the data.

Among the soft attention methods, channel attention, spatial attention and self-attention are preferred according to the problem types. In channel attention, weight parameters of the model are calculated on a channel basis by evaluating the contribution of each feature channel to certain parts of the input [84]. Spatial attention is a method for enhancing the features of facial regions in the image and ignoring the features of other parts of the image using special residual blocks. It can be used as a separate mechanism or as a complement to channel attention [85]. Self-attention allows for a sharper and more accurate reconstruction of the face by modeling the relationships between the face image's low- and high-level feature spaces better [59], [60], [86]. Soft attention mechanisms are constructs that consider all elements of the input data. However, hard attention mechanisms are preferred when the model needs to focus its attention on a specific region, element or feature.

In the Statistical Hard Attention type of hard attention networks, the model uses statistical information to focus its attention on a particular learning task [87]. For example, a given data point is given more attention based on its performance at previous times in the model's operating period. In Gaussian Hard Attention, Gaussian or similar probability density functions can be used to focus the model's attention on a given task and to make this focus have a smoother transition [61]. In Clustering-based Hard Attention, in order to focus on data points with similar characteristics, the model identifies certain features using clustering algorithms and focuses its attention on these clusters [88].

There have been DRL-based studies that have produced good results in local evaluations [30], [31], [32]. In this study, a method that combines the concepts of DRL and attention is developed.

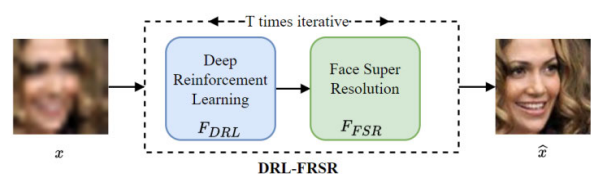


FIGURE 1. Basic functioning of the model.

The method identifies attention regions (a face patch) and enhances the attention region with simple deep network models.

III. METHODOLOGY

Some abbreviations are given in Table 1 for the reader to follow the paper.

TABLE 1. Summary of abbreviations.

Abbreviation	Explanation
DRL	Deep Reinforcement Learning
DRL-SRFR	Deep Reinforcement Learning Based Super Resolution of Face Regions
FH	Facial Hallucination
FSR	Facial Super-Resolution
GRU	Gated Recurrent Unit
RAM	Recurrent Attention Model
RDB	Residual Dense Block
POMDP	Partially Observable Markov Decision Process
RRDB	Deep Residue Dense Block

A. PROBLEM DEFINITION

The aim of DL-based FSR methods is to transform a low-resolution face image x into a high-resolution target face image y . For this purpose, x is processed by the DL method \mathcal{F} with parameter θ to obtain a face image \hat{x} that predicts y . This process can be represented as follows.

$$\hat{x} = \mathcal{F}(x|\theta) \quad (1)$$

In this study, a DRL-based patch selection and RRDB-based FSR method is proposed using the model given in Figure 1. A certain number (T) of regions (patches) are sequentially identified from a face image. At each step, a patch is enhanced and embedded into the image to create \hat{x} . The patch enhancement is expressed as follows

$$\hat{x} = \mathcal{F}_{FSR}((\mathcal{F}_{DRL}(x | \theta_r)) | \theta_h) \quad (2)$$

where \mathcal{F}_{DRL} and \mathcal{F}_{FSR} represent the DRL model and the resolution-enhancing model, respectively. θ_r and θ_h are the parameters of the models. Since DRL is used to select the face regions and the selected region is enhanced by the SR method, the proposed method is named Deep Reinforcement Learning Based Super Resolution of Face Regions (DRL-SRFR).

The \mathcal{F}_{DRL} model, an attention-based approach, is applied to the face image. The output of the model is the likely location and region that needs attention (enhancement). The selected patch is enhanced with the \mathcal{F}_{FSR} model, which includes residual connections and dense layers.

There are some challenges in the \mathcal{F}_{DRL} and \mathcal{F}_{FSR} models. The first one is to select regions with missing details and identify other target regions on the face. Second, how to create the reward mechanism for \mathcal{F}_{DRL} needs to be determined. In a face image, there may be many regions whose details need to be improved. For this purpose, an iterative attention model is used to select T regions. Each step aims to reveal high-frequency regions and details of the face. Attention models require prior knowledge. Therefore, storing the improved regions and images in iterative steps as well as utilizing past model experiences is another challenge in

the iterative nature of the process. In the SR method used in the refinement, the implementation of the RRDB model is modified [45]. RRDB has many layers and the model input is an image. However, in our work, the input of the FSR model is the refined image obtained from the iterative structure and the selected region obtained in the new step. Due to the change of the model input, the residual learning structure in the RRDB layered network structure needs to be modified. Since the output of \mathcal{F}_{DRL} is the input to \mathcal{F}_{FSR} , continuous backpropagation is required to update the model parameters. Finally, choice of T in the iterative structure, size of the regions to be selected, the number of layers of the residual dense blocks, the filter sizes, the parameters of the connection density are also challenges.

B. DEEP REINFORCEMENT LEARNING BASED SUPER RESOLUTION OF FACE REGIONS (DRL-SRFR) NETWORK

Figure 2 shows the general structure of the proposed method. Unlike previous DL-based FSR methods, the SR problem is treated as a Partially Observable Markov Decision Process (POMDP) [89]. The \mathcal{F}_{DRL} model is based on the Recurrent Attention Model (RAM). A gradient-based recurrent policy network is used for the RAM model. At each step t , the decision maker gets a partial observation of the environment as input (current state). Depending on the current state, the decision maker determines the patches that need to be improved.

In the \mathcal{F}_{FSR} model, a SR method is created that is consistent against low-resolution inputs, emphasizes fine details, improves gradient flow with dense connections, and uses a hierarchical and nested residual structure. In DRL-SRFR, the enhanced face image is obtained at the end of iterative process. During the iterative process, each state is the input of the policy layer. The policy layer carefully determines a position and a face patch from the face image. The patch is enhanced with \mathcal{F}_{FSR} and embedded into the enhanced image obtained from the previous steps. When all the selected patches are processed, a high-resolution face image is obtained. At each step of the process, the reward mechanism is activated and the parameters of \mathcal{F}_{DRL} and \mathcal{F}_{FSR} are updated with the help of the back propagation algorithm and the accumulated rewards.

1) DETAILS OF MODEL

Most FSR and hallucination methods do not utilize the inter-pixel correlation in the face image and consider facial regions independently. The proposed method performs sequential learning to generate high-resolution face images, focuses on the distorted parts of the image, and uses inter-region correlations using the enhanced regions.

As shown in Figure 2, the proposed framework consists of three parts. The first part is an agent network and a reward function that provides the current state and rewards for the policy network. The second part serves as the decision maker (policy network) that generates the location and patch

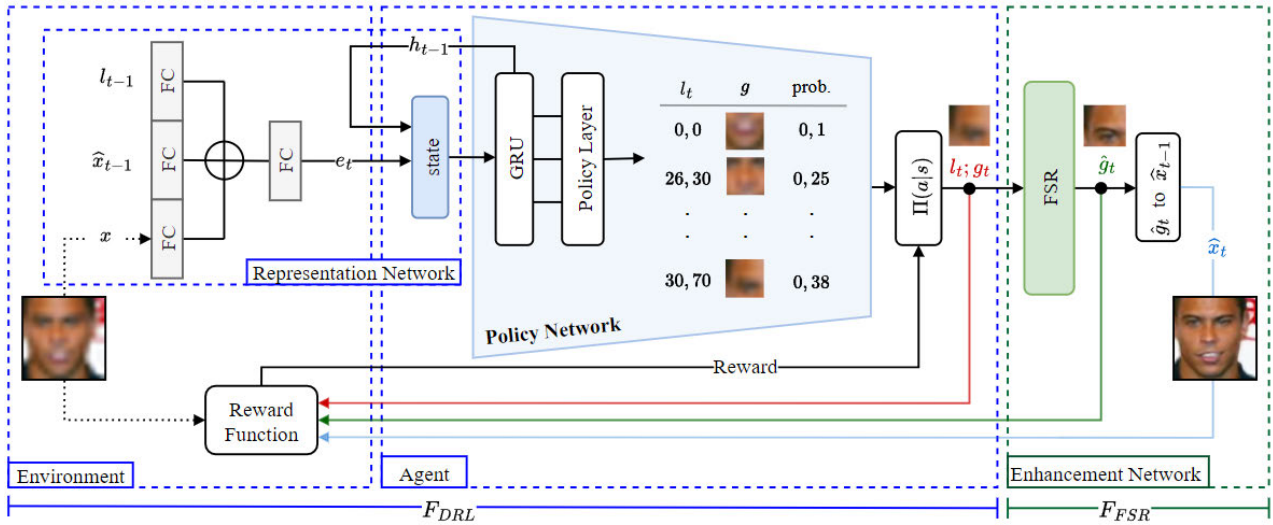


FIGURE 2. The general structure of the proposed method. The policy network acts as a decision-maker. The policy network \mathcal{F}_{DRL} directs the sequential training of the reward function while determining the patches to be considered. \mathcal{F}_{FSR} enhances the patches with deep RRDB layers and adds them to the face image.

according to the current state. The last part is a deep network responsible for the refinement of the selected patch.

The agent network extracts the features of the observations obtained in the iterative process with linear layers. In \mathcal{F}_{DRL} , the policy network includes a gated recurrent unit (GRU) network and a location generator network. With the GRU, information from the past is stored and used to determine new locations. In \mathcal{F}_{FSR} , the RRDB method is implemented using long-range layer relationships using residual links. Details are given in the following sections.

Notations: The objective of the FSR method for a given image $x \in \mathbb{R}^{h,w}$ with steps in iterative processes $t = \{1, 2, 3, \dots, T\}$ is to determine the positions $l = \{l_1, l_2, \dots, l_T\} \in \mathbb{R}^2$ and patches at those positions $g = \{g_1, g_2, \dots, g_T\} \in \mathbb{R}^{u,v}$ sequentially. Here, h and w denote the dimensions of the image, and u and v represent the dimensions of the patches selected from the image. The determined patches are processed in each step to obtain enhanced patches $\hat{g} = \{\hat{g}_1, \hat{g}_2, \dots, \hat{g}_T\}$, and a hallucinated image $\hat{x} = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_T\}$ is formed by embedding each patch into the previous image at the same position in each step. The image formed at iteration T is the output image. The operation for each step is expressed with the following equations. Table 2 provides intermediate structures and outputs generated during the iterative process.

The iterative policy layer $f_\pi(s_{t-1}; \theta_r)$ generates a position information $l_t = \{m, n\}$ for a state s_{t-1} at step $t - 1$ as in Equation (3). With $\{m, n\}$ location information, a patch g_t of size (u, v) is selected from the face image.

$$l_t = \{m, n\} = f_\pi(s_{t-1}; \theta_r) \quad (3)$$

Patch g_t is refined with \mathcal{F}_{FSR} to obtain the refined patch \hat{g} as in Equation (4).

$$\hat{g}_t = \mathcal{F}_{FSR}(g, x, \hat{x}_{t-1}; \theta_h) \quad (4)$$

TABLE 2. Inputs and outputs OF DRL and FSR models.

Step (t)	DRL Input→state	DRL Out→	FSR In→	FSR Out→	Embed \hat{g} to \hat{x}
1	x, \hat{x}_0, l_0, h_0	l_1, g_1, h_1	g_1	\hat{g}_1	\hat{x}_1
2	x, \hat{x}_1, l_1, h_1	l_2, g_2, h_2	g_2	\hat{g}_2	\hat{x}_2
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
$T - 1$	$x, \hat{x}_{T-2}, l_{T-2}, h_{T-2}$	$l_{T-1}, g_{T-1}, h_{T-1}$	g_{T-1}	\hat{g}_{T-1}	\hat{x}_{T-1}
T	$x, \hat{x}_{T-1}, l_{T-1}, h_{T-1}$	l_T, g_T, h_T	g_T	\hat{g}_T	\hat{x}_T

The states that occur at each step in the iterative process, the output of the policy network, the input and output of the enhancement network

At each step, \hat{g}_t is embedded into the face image \hat{x} that was enhanced in the previous step. In Equation (5), the process proceeds in a similar way, resulting in a final image \hat{x} at the end of the T steps.

$$\hat{x}_t = \hat{x}_{t-1} \oplus \hat{g}_t \text{ where } t = 1, 2, \dots, T \quad (5)$$

where \oplus represents embedding of \hat{g} into the previous enhanced image. During the process, the subregion \hat{g} is determined from \hat{x}_{t-1} .

2) DESCRIPTION OF STATE, ACTION AND REWARD IN REINFORCEMENT LEARNING FOR OUR METHOD

RAM is a computational framework designed to understand and utilize attention in a visual environment, especially in the context of activities aimed at achieving specific goals. Theoretically, it builds on the paradigm of sequential decision-making processes in cognitive science and artificial intelligence. RAM outperforms CNN in dealing with clutter and scaling to large input images. RAM is therefore more efficient and effective than CNN in real-world applications with limited computational resources, especially in complex visual

environments. The decision maker uses a limited bandwidth network in RAM to observe its environment. In other words, there is an input (state) that can only be of a specific resolution or cover a particular area. Thus, the decision maker does not consider the whole environment in a step but only the parts it focuses on. Therefore, the decision-maker must integrate information between steps. In response to the actions taken, rewards are generated by feedback from the environment. The reward, reflecting the results of its interaction with the environment, can be immediate or delayed. The main goal of the process is to optimize the total reward at the end of the iterative process. Below, we define the state, action and rewards for the RAM mechanism in \mathcal{F}_{DRL} . Note that while the learning parameters θ should be specified with different indices in both the linearization layers and models, we do not use indices for each function for the sake of simplicity.

State: Each new patch should be able to identify incorrect attention locations resulting from previous steps. Therefore, a state should reflect the previous location information l_{t-1} and the historical information h_{t-1} of the previous locations. In this way, the attention mechanism learns the correct preferences and locations during the training process. Furthermore, a low-resolution face image x is also included as part of the state at each step to select a new location and preserve the holistic characteristics of the image. The sequential inclusion of the improved image \hat{x}_{t-1} in the state preserves the relationship with the past and contributes to both local and global improvement. For the given objectives, the input data x , \hat{x}_{t-1} and l_{t-1} are transformed into the vector e_t by the representation network $f_s(x, \hat{x}_{t-1}, l_{t-1}; \theta)$. Then, current state denoted by s_t is obtained by appending h_{t-1} to e_t as in Equation (6).

$$s_t = (e_t, h_{t-1}) \quad (6)$$

At the beginning of the iterative process, the improved image \hat{x} is assigned the original low-resolution image x and the initial position is the center of the image. f_s is the representation network that linearizes the input parameters. The policy network uses the state to compute the probability of attention regions.

Action: The main task of the policy network is to determine the location of the new patch to be improved. For a state s , all location probabilities of patches with size (u, v) are calculated. From the probabilities, the coordinate of the new patch is determined from the final hallucination image \hat{x} . Due to this structure, when determining the next location, the policy network evaluates the current situation together with the information obtained from past observations and determines where in the image attention should be focused.

An action is realized by a GRU and a linear network. Like RAM methods, A GRU is a memory element with two inputs and one output. The first input is the state, and the other is the history information. The output is a vector of new history information. With GRU, past experiences and current situations are combined to provide a comprehensive representation of which location to choose. A linear network is then used to determine the new location using the

new hidden vector. All possible locations are defined with $l_t = (m, n | m \leq h, n \leq w)$. The location is treated as a random variable with a probability mass function given in Equation (7).

$$\mathcal{P}(a | s, \theta) = \mathcal{P}(l_t | s_t, \theta) = \mathcal{P}((m, n) | s_t, \theta) \quad (7)$$

Reward: DRL methods use reward functions to guide the decision-maker toward the right actions. In the proposed method, local and global reward functions guide the enhancement of both image patches and the entire image. For this purpose, four reward functions and an auxiliary reward parameter E_{pc} , which uses the ratio ‘‘Number of improved pixels/number of all pixels’’ of an image, are employed. Recent studies have shown that using only pixel-wise rewards leads to underestimation of structural differences and low perceptual quality [45]. The perceptual loss function introduced in the ESRGAN method improves the perceptual quality by prioritizing the structural similarity of the image. In this study, the perceptual loss function is also used. Two local reward values over the patch regions and two global reward values over the whole image contribute to determine the total reward. Local rewards are based on the patch \hat{g}_t at each step and the target patch g_t^y that is the target image counterpart of the selected patch g_t from the low-resolution image. Over the two patches, two rewards are calculated with l_2 -norms between \hat{g}_t and g_t^y and the features obtained from VGG19 (\hat{g}_t) and VGG19(g_t^y). (VGG19 is a CNN consisting of 19 layers that performed well in the 2014 ImageNet competition. The features extracted from the feature layers of VGG19 are used to calculate content loss.) Global rewards are generated at the end of the iterative process. Given the final image \hat{x} and the target image y , the global rewards are calculated as mean square error between \hat{x} and y and the l_2 -norm between the features obtained from VGG19(\hat{x}) and VGG19(y). The parameter E_{pc} benefits the global enhancement. It is an indicator of how many pixel values have contributed to the enhancement. During a cycle, all selected pixels are assigned a value of 0 and 1 is assigned to others. The resulting (0,1) map for E_{pc} is given in the ablation studies. The expressions for local and global rewards are given below:

$$rl = \frac{1}{T \times c \times u \times v} \sum_{t=1}^T (g_t^y - \hat{g}_t)^2 \quad (1^{st})$$

$$rl_{per} = \lambda \frac{1}{T \times c \times u \times v} \sum_{t=1}^T \|\phi^{vgg} \hat{g}_t - \phi^{vgg} g_t^y\|_2^2 \quad (2^{st})$$

$$rx = \frac{1}{c \times h \times w} (\hat{x} - y)^2 \times E_{pc} \quad (3^{st})$$

$$rx_{per} = \lambda \frac{1}{c \times h \times w} \|\phi^{vgg} \hat{x} - \phi^{vgg} y\|_2^2 \times E_{pc} \quad (4^{st})$$

Features are extracted using the VGG19 model for the function ϕ^{vgg} . The feature layers are obtained by removing the adaptive mean pooling and classification at the end of the pre-trained VGG19 model. In addition, all parameters of the VGG19 model are frozen and features are extracted based on the trained model parameters.

The total reward R is generated as in Equation (8) and used to train the decision-maker.

$$R = -(rl + rl_{per} + rx + rx_{per}) \quad (8)$$

3) FSR STRUCTURE

Deep networks have been proposed to reveal image details and preserve global structure in SR structures [77]. In \mathcal{F}_{DRL} , the patch obtained from the repeated attention mechanism is enhanced with residual connections, hierarchical features and deep dense layers. Classical CNN-based methods have limitations in extracting image features. RDB structure shown in Figure 3 has been developed to obtain richer image features, increase the use of hierarchical features and multi-level features, and provide better preservation and enhancement of details [43]. RDB structures consist of residual links and dense links. In the ESRGAN method, RRDB structures, which are more complex and efficient than RDB, were developed by combining RDB structures. In RRDB, the basic features of RDB are preserved, while adopting a residue-on-residue structure. An RRDB block contains more than one RDB block. It is particularly effective in preserving details, textures and offers significant improvements in realism and detail in SR tasks [46], [67]. At the same time, RRDB provides a more efficient computation of gradients to deep network layers. Thus, the problem of gradient loss is avoided. The RRDB structure of the ESRGAN method is given in Figure 4. The ESRGAN method employs one input image and includes structures to perform upsampling in the output layers. In our proposed RRDB structure, there are two image inputs and no upsampling is used. RRDB structure is formed as shown in Figure 3 with the previous enhanced image \hat{x} and the patch g obtained from the repeated attention mechanism during the iterative process. An important advantage of the proposed structure is that it provides the gradient flow in a seamless manner. However, a bridge between \mathcal{F}_{DRL} and \mathcal{F}_{FSR} model structures is required. The integration of the new region with the enhanced face image obtained from the $(t - 1)$ th step acts as a bridge between the models. In this way, the network parameters of the models can be updated with the backpropagation algorithm.

There are many RDB layers in an RRDB. In addition, there are residual links in RRDB and RDB structures. Thus, in the initial layers of the network, features such as edges, texture, and color of basic elements in the face image begin to appear while the network delves deeper to reveal more complex and abstract features. As a result, low-level and high-level features are used in deep layers to preserve facial features.

Figure 3 shows the FSR structure of the model for an instant t . Convolution and ReLu activation are applied to $\hat{x}_{t-1} \in R^{c,h,w}$ and $g_t \in R^{c,u,v}$ to obtain the shallow feature vectors $\hat{x}_{t-1}^f \in R^{c,u,v}$ and $g_t^f \in R^{c,u,v}$ respectively. Both vectors are combined to form the input vector of the deep RRDB denoted by $F_1 \in R^{c,u,v}$. Hierarchical features from the input vector are used in the deep RRDB layers. A residual connection is added at two points at the end of the

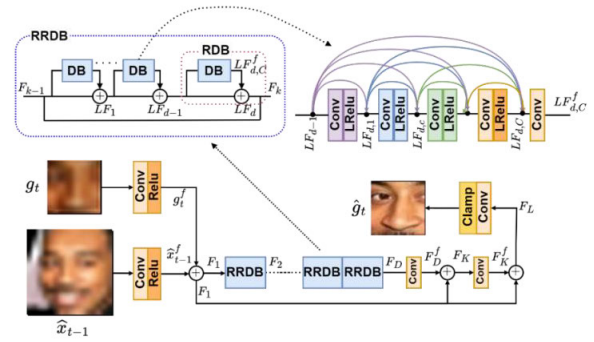


FIGURE 3. The enhancement network in the \mathcal{F}_{FSR} model. \hat{x}_{t-1} is the face image enhanced in the previous step and g_t is the patch region determined in \mathcal{F}_{DRL} . Features are extracted from \hat{x}_{t-1} and g_t and passed through RRDB structures. A residual learning structure is applied at two points. In the RRDB structure, residual links are used in the RDB. In which features are again determined with residual links.

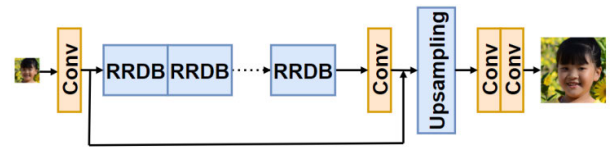


FIGURE 4. RRDB structure in the ESRGAN method.

sequential RRDB structure. With the last convolution layers, the improved patch \hat{g}_t is obtained. There are D times RDBs in an RRDB. In each RDB structure, feature extractions are performed using residual links.

Table 3 summarizes the operations and formulas in Figure 3. Except for the RDB block, the convolution weights of the other layers are represented by W and bias terms are omitted for simplicity. Both RDB and deep RRDB have a continuous memory structure and local feature fusion. σ represents the ReLU activation function, and $W_{d,c}$ and $W_{d,LF}$ in RDB represent the weights of the c -th convolution layer and the final layer, respectively. The clamp function scales the output of the last convolution layer to the $\{0, 1\}$ range to produce \hat{g}_t .

4) DRL STRUCTURE

The representation of the states in the DL structure and the structure of the RAM for the policy network are detailed below.

α : STATE REPRESENTATION

The representation network is shown in Figure 5. It acts as a feature extractor and consists of several connected linear layers. The representation network consists of past information h_{t-1} and new observation features e . While the past information is extracted from the GRU layer, the new observation is generated by sequential linear layers and activation functions as in Equation (9).

$$e_t = f_s(x, \hat{x}_{t-1}, l_{t-1}) = x^f + \hat{x}_{t-1}^f + l_{t-1}^f \quad (9)$$

TABLE 3. RRDB process formulas.

	Process	Formula
RDB	Continuous memory mechanism	$LF_{d,c} = \sigma(W_{D,c}[LF_{d-1}, LF_{d,1}, \dots, LF_{d,c-1}])$
	Fusion of local specialties	$LF_{d,c}^f = W_{d,f}([LF_{d-1}, LF_{d,1}, \dots, F_{d,c}, \dots, F_{d,c}])$
	Local residual learning	$LF_d = LF_{d-1} + LF_{d,c}^f$
Deep RRDB	RRDB Block representation	$F_k = F_{k-1} + LF_d$
	Continuous memory mechanism	$F_D = RRDB(D \text{ times } \dots RRDB(RRDB(F_1)))$
	Fusion of local specialties	$F_D^f = W_{D,f}(F_D)$
Residual link of Deep RRDB	Sequential residual learning	$F_k = F_1 + F_D^f$ $F_k^f = W_{k,f}(F_k)$ $F_L = F_1 + F_k^f$
Input of Deep RRDB	Combination of shallow features of g_t and \hat{x}_{t-1}	$F_1 = g_t^f + \hat{x}_{t-1}^f$
Shallow Features of Inputs	Feature extraction of g_t and \hat{x}_{t-1}	$g_t^f = \sigma(W \times g_t)$ $\hat{x}_{t-1}^f = \sigma(W \times \hat{x}_{t-1})$

Details of the \mathcal{F}_{FSR} network. The RDB has a dense block (DB) structure. Features are combined by creating a continuous memory structure in the DB. The RDB structure is built by collecting the residual link of the input of the DB and the residual link of the output of the DB. With sequential RDBs, the RRDB structure is created with the residual link of the first RDB entry. For RRDBs, a continuous memory mechanism is also established, and feature fusions are performed. \hat{x}_{t-1} and g_t are input to the Deep RRDB network with linear layers and activation function to obtain an improved patch \hat{g}_t using both surface, intrinsic and global features.

Terms having top index f in Equation (9) are generated by applying two consecutive linear layers to the inputs x , \hat{x}_{t-1} and l_{t-1} . Location l_{t-1} is 2×1 vector. x and \hat{x} are smoothed into a single-column vector. Then, vectors of size 256 in the first linearization layer and 512 in the next layer are obtained and summed to obtain e . The other component of the state, h_{t-1} , is the history information in the GRU structure and is a vector of size 512.

b: POLICY NETWORK

Recursive links are used to handle the iterative process and to transfer the information about the past to the following moments. The policy network consists of the GRU layer and the policy layers. The GRU layer takes state (s_t) features and transforms them into an internal state (memory). This internal state stores time information and the accumulation of previous steps. The policy layer transforms the internal state into a given policy probability. The policy probability is

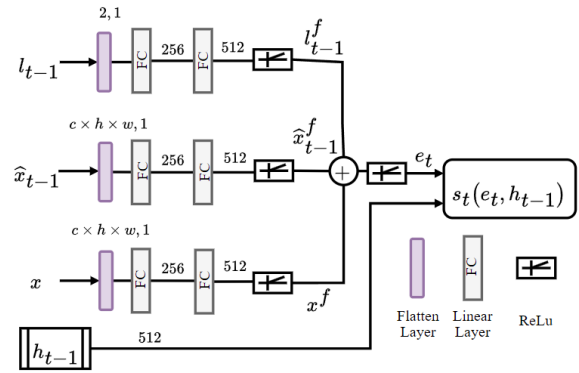


FIGURE 5. A state consists of e and h_{t-1} . The position and the recovered image from the step ($t - 1$) and the original image are combined by extracting features from the linearization layers. A state is determined by adding the RAM's historical information.

a normal distribution that determines which action the model should take next.

Figure 6 shows details of the policy network. A GRU memory cell converts $s_t = (e_t, h_{t-1})$ into the probability of an action group. The memory cell integrates the previous cell memory unit h_{t-1} and the input vector e_t from the current step. For a given state (e_t, h_{t-1}) , the GRU layer $h_t = GRU(e_t, h_{t-1})$ computes an output sequence h_t by recursively computing the activations of the units in the network with Equation (10).

$$h_t = \tanh(W_{si}e_t + b_{si} + W_{hi}h_{t-1} + b_{hi}) \quad (10)$$

where t , e_t , h_t , W_{si} , W_{hi} and b denote step t , the input vector, the hidden vector, the weight matrix corresponding to input e_t , the hidden vector corresponding to the new hidden state and the bias terms, respectively. The position initial values at e_t are $l_0 = [-1, 1]$. In the other steps, they are chosen with a uniform distribution function. In the first step, $\hat{x} = x$. In this way, the first enhanced image is actually the low-resolution image. The parameter \hat{x}_{t-1} in the other e_t states is the image whose sub-region has been enhanced at time $t - 1$. In the sequential learning process, the enhanced image information and the location information from the previous steps are stored in the past action information. Therefore, the new patch selection is influenced by the past information and adjusted to correct the errors caused by previous actions. For GRU, initial values are used as $h_0 \in \mathbb{R}^{512} = 0$.

The policy layer is used to model the decision-making process. At each step in the process, a policy function is defined that determines which action to take. The output of the policy function, which consists of a neural network, is the probability distribution of actions. The policy is implemented by sampling from a Gaussian distribution defined by the parameters mean value (μ) and standard deviation (σ). From the Gaussian distribution, a new location is estimated by a stochastic process. Thus, the aim is to strike a balance between using the current state to predict future phases and taking actions that have not been tried before.

Mean value and standard deviation are critical statistical parameters in RL algorithms. The mean value indicates the central tendency of a data set, while the standard deviation measures how much that data deviates from the mean and determines the spread of the distribution. In a normal distribution, the mean value forms the center of the distribution and represents the most likely action, while the standard deviation determines the width of the distribution, managing uncertainty and the process of exploration. In stochastic processes, the combination of mean value and standard deviation allows the decision-maker to optimize rewards by learning about the environment. When the standard deviation is high, the decision maker is encouraged to sample over a wider range of actions and make an untested decision, while a low standard deviation allows the decision maker to take the confirmed best action in the current situation. This balance allows the decision maker to both maximize rewards in the short term and develop generalizable and adaptive strategies in the long term. Dynamically adjusting the mean value and standard deviation allows the decision-maker to adapt to changing environmental conditions and optimize performance. As a result, using mean value and standard deviation in RL and stochastic processes provides a flexible learning experience for the decision-maker by making the learning process more efficient.

μ is calculated from the relation $\tanh(W_2 \text{ReLU}(W_1 h_t + b_1) + b_2)$ using two linear layers and an activation function. Standard deviation σ is calculated from the reparameterization trick approach defined by $\sigma = \exp(W_\sigma h_t + b_\sigma)$, where $W_1, W_2, W_\sigma, b_1, b_2, b_\sigma$ are the weight and bias parameters of the linear layers. With the calculated parameters probability density function (PDF) defined by Equation (11), is generated

$$P(l_t | \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(l_t - \mu)^2}{2\sigma^2}\right) \quad (11)$$

where l_t is an indicator of the action produced by the decision-maker in a given situation, and $P(l_t | \mu, \sigma)$ gives the probability of the actions sampled by the model. A random location point is selected from the calculated probability distribution. Thus, a stochastic behavior layer is created that allows the decision-maker to explore the current situation and make decisions under uncertainty.

C. MODEL TRAINING

In the \mathcal{F}_{DRL} part of the proposed method, the POMDP model is used to efficiently apply location selection in uncertain environments. The objective of POMDP is to determine the best location selection sequence under uncertainty. In this context, a decision-making strategy should be developed using the belief state so that the decision-maker can choose the best actions in situations that are not fully observable. The belief state represents the probabilities of the decision maker being in the current situation, and the probabilities are continuously updated by observations received and

actions taken. The belief state plays a critical role for the decision-maker to determine the policy and choose the best actions in line with the policy. However, learning the best policy for complex and uncertain environments is very difficult with traditional methods. At this stage, RL methods come into play. REINFORCE algorithm, a gradient-based Monte Carlo Policy Gradient method, is used to determine the best policy based on the belief state. The REINFORCE algorithm updates the policy parameters based on reward signals. Thus, the decision-maker learns to achieve the highest total reward over time. In the process, belief states are used to allow the decision-maker to make the best decisions under uncertainty.

Figure 6 shows the three steps of the sequential process. At each step t , the environment generates a state, denoted by s_{t+1} . Transitions between states are governed by an unknown function $p(s_{t+1} | l_{1:t}, s_{1:t})$ which depends on all previous choices $l_{1:t}$ and previous states ($s_{1:t}$). In this view, the policy network learns a stochastic policy $\pi(l_t | s_{1:t}, \theta)$ that maps a distribution over instantaneous actions onto the history of past interactions $H_{1:t} = s_1, l_1, s_2, l_2, \dots, s_{t-1}, l_{t-1}, s_t, l_t$. At an instant t , choice l_t interacts only with the observation at instant s_t and the delayed reward signals are computed during the previous cycle.

After the action is taken, a set of local rewards (rl_t and rl_{per_t}) is generated for the new observation s_t , and two more global rewards (rx ve rx_{per}) are generated in the last step T . Let the sum of the reward values for step t be denoted by r_t . The objective of the policy network is to maximize the reward sums $R = \sum_{t=1}^T r_t$.

In Figure 6, for a state s_t , g_t at position l_t is selected as the attention patch. The selected patch is enhanced with the \mathcal{F}_{FSR} to obtain \hat{g}_t and embedded into the previous enhanced face image \hat{x}_{t-1} . The reward mechanism is run for the patch and two local rewards (rl , rl_{per}) are calculated. In the last step, two global rewards (rx and rx_{per}) are added based on the final face image and the target image. The E_{pc} in the reward process aims to reach all regions of the face, while the \mathcal{F}_{DRL} model training is intended to target the bad patches that need to be improved.

The expected total reward $J(\theta)$ is expressed as the expected value of sums of r_t under the aggregate probability density function $p(s_{1:t}; \theta)$, which is calculated by the policy parameter θ :

$$J(\theta) = \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^T R_t \right] = \mathbb{E}_{p(s_{1:T}; \theta)} \left[\sum_{t=1}^T R_t \right] \quad (12)$$

To maximize $J(\theta)$, the parameters θ that maximize the expected reward should be calculated. For this purpose, the effect of the policy parameters on the total reward R should be analyzed by taking a gradient on the expected value with the Monte-Carlo Policy Gradient rule [90]. For the derivative of the expected value, the logarithmic derivative rule is applied to form Equation (12). Thus, the agent is trained to obtain the

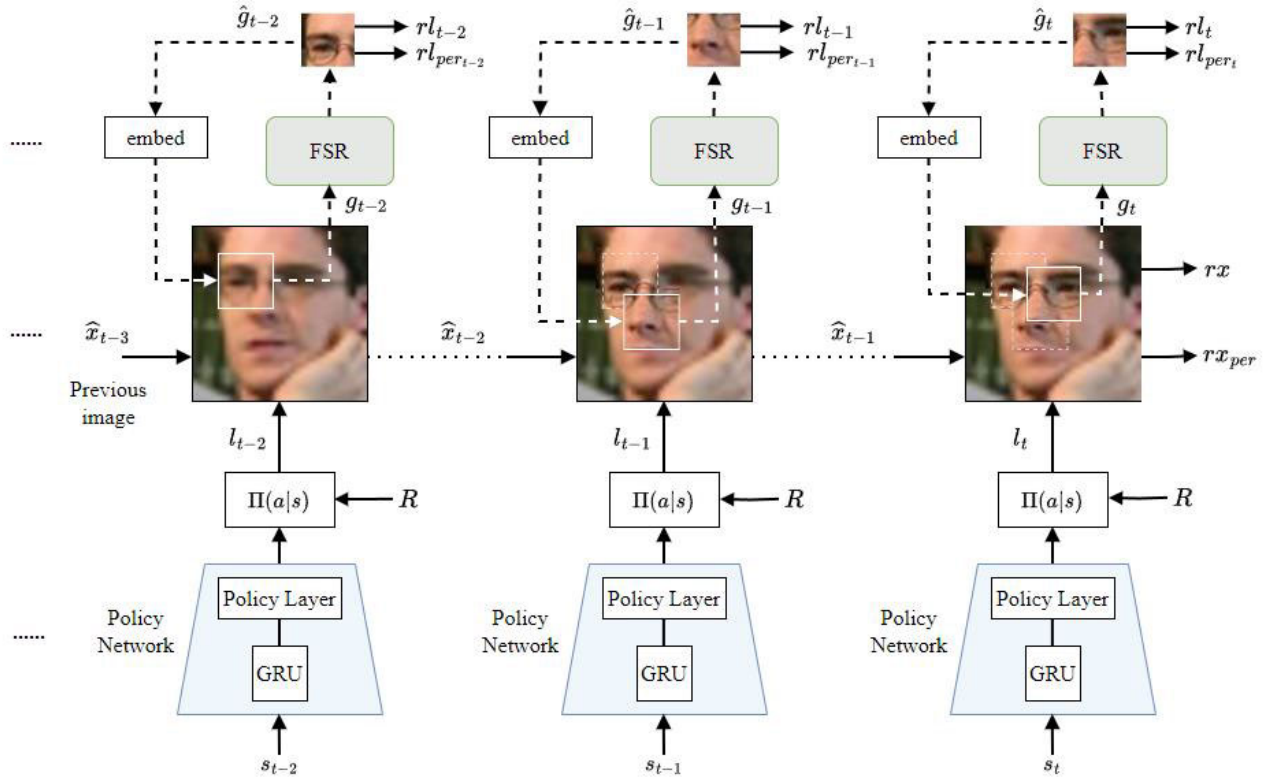


FIGURE 6. The last three steps of the recurrent structure are shown. The patch from the previous image is enhanced with the FSR model and embedded into the incoming image. Rewards are calculated from both the patch and the resulting image. In the last step, the enhanced image is output.

highest total reward R:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{(s_{1:T}; \theta)} \left[\nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \left[\sum_{t=1}^T R_t - b_t \right] \right] \quad (13)$$

In Equation (13), b_t is a base that depends on the state independent of the action. The basis is a function of the value in the corresponding batch, increasing the probability of choosing rewarding locations and decreasing the probability of choosing the opposite. As a result, gradients of the total expected reward over each batch and the number of patches are calculated. Let M be the number of batches, the final gradient rule is given in Equation (14) and is also known as the REINFORCE rule.

$$\nabla_{\theta} J(\theta) = \frac{1}{M} \sum_{i=1}^M \sum_{t=1}^T \nabla_{\theta} \log \pi(a_t^i | s_{1:t}^i; \theta) (R_t^i - b_t) \quad (14)$$

Using the current policy, the goal is to obtain samples of $s_{1:T}$ interaction sequences and then increase the logarithmic probability of selected actions by adjusting the parameter θ and selecting actions with high cumulative reward, avoiding selecting actions with low reward.

The training processes explained so far are in the scope of RL. Mean Squared Error (MSE) is used for the training processes of local regions in the RRDB network and

for updating the RRDB network parameters. The difference between the improved image and the ground truth image is used as the loss function. The loss function for patches and the global loss function are calculated using MSE. The loss function for the i -th patch is calculated as $L_i(g) = \frac{1}{n_i} (\hat{g}_i - g_i^y)^2$ where $n_i = c \times u \times v$. The total loss function for T patches is given as Equation (15).

$$L(g) = \sum_{i=1}^T L_i(g) \quad (15)$$

At the end of the T steps, the global loss function is calculated by Equation (16).

$$L_G = \frac{1}{k_i} (\hat{x} - y)^2 \quad (16)$$

where $k_i = c \times h \times w$ is the number of pixels in the image. In the \mathcal{F}_{FSR} model, improving only \hat{g} with RRDB and embedding it in the spatial space causes the gradients to disappear due to the global loss function is between \hat{x} and y . To solve this problem, \hat{x}_{t-1} is included in the model in the formation of \hat{g} by establishing a relationship to ensure gradient flow. Not only the selected patch g but also \hat{x}_{t-1} are applied to the RRDB input in the FSR structure. Therefore, the loss function between \hat{x} and y allows the parameters of \mathcal{F}_{FSR} to be updated. The final loss function used to determine the optimal

parameters of the method is calculated from Equation (17).

$$L = L(g) + L_G \quad (17)$$

When calculating the derivative of the final loss function with respect to the model parameter θ , the chain rule is applied and the gradients for each $L(g)$ and L_G are calculated using Equation (18).

$$\frac{\partial L}{\partial(\hat{g}, \hat{x})} = \frac{2}{n_i T} \sum_{i=1}^T (\hat{g}_i - g_i^y) + \frac{2}{k_i} (\hat{x} - y) \quad (18)$$

IV. EXPERIMENTS

Several benchmark experiments were conducted to demonstrate the superiority of the proposed model. First, the training data and implementation details will be introduced. Then, the proposed DRL-SRFR model will be comprehensively compared with other recent methods in the literature. In the comparisons, standard performance metrics in SR and FH will be used. In particular, the performance of the proposed model depending on different parameters will be analyzed in detail, and its advantages over previous methods will be emphasized.

In addition, several analyses were performed to evaluate the performance of our model on different datasets. The analyses were aimed at assessing the general applicability of the model and its ability to adapt to various conditions. Finally, detailed ablation studies were performed to further elucidate the contribution of each component in the model.

The proposed method is compared with SRCNN [66], VDSR [70], SRGAN [42], IPFH [30], Attention-FH [31], SFMNet [40], ESRGAN [45], SR3 [48], and SwinIR [91]. These methods can be categorized into six groups: (i) general image super-resolution: SRCNN and VDSR; (ii) facial hallucination: SFMNet; (iii) Diffusion: SR3; (iv) Transformer: SwinIR; (v) generative adversarial learning: SRGAN and ESRGAN; (vi) RL: IPFH and Attention-FH. The methods in the first group address normal image restoration, while the algorithms in the second to sixth groups are widely applied for face image restoration. The third group includes the diffusion-based method, which produces high-quality images by gradually removing noise. The fourth group deals with the Transformer-based method, which obtains high-resolution images using attention mechanisms in multiple layers. The methods in the fifth group are widely used in image reconstruction and have achieved impressive results.

Except IPFH [30] and Attention-FH [31], the other methods were implemented by us. While training the compared methods, the training parameters given in the related papers were used. All methods were trained on a computer with NVIDIA 3090 24GB GPU, Ryzen 5 5600 CPU and 32GB RAM.

A. DATASETS AND IMPLEMENTATION DETAILS

1) TRAINING SETS

The following four training data sets were used to train and test our model.

- CelebA [92]: A large-scale dataset containing 202,599 wildlife face images with 10,177 identities. 40000 images were used for training and 4000 images were used for testing.
- BioID [93]: an open dataset containing 1,521 gray-scale face images taken under laboratory conditions. 1,293 images were used for training and 228 images were used for testing.
- LFW [94]: 13,233 real face images with 5,749 identities. 11248 images were used for training and 1985 images were used for testing.
- PubFig [95]: It is a large dataset containing 58,797 real-world face images collected from the web. We used 12000 images for the training set and 2000 images for the test set.

2) IMPLEMENTATION DETAILS

In all methods and datasets, RGB format images are cropped to (128,128) with the center region centered. The training batch size is 20. For accurate and reliable benchmarking, all methods are trained only on the corresponding training set without using other datasets in the pre-training. The method was evaluated with scaling factors of 4 and 8 to evaluate different cases and to understand the competence of the model. We also normalize the input images in the range [0, 1]. T for the policy network was set to 15 to achieve a balance between model speed, sufficient area coverage and accuracy. The DRL-SRFR model was trained using ADAM gradient descent with a base learning rate of 10^{-3} , momentum term of 0.9. Every 15 epochs the learning rate was reduced by 0.5. The σ was set to 0.25, the number of RDB blocks was set to 3 and the number of RRDB blocks was set to 8. The number of epochs was chosen as 150 and the batch size as 10. Based on the given parameters, the amount of RAM required for the method to run is approximately 6 GB. To give an idea, the training time for the LFW dataset is 925 s for one epoch. The training time increases as the number of RRDB blocks increases. For example, with a 12-block structure, the training time is 1392 seconds.

The compared methods were implemented using the following parameters: ESRGAN: number of blocks 23, batch size 16, number of epochs 1000; SWINIR: window size 8, number of transfer blocks 6, embedded size 180, number of num heads 6, multi-layer perceptron layer size 2, batch size 8, number of epochs continued until the approximate performance value given in the method. SR3: the number of internal channels is 64, the channel expansion factor is [1,2,4,8,8,8], the resolution at which the attention mechanism is applied is 16, the number of residual blocks at each resolution level is 0.2, the noise addition and subtraction processes are repeated 2000 times during the training process of the model, the noise addition and subtraction rate is linearly increased and decreased from beta value $1e-6$ to $1e-2$, the total number of iterations is maximum 500000. SFMNET has a two-branch structure, an 8-layer structure is used and fused

TABLE 4. Comparisons of PSNR values according to scale ratio of 4 and 8.

Dataset	Scale	Bicubic	SR		Face Hallucination	Diffusion	Transformer	GAN		Reinforcement Learning		Ours
			SRCNN	VDSR	SFMNet	SR3	SwinIR	ESRGAN	SRGAN	IPFH	Attention-FH	
BioID	x 4	24.59	25.99	29.56	31.79	31.67	30.59	31.66	28.29	-	33.38	34.12
	x 8	20.24	21.15	23.12	26.50	28.29	25.17	28.35	24.05	26.59	27.81	28.86
CelebA	x 4	25.76	26.42	28.08	29.25	31.59	29.89	30.75	28.44	-	30.58	30.42
	x 8	21.84	22.05	23.96	24.80	30.04	23.41	28.50	24.61	26.42	26.14	27.26
LFW	x 4	26.79	28.13	31.73	35.01	32.68	33.69	33.05	32.23	-	32.93	34.29
	x 8	21.92	23.22	24.07	28.15	27.35	26.56	28.49	25.99	24.93	27.81	28.66
PubFig	x 4	24.76	25.21	28.05	33.04	34.20	32.80	32.70	27.44	-	28.87	33.23
	x 8	20.75	21.31	21.94	27.72	28.64	26.42	29.01	23.45	-	23.59	29.30

Comparison of the developed model with state-of-the-art methods of the same and different constructs

TABLE 5. Comparisons of SSIM values according to scale ratio of 4 and 8.

Dataset	Scale	Bicubic	SR		Face Hallucination	Diffusion	Transformer	GAN		Reinforcement Learning		Ours
			SRCNN	VDSR	SFMNet	SR3	SwinIR	ESRGAN	SRGAN	IPFH	Attention-FH	
BioID	x 4	0.8056	0.8268	0.8627	0.8971	0.9518	0.8860	0.8568	0.8630	-	0.9418	0.9524
	x 8	0.6189	0.6015	0.6601	0.7174	0.7895	0.7011	0.6359	0.6881	0.78	0.8568	0.8627
CelebA	x 4	0.7672	0.7634	0.8075	0.8670	0.7902	0.8747	0.7819	0.8300	-	0.8711	0.8705
	x 8	0.6107	0.6120	0.6665	0.7001	0.6324	0.6662	0.5318	0.6753	0.80	0.7441	0.7640
LFW	x 4	0.8469	0.8599	0.8945	0.9353	0.9011	0.9215	0.8915	0.9266	-	0.9418	0.9573
	x 8	0.6712	0.6907	0.7074	0.7859	0.7857	0.7433	0.6679	0.7381	0.69	0.8568	0.8643
PubFig	x 4	0.7600	0.7708	0.8262	0.9262	0.9113	0.9141	0.8827	0.8197	-	0.8587	0.9297
	x 8	0.5782	0.5819	0.5822	0.7758	0.7057	0.7507	0.6582	0.6747	-	0.6805	0.8341

Comparison of the developed model with state-of-the-art methods of the same and different constructs

for up and down feature extraction in frequency branch and spatial branch operations, batch size is 8 and the number of epochs is 400 during training. IPFH and Attention-FH: since there is no official code page and no other code implementation, the values and images given in their articles were used. SRCNN: the images were first trained using 33×33 patches with epochs of 50000 and batch size 16, the feature layer was followed by a non-linear layer, and the SR image was reconstructed in the last convolution layer. VDSR: there are a total of 20 convolution layers and each layer uses filters of size 3×3 , each convolution layer contains 64 filters, the batch size is set to 16 and the model is trained for 100 epochs.

3) EVALUATION METRICS

Like many previous works, we adopt the widely accepted visual evaluation metrics of Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) [30], [31], [42]. The mathematical expressions for PSNR and SSIM are given in Equations (19) and (20).

$$PSNR(\hat{x}, y) = 10 \times \log\left(\frac{255^2}{\|\hat{x} - y\|_2^2}\right) \quad (19)$$

$$SSIM(\hat{x}, y) = \frac{(2\mu_{\hat{x}}\mu_y + C_1)(2\sigma_{\hat{x}y} + C_2)}{(\mu_{\hat{x}}^2 + \mu_y^2 + C_1)(\sigma_{\hat{x}}^2 + \sigma_y^2 + C_2)} \quad (20)$$

where \hat{x} is our resolution-enhanced model output and y is the target image. $\mu_{\hat{x}}$ and μ_y denote the averages of the pixel values for \hat{x} and y while $\sigma_{\hat{x}}^2$ and σ_y^2 stand for the variances

of \hat{x} and y respectively. $\sigma_{\hat{x}y}$ denotes the covariance of the two images. C_1 and C_2 are small constants and are used to avoid division by zero.

Higher PSNR values indicate better quality. However, PSNR should be interpreted with caution as it does not adequately reflect the perceptual characteristics of the human eye. SSIM attempts to measure image quality more closely to the perception of the human eye and is therefore often used in image processing and quality assessment. The SSIM score takes a value between 0 and 1. An SSIM score approaching 1 indicates a high similarity between two images, while a score approaching 0 means low similarity.

We compared our method visually and semantically with other existing methods. In the comparisons, PSNR and SSIM values are presented in tables to evaluate the model output results with increased resolution of the face images. Visual comparisons were also made between various methods for different datasets. Since the training model files of the Attention-FH model were not available, ready-made images from the article were used for comparison purposes.

B. COMPARISON OF METHODS

Applying SR to facial images requires attention to several critical factors. First, the algorithms must ensure that the high-resolution image has a natural and realistic appearance, preserving skin tones, textures and facial features. Special emphasis should be placed on enhancing details, accurately reconstructing crucial details such as eyes, lips and hair

strands. Second, color accuracy is important to preserve original colors and avoid chromatic aberration. Third, the ability to perform SR operations quickly and effectively in a variety of image scenarios is also crucial. Fourth, the training data should include a variety of examples for the model to understand facial details.

In addition, the level of detail is a key factor; in particular, details related to the main components of the face such as the mouth, eyes, nose, hairstyle may have different priorities depending on the intended use. Local details such as gestures, eyebrow movements, ears and accessories can also be effective in determining the overall facial expression. However, it is also important to understand how the human eye perceives these details rather than focus only on technical details. Examining the global structure is significant when considering perceptual conservation. The image obtained as a result of the face SR process should maintain perceptual integrity in terms of its overall structure. Taking the global structure into account allows us to capture the naturalness and proportionality of facial features that cannot be achieved with local details alone. In some applications such as security, details may be important while in another context the overall form and aesthetics of the face may take precedence. In conclusion, in addition to technical metrics such as PSNR and SSIM, understanding human perception and setting priorities based on the intended use plays a critical role in the successful evaluation of facial SR technologies.

From the quantitative and visual comparison results, it is observed that the model behavior varies depending on the scaling factor and data sets. For a scaling factor of 4x, the current methods and the proposed method give similar results. Moreover, generative network-based methods sometimes detect additional artifacts that are not present in the image. Although these artifacts improved visual perception, they resulted in both the lack of preservation of personal features and deterioration in comparison metrics. For a scaling factor of 8x, the artifacts become more apparent in the output. In terms of visual perception, the results leave a positive impression, such as people looking in different directions or facial expression changes compared to their original state. The proposed model has produced successful outputs both locally and in the whole image and is a competitive method in terms of visual perception performance. From the results, it is seen that the proposed method produces quality outputs in terms of facial expression in terms of gaze, facial expression and emotion conveyance. In addition, the proposed method produces better results in quantitative and especially SSIM evaluations.

Table 4 and Table 5 show the superiority of the proposed method over the compared methods in terms of PSNR and SSIM metrics at different scaling factors (4x / 8x) on BioID, CelebA, LFW and PubFig datasets. In the empirical tests of different models on different datasets, some results stand out. Results of the classical FSR methods are observed to be quite low compared to the other methods. There is competition among other methods in various aspects. The SFMNet model

produces satisfactory results at a scaling factor of 4x, but poor results at a scaling factor of 8x.

The ESRRGAN model is the second-best method in Table 4 in terms of PSNR, but worse than most methods in Table 5 in terms of SSIM. The variation in the amount of datasets significantly affects the output of the models. Even though the SwinIR model produces good results, it lags behind the other models due to the large number of datasets required for transformer model training. The diffusion-based SR3 method produces better results with large datasets due to its generative nature, but it does not perform as well as our method on small datasets. DRL-SRFR gives good results in terms of local detail extraction and global preservation even when trained with a small dataset. As the number of datasets increases, the metric and perceptual performance of the DRL-SRFR method increases.

The variability in the performance depending on the dataset size is evident from the significant difference between the results of BioID with the least data and CelebA with the most data. However, in terms of the SSIM metric, DRL-SRFR gives the best results regardless of the dataset size. Compared to the second-best method, DRL-SRFR achieved an average PSNR gain of 0.74 dB for the BioID dataset at 4x scaling factor.

For the CelebA dataset, it gives the third-best result with a difference of 1.17 dB, for the LFW dataset with a difference of 0.72 dB and for the PubFig dataset with a difference of 0.97 dB. In terms of the PSNR measure at 8x scaling factor, the DRL-SRFR method offered an average gain of 0.51 dB, 0.17 dB and 0.29 dB on the BioID, LFW and PubFig datasets, respectively, compared to the second closest model. In the CelebA dataset, it ranks third with a difference of 2.78 dB. Similarly, for a scaling factor of 4x, the DRL-SRFR method yielded average SSIM gains of 0.0006, 0.0155 and 0.0035 for the BioID, LFW and PubFig datasets, respectively, compared to the second-best model. It gives the third-best result only on the CelebA dataset with a SSIM gain difference of 0.0039. According to the 8x scaling factor SSIM analysis, DRL-SRFR outperforms all other methods, producing average SSIM gain differences of 0.0059, 0.0199, 0.0075 and 0.0583 compared to the second-best model.

The performance of the proposed method in preserving local details and global view compared to the other methods can be seen in Figure 7. When the figures and quantitative values are evaluated in general, it is seen that Attention-FH and DRL-SRFR give better results. The other methods produce either significantly lower quantitative values or images with poor visual perception (blurred and artifactual output). In Figure 7, the glasses of the person in column are almost imperceptible in the other methods, whereas with DRL-SRFR they are clearly visible in the SR3 model. Compared to the output of the ESRRGAN method, which gives the second-best result according to quantitative metrics, DRL-SRFR gives an output closer to the original image in terms of holistic and visual perception. Although the visual perception of the SR3 model was high, the squinting of the right eye and

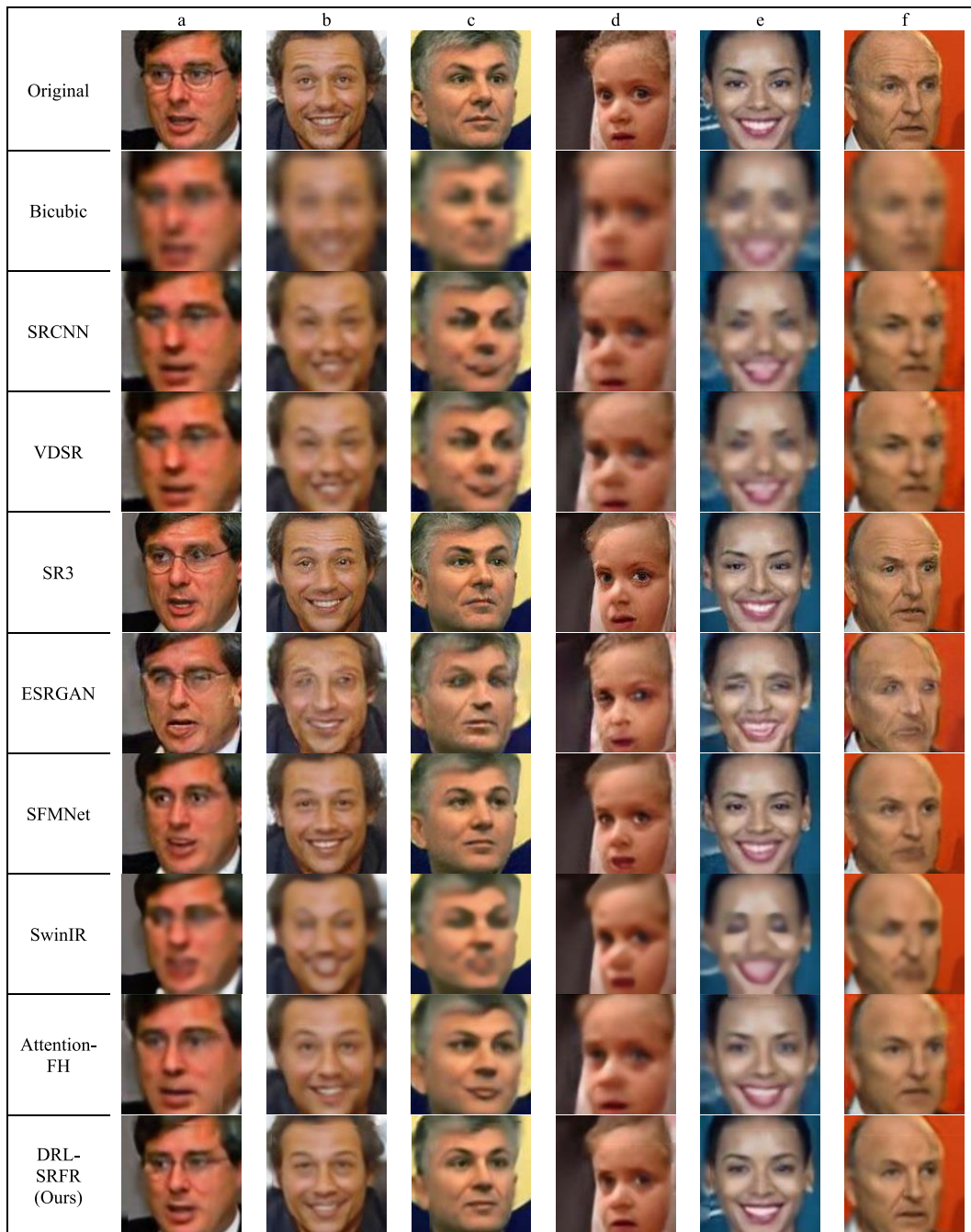


FIGURE 7. Outputs of different methods on the LFW-funneled dataset for scaling factor 8.

the more open left eye caused a deviation from the original image. In the Attention-FH output, an expression close to the expression of surprise is perceived on the face, while there is no such perception in the original image. When the person in column b is analyzed, all the other methods produce eye and facial structures that are far from the original. In contrast, the DRL-SRFR method produces an image with a lower resolution of the eye region compared to Attention-FH. Still, the general eye structure and facial expression are more reminiscent of the original. In the S3 method,

although the visual perception is high, the shift of the person's gaze in different directions, artificial formations in the eyes, teeth and mouth caused a departure from the original image. In column c, while the blackness in the eye and eyebrow area is more intense and mixed in the outputs of the other methods or the whiter tone pupil artificial formations are observed in the SR3 method, it is more discrete and clearer in the DRL-SRFR output, the lip structure and nose details are closer to the original. In column d, the lines formed in the eye bags in the SR3 model caused differentiation. In the eye

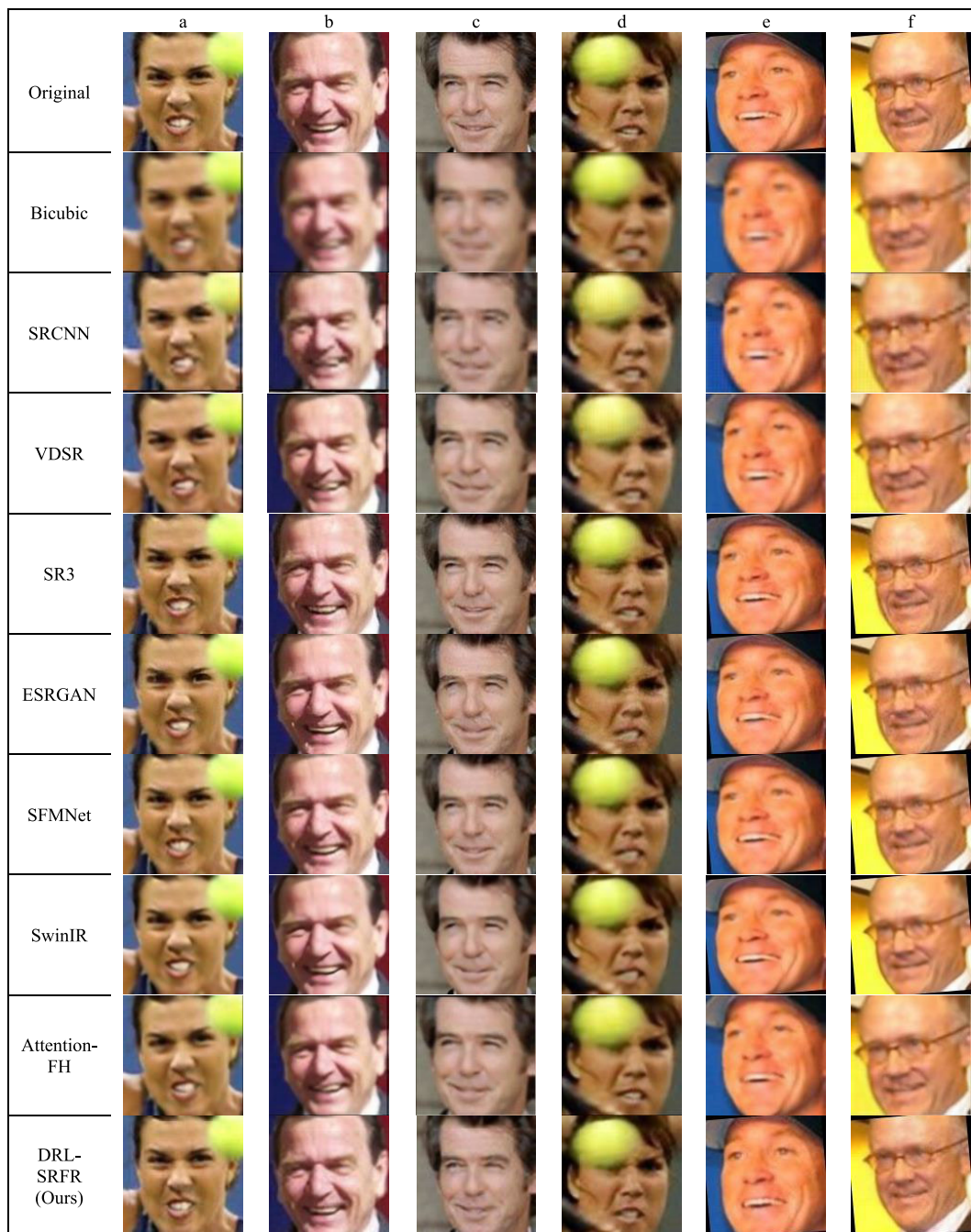


FIGURE 8. Outputs of different methods on the LFW-funneled dataset for scaling factor 4.

region, our model performed better than the other methods, but not in the mouth region. However, the global structure is closer to the original. In column e, while the other methods tried to increase the resolution, the length and width of the face changed considerably compared to the originals. In the Attention-FH model, the local eye resolution is improved and the eye is prominent, but the original structure is not preserved. In the SR3 model, although visual perception was good, a different emotion was conveyed due to the squinting of the eyes. In the DRL-SRFR output, although the eye

resolution appears lower, the output is close to the original in terms of eye region, general facial features and size. In the Attention-FH output, which gives the best result close to DRL-SRFR in column f, the resolution of the right eye looks better, but there is a confused expression that is not present in the original. There is also a change in gaze direction in the left eye. In the SR3 model, details such as the line on the forehead are clearer than all methods, but the person’s gaze sense has changed due to the structural change in the right eye. In the output of the proposed method, the expression is close to the

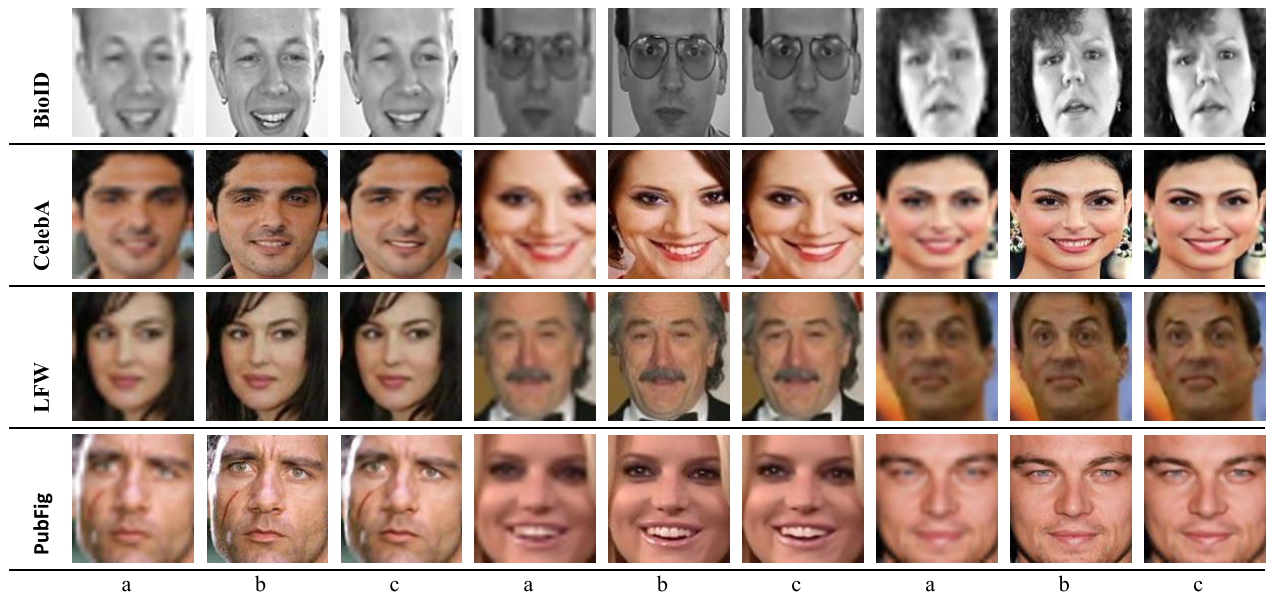


FIGURE 9. Outputs of our DRL-SRFR model for scale 4 on BioID, CelebA, LFW and PubFig datasets. (a) LR inputs. (b) Ground truth. (c) Ours.

original and the eye directions are correct. In addition, when the region between the nose and the mouth is examined, the facial contours are clearer and closer to the original. Figure 8 shows that the resolutions provided by Attention-FH, SR3, SFMNet, SFMNet, ESRGAN and methods other than our model are significantly lower for 4x scaling factor in the LFW dataset. Although Attention-FH, SR3 and ESRGAN give the second-best results, they can provide good results in visual perception globally at 8x scaling factor. The results show that the DRL-SRFR method outperforms all competing methods in terms of global visual perception and quantitative values and achieves impressive success in preserving local details and overall appearance.

Figure 9 shows the perceptual closeness of the proposed method output to the original image for a scaling factor of 4. In the BioID dataset, local details and global structure are preserved in black and white contrast values. For the first person in the BioID dataset, the mouth structure is preserved, and the same smile perception is captured. Similarly, in the second person, the glare details on the glasses are successfully captured and the global similarity is quite close. In the third person, the hair curl detail is distinct enough to be detected and the overall similarity is quite close to the original. In the CelebA dataset, the facial expression of the first person is successfully captured in the same way and the stubble details are noteworthy. In the second person, the dimple lines caused by the smile are created quite successfully, while the general facial expressions are preserved. In the third person, earring details as accessories and eyebrow structure details are successfully created. Finally, in the LFW dataset, the hair separation detail and the similarity of the shape of the lip structure stand out in the first person. In the second person, despite the presence of a mustache, the general expression of

the face is preserved and the hollows on the sides of the chin are equally prominent. In the third person, it is noteworthy that the surprised expression on the face is preserved and the personal posture in the mouth structure can be created in the same way. Finally, in the PubFig dataset, in the first person, the facial scar is significantly similar to the original. In the second person, it is noteworthy that the alignment of the teeth is dimpled and the flat areas are rendered to be close to the original. In the third person, the lines on the cheek and the eye bags are very close to the original, with preservation of the emotional expression in the gaze. In conclusion, the DRL-SRFR method produces an output that has perceptual similarity to the original image in terms of local and global details.

Figure 10 presents a comparison of the original high-resolution (HR) and low-resolution (LR) images and their enhanced versions by various super-resolution methods (ESRGAN, SwinIR, SR3 and DRL-SRFR). Images of selected individuals are from the PubFig dataset. The values in the row below the images refer to the different local regions of the face images (eyes, nose, mouth) and show the PSNR and SSIM values calculated from the enhanced versions of the local regions and the versions generated from the HR image. For the first face, the DRL-SRFR method outperforms the other methods (except SwinIR) with a PSNR of 33.52 and SSIM of 0.924 in the overall image, and gives very good quantitative values, especially in the mouth region. However, only quantitative evaluation or visual perception should not be used to contrast models. For example, SwinIR, which gives the highest quantitative value, produces worse images than SR3 and our model in terms of visual perception. On the other hand, the ESRGAN model failed to produce favorable results in terms of quantitative values and visual







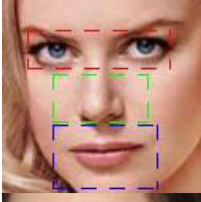
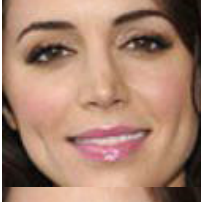


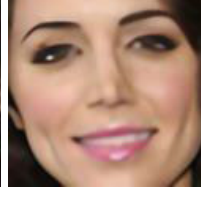
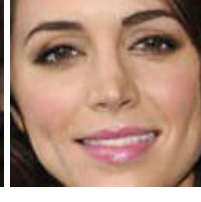
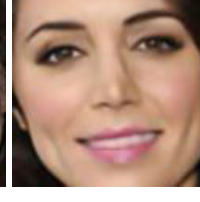
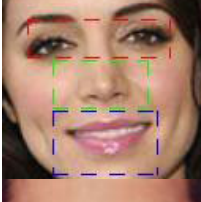


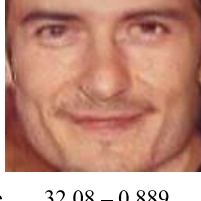
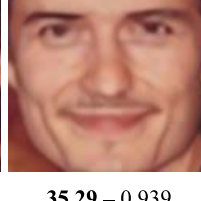
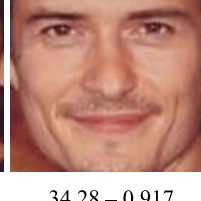
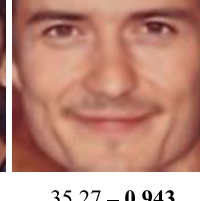
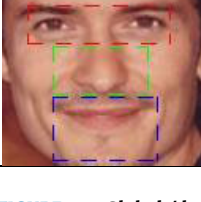
GT	LR	ESRGAN	SwinIR	SR3	DRL-SRFR
					
	Image	31.27 – 0.874	33.79 – 0.931	33.19 – 0.918	33.52 – 0.924
	Mouth	30.84 – 0.882	32.94 – 0.931	32.70 – 0.922	32.72 – 0.905
	Nose	30.64 – 0.888	34.80 – 0.930	34.12 – 0.922	35.28 – 0.944
	Eyes	29.65 – 0.837	30.85 – 0.886	30.68 – 0.925	30.76 – 0.887
					
	Image	31.13 – 0.862	34.37 – 0.929	33.19 – 0.906	34.45 – 0.932
	Mouth	30.47 – 0.810	33.03 – 0.882	32.72 – 0.877	33.08 – 0.872
	Nose	31.82 – 0.890	35.86 – 0.923	35.04 – 0.919	35.95 – 0.933
	Eyes	29.69 – 0.791	31.33 – 0.873	30.75 – 0.868	31.90 – 0.925
					
	Image	32.08 – 0.889	35.29 – 0.939	34.28 – 0.917	35.27 – 0.943
	Mouth	31.29 – 0.841	33.86 – 0.897	33.41 – 0.884	34.45 – 0.912
	Nose	32.30 – 0.883	36.35 – 0.940	34.90 – 0.919	36.7 – 0.943
	Eyes	30.25 – 0.863	31.76 – 0.873	31.87 – 0.922	32.43 – 0.933

FIGURE 10. Global / local visual and quantitative outputs of our DRL-SRFR model for scale 4 on PubFig datasets.

perception. For the second face, our model outperformed the other methods with 34.45 PSNR and 0.932 SSIM values in the overall image. In the quantitative evaluations of mouth, nose and eye parts, the model outperformed the other methods. For the third face, high performance was achieved with 35.27 PSNR and 0.943 SSIM values in the overall image. Our model outperformed the other methods in PSNR and SSIM evaluations of all local regions. The findings in Figure 10 demonstrate the superiority of DRL-SRFR in visual enhancement performance by providing higher PSNR and SSIM values, especially in important regions of the face. In summary, the proposed method allows for sharp and detailed facial images while preserving detail and structure quality.

Considering the parameters used for the 4x scaling factor and batch size = 1, the computational load and parameter counts for the various methods are given in Table 6. FLOPs (Floating Point Operations per Second) refer to the number of floating-point operations a model performs during inference, while computational load GMAC (Giga Multiply-Accumulate Operations) measures the total number of multiplication and addition operations in billions that the model needs to process an input. The number of parameters, flops and computational load of our method are low compared to current SR studies. The time to render an image is similar for all methods except diffusion. In the diffusion method, an image can be created in 61 seconds.

TABLE 6. Process loads.

Model	TFlops	Parameters Count	Test Time (Second)	Computational load (GMAC)
ESRGAN	84,1	43,242,820	0.99	42.06
SWINIR	60,3	11,900,199	1.39	30.17
SR3	91,7	97,807,491	61	45.89
DRL-SRFR	26,7	16,579,942	1.55	13.35

Comparison of our method (DRL-SRFR) with competing models in terms of computational load.

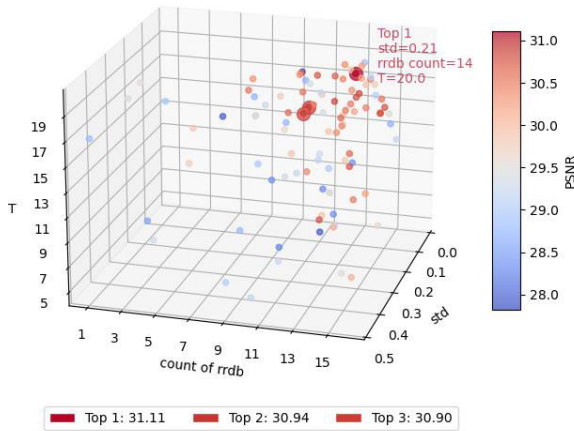


FIGURE 11. PSNR performance graph according to T , RRDB number and standard deviation values. The best result is realized for $T = 20$, RRDB number 14 and $std = 0.21$.

C. TUNING OPTIMAL HYPERPARAMETERS

There are four hyperparameters in the DRL-SRFR method, namely the number of patches T in the process of enhancing an image, the number of RRDBs, the patch size and the standard deviation in the REINFORCE method. Ablation studies related to the parameters are given in the next section. In this section, we discuss the determination of T , the number of RRDBs and the standard deviation σ to maximize the PSNR.

The best values of the parameters were determined using the Tree-Structured Parzen Estimator (TPE) method. TPE is a Bayesian-based optimization method. Our conjecture is that a given combination of parameters will maximize the PSNR. In TPE, a priori probabilities are first determined with random initial values. The initial values provide a preliminary idea of the model’s performance. Then, using the results obtained after each new trial as evidence, the likelihood of a given hyperparameter combination with respect to the observed PSNR is calculated. Finally, new parameter combinations are tried in accordance with the updated likelihood. In our study, probability distributions were generated by determining the threshold values corresponding to the best 10% PSNR. To determine the parameters, 3000 training and 600 test data were randomly selected from the LFW and Celeba datasets. Figure 11 displays the outcomes achieved from executing the model 100 times. Each execution has 100 epochs and involves estimating the parameters. The parameter estimation intervals

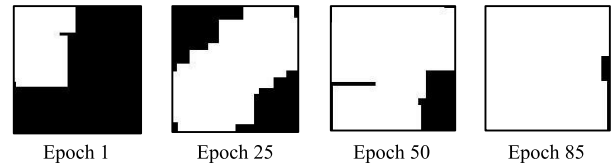


FIGURE 12. Selected patch map depending on some epochs in the enhancement phase.

are [0.1, 0.5] for σ , [1], [16] for RRDB number and [5], [20] for T . As can be seen from the graph, the best results are obtained if enough patches are optimized and the RRDB block is used. The σ value determines the trade-off between making a decision that has never been tried or taking the best action found from previous experience. The best result was obtained with $T = 20$, RRDB number 14 and $\sigma = 0.21$. However, satisfactory results were also obtained with $\sigma = 0.37$, $T = 20$ and $rrdb = 10$. The effect of RRDB and T cannot be ignored in our method. Small values of T and RRDB produce low PSNR values. With T in the range [14], [20] and RRDB in the range [6], [12], the PSNR value is very close to the best result. Figure 11 shows the concentration in these regions.

D. ABLATION STUDIES

Several ablation studies have evaluated the effects of different important components in our study.

1) EFFECT OF PATCH COUNTS

First, we investigated the effect of using different recursive steps (T) at a fixed patch size. The model was trained with four different settings ($T = [5, 15, 25, 35]$), keeping the overall number of parameters constant. Table 7 shows that the model performance improves first as T increases, but then decreases after a certain level. Since the extracted patches cannot cover the entire image, the low number of recursion steps results in low PSNR values. From the experiments, it was found that the number of recursion steps should be at least 10 in order for the patches to cover the entire image. As T increases, the increase in PSNR values is limited. $T = 15$ was found to be optimal in terms of PSNR, speed and computational load.

2) EFFECT OF GLOBAL ENHANCEMENT PARAMETER (E_{pc})

How the decision maker works with the DRL-SRFR model is visualized on the image map. As shown in Figure 12, DRL-SRFR first concentrates on specific areas of the image, and then learns to enhance the whole image and extends to every region of the image. The selected patch enhancements provide better learning of local features, helping to better extract the features of the relevant area of the person’s face. The corners of the face image are usually flat backgrounds with no personal features of interest and are easy to enhance without knowledge of specific facial characteristics. In a piecewise progressive model, the features of the facial components such

TABLE 7. Comparison of different values of T .

	LFW 4 x	LFW 8 x
T=5	28.96	23.62
T=15	34.88	28.53
T=25	35.92	29.22
T=35	34.12	28.01

PSNR comparison of different values of T with respect to 4 and 8 scale factors

as ears, eyes, nose and mouth are better recognized, resulting in a better perceptual enhancement that does not distort the person's personal features.

The parameter used to determine the map takes a value in the range (0;1) since it is calculated by the ratio of the selected local and global pixels. An increase in E_{pc} indicates a wider coverage of the map. In addition to the rewards obtained with local details, the use of the reward obtained with the increase in E_{pc} at the end of the loop also benefits the global improvement.

3) EFFECT OF PATCH SELECTION

The policy network can adaptively select a set of patches. To evaluate the impact of patch selection strategies, Figure 7 compares different methods. The comparison shows that there are significant differences between standard resolution enhancement methods and patch selection. There are two reasons for better patch selection performance with DRL. At first in standard methods, deformable facial patches (e.g., the nose or mouth shown in Figure 7 (c) and (d) may appear because the whole image or the patches used include broken facial patches (e.g., crooked nose, misaligned eyebrows, separated eyes). Secondly, patches selected using DRL have the ability to discover surface relationships between patches that are close to each other, helping to select intact surface fragments and enhancing the local restoration network, allowing credential recovery.

4) EFFECT OF PATCH SIZE

Patch selection is one of the key strategies of SR methods. The effect of the patch on the resolution increase depends on the patch size and attributes. Small patch size preserves details and increases resolution. However, a small patch size requires more computation and negatively affects the processing time. On the other hand, a large patch size reduces the processing time but makes it difficult to preserve details because a large patch size focuses on the general features of the image, giving a general idea of the image. The choice of patch size is a trade-off between detail preservation and processing time. The ideal patch size should be determined by considering the targeted resolution, available resources and application requirements. The improvement in resolution as a function of patch size for a scaling factor of 4 is given in Table 8.

TABLE 8. Patch size comparison.

Patch size	PSNR	SSIM
32 x 32	31.33	0.9025
52 x 52	34.88	0.9691
64 x 64	32.71	0.9187

PSNR and SSIM variation with patch size on LFW dataset for scaling factor 4.

TABLE 9. Comparison of different number of RRDB layers.

Number of blocks	PSNR	SSIM	Time(millisecond)
4	31.09	0.8820	0,114
8	34.45	0.9688	0,158
12	34.89	0.9710	0,213

PSNR and SSIM variation with respect to the number of RRDB blocks on LFW dataset for scaling factor 4.

5) EFFECT OF RRDB COUNTS

Different models were created by varying the number of blocks in the RRDB structure. Initially, there is a basic RRDM model with 8 blocks. By adapting the basic block, models were derived, one with 12 blocks and the other with 4 blocks. All models are compared using performance metrics and the results are presented in Table 9. Increasing the number of RRDB blocks does not affect the SR performance satisfactorily. However, it is observed that the 12-block structure provides significant PSNR and SSIM improvement compared to the 4-block structure but provides marginal improvement in the related metrics compared to the 8-block structure. It is concluded that as the number of blocks in the RRDB structure increases, the improvement in performance is limited due to the decrease in feature extraction after a certain threshold.

The findings provide an important perspective on the effectiveness of the RRDB structure in SR applications. Adding more blocks initially increases the feature representation of the model, but from a certain point the gains become negligible because the basic features are already captured by the previous layers. This observation is consistent with the principle of diminishing returns [96]. It is clear from the results that additional layers provide marginal improvements when the model is deep enough. Furthermore, using more blocks would increase the number of parameters, which may cause the model to overfit the training data and thus not generalize to new data. When ESRGAN was applied for an 8x scaling factor, an overfitting problem was observed. Another problem is that the gradients gradually grow excessively during backpropagation as the network depth increases. This problem was experienced both in the ESRGAN method and in our proposed method. Even if gradient clipping is used to overcome this problem, the gain will be very small after a certain number of blocks.

V. CONCLUSION

This study proposed an original approach called DRL-SRFR to solve the facial hallucination problem. The utilization of a DRL-based framework for patch selection in the image enhances the efficiency of evaluating the interconnections among facial components. The proposed model includes a resolution enhancement mechanism based on DRL and RRDB. A more specialized solution is obtained instead of the classical reward mechanism by establishing a connected quintuple structure. The DRL-SRFR structure includes a recurrent policy network and a local enhancement network for face patch SR. Extensive simulation results show that this approach can produce high-resolution face images that are recognizable by human visual perception and outperform the state-of-the-art face hallucination methods in terms of PSNR and SSIM. The proposed method can potentially improve training efficiency by using lighter network structures. This strategy minimizes the computational resources required and adjusts the model to suit a broader range of applications. Additionally, integrating different attention mechanisms can provide a more detailed analysis of the relationships between facial components, significantly enhancing model performance. The findings also suggest that patch-based diffusion methods can be successful in solving the facial hallucination problem, particularly in preserving and enhancing local details. These methods can potentially create more natural and realistic high-resolution facial images by harmonizing facial components with each other.

REFERENCES

- [1] J. Jiang, C. Wang, X. Liu, and J. Ma, "Deep learning-based face super-resolution: A survey," *ACM Comput. Surveys*, vol. 55, no. 1, pp. 1–36, Nov. 2021, doi: [10.1145/3485132](https://doi.org/10.1145/3485132).
- [2] S. Kanakaraj, V. K. Govindan, and S. Kalady, "Face super resolution: A survey," *Int. J. Image, Graph. Signal Process.*, vol. 9, no. 5, pp. 54–67, May 2017, doi: [10.5815/ijgisp.2017.05.06](https://doi.org/10.5815/ijgisp.2017.05.06).
- [3] A. Sharma, B. P. Srivastava, and P. N. Shankar, "Facial image super-resolution with CNN 'A review,'" in *Proc. IEEE Int. Students' Conf. Electr., Electron. Comput. Sci. (SCEECS)*, Feb. 2023, pp. 1–6, doi: [10.1109/SCEECS57921.2023.10063110](https://doi.org/10.1109/SCEECS57921.2023.10063110).
- [4] A. J. Shah and S. B. Gupta, "Image super resolution—A survey," in *Proc. 1st Int. Conf. Emerg. Technol. Trends Electron., Commun. Netw.*, Dec. 2012, pp. 1–6, doi: [10.1109/ETTECN.2012.6470098](https://doi.org/10.1109/ETTECN.2012.6470098).
- [5] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1167–1183, Sep. 2002, doi: [10.1109/TPAMI.2002.1033210](https://doi.org/10.1109/TPAMI.2002.1033210).
- [6] T. Lu, X. Hao, Y. Zhang, K. Liu, and Z. Xiong, "Parallel region-based deep residual networks for face hallucination," *IEEE Access*, vol. 7, pp. 81266–81278, 2019, doi: [10.1109/ACCESS.2019.2923023](https://doi.org/10.1109/ACCESS.2019.2923023).
- [7] C. Liu, H.-Y. Shum, and C.-S. Zhang, "A two-step approach to hallucinating faces: Global parametric model and local nonparametric model," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2001, pp. 192–198, doi: [10.1109/CVPR.2001.990475](https://doi.org/10.1109/CVPR.2001.990475).
- [8] X. Wang and X. Tang, "Hallucinating face by eigentransformation," *IEEE Trans. Syst. Man Cybern., C, Appl. Rev.*, vol. 35, no. 3, pp. 425–434, Aug. 2005, doi: [10.1109/TSMCC.2005.848171](https://doi.org/10.1109/TSMCC.2005.848171).
- [9] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000, doi: [10.1126/science.290.5500.2323](https://doi.org/10.1126/science.290.5500.2323).
- [10] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2004, pp. 275–282, doi: [10.1109/CVPR.2004.1315043](https://doi.org/10.1109/CVPR.2004.1315043).
- [11] Y. Zhuang, J. Zhang, and F. Wu, "Hallucinating faces: LPH super-resolution and neighbor reconstruction for residue compensation," *Pattern Recognit.*, vol. 40, no. 11, pp. 3178–3194, Nov. 2007, doi: [10.1016/j.patcog.2007.03.011](https://doi.org/10.1016/j.patcog.2007.03.011).
- [12] W. Liu, D. Lin, and X. Tang, "Neighbor combination and transformation for hallucinating faces," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2005, p. 4, doi: [10.1109/ICME.2005.1521381](https://doi.org/10.1109/ICME.2005.1521381).
- [13] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8, doi: [10.1109/CVPR.2008.4587659](https://doi.org/10.1109/CVPR.2008.4587659).
- [14] S. Dai, M. Han, W. Xu, Y. Wu, Y. Gong, and A. K. Katsaggelos, "Soft-Cuts: A soft edge smoothness prior for color image super-resolution," *IEEE Trans. Image Process.*, vol. 18, no. 5, pp. 969–981, May 2009, doi: [10.1109/TIP.2009.2012908](https://doi.org/10.1109/TIP.2009.2012908).
- [15] Q. Yan, Y. Xu, X. Yang, and T. Q. Nguyen, "Single image superresolution based on gradient profile sharpness," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3187–3202, Oct. 2015, doi: [10.1109/TIP.2015.2414877](https://doi.org/10.1109/TIP.2015.2414877).
- [16] A. Marquina and S. J. Osher, "Image super-resolution by TV-regularization and Bregman iteration," *J. Sci. Comput.*, vol. 37, no. 3, pp. 367–382, Dec. 2008, doi: [10.1007/s10915-008-9214-8](https://doi.org/10.1007/s10915-008-9214-8).
- [17] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2005, pp. 60–65, doi: [10.1109/CVPR.2005.38](https://doi.org/10.1109/CVPR.2005.38).
- [18] R. C. Hardie, K. J. Barnard, and E. E. Armstrong, "Joint MAP registration and high-resolution image estimation using a sequence of undersampled images," *IEEE Trans. Image Process.*, vol. 6, no. 12, pp. 1621–1633, Dec. 1997, doi: [10.1109/83.650116](https://doi.org/10.1109/83.650116).
- [19] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1327–1344, Oct. 2004, doi: [10.1109/TIP.2004.834669](https://doi.org/10.1109/TIP.2004.834669).
- [20] M. Irani and S. Peleg, "Motion analysis for image enhancement: Resolution, occlusion, and transparency," *J. Vis. Commun. Image Represent.*, vol. 4, no. 4, pp. 324–335, Dec. 1993, doi: [10.1006/jvci.1993.1030](https://doi.org/10.1006/jvci.1993.1030).
- [21] Z. Lin and H.-Y. Shum, "Fundamental limits of reconstruction-based superresolution algorithms under local translation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 83–97, Jan. 2004, doi: [10.1109/TPAMI.2004.1261081](https://doi.org/10.1109/TPAMI.2004.1261081).
- [22] H. Huang and H. He, "Super-resolution method for face recognition using nonlinear mappings on coherent features," *IEEE Trans. Neural Netw.*, vol. 22, no. 1, pp. 121–130, Jan. 2011, doi: [10.1109/TNN.2010.2089470](https://doi.org/10.1109/TNN.2010.2089470).
- [23] K. Nguyen, S. Sridharan, S. Denman, and C. Fookes, "Feature-domain super-resolution framework for Gabor-based face and iris recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2642–2649, doi: [10.1109/CVPR.2012.6247984](https://doi.org/10.1109/CVPR.2012.6247984).
- [24] L. Wu and X. Wang, "A fast algorithm for learning-based super-resolution reconstruction of face image," in *Proc. 4th Int. Congr. Image Signal Process.*, vol. 2, Oct. 2011, pp. 1049–1053, doi: [10.1109/CISP.2011.6100303](https://doi.org/10.1109/CISP.2011.6100303).
- [25] W. W. Zou and P. C. Yuen, "Learning the relationship between high and low resolution images in kernel space for face super resolution," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 1152–1155, doi: [10.1109/ICPR.2010.288](https://doi.org/10.1109/ICPR.2010.288).
- [26] Y. He, K.-H. Yap, and L.-P. Chau, "A learning approach for single-frame face super-resolution," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2009, pp. 770–773, doi: [10.1109/ISCAS.2009.5117862](https://doi.org/10.1109/ISCAS.2009.5117862).
- [27] Y. Kong, S. Zhang, and P. Cheng, "Super-resolution reconstruction face recognition based on multi-level FFD registration," *Optik*, vol. 124, no. 24, pp. 6926–6931, Dec. 2013, doi: [10.1016/j.ijleo.2013.05.175](https://doi.org/10.1016/j.ijleo.2013.05.175).
- [28] T. Lu, H. Wang, Z. Xiong, J. Jiang, Y. Zhang, H. Zhou, and Z. Wang, "Face hallucination using region-based deep convolutional networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1657–1661, doi: [10.1109/ICIP.2017.8296563](https://doi.org/10.1109/ICIP.2017.8296563).
- [29] T. Lu, Y. Wang, Y. Zhang, Y. Wang, L. Wei, Z. Wang, and J. Jiang, "Face hallucination via split-attention in split-attention network," in *Proc. 29th ACM Int. Conf. Multimedia*. New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 5501–5509, doi: [10.1145/3474085.3475682](https://doi.org/10.1145/3474085.3475682).
- [30] X. Cheng, J. Lu, B. Yuan, and J. Zhou, "Identity-preserving face hallucination via deep reinforcement learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4796–4809, Dec. 2020, doi: [10.1109/TCSVT.2019.2961629](https://doi.org/10.1109/TCSVT.2019.2961629).

- [31] Y. Shi, G. Li, Q. Cao, K. Wang, and L. Lin, "Face hallucination by attentive sequence optimization with reinforcement learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 11, pp. 2809–2824, Nov. 2020, doi: [10.1109/TPAMI.2019.2915301](https://doi.org/10.1109/TPAMI.2019.2915301).
- [32] Q. Cao, L. Lin, Y. Shi, X. Liang, and G. Li, "Attention-aware face hallucination via deep reinforcement learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1656–1664, doi: [10.1109/CVPR.2017.180](https://doi.org/10.1109/CVPR.2017.180).
- [33] J. Pan, D. Sun, J. Zhang, J. Tang, J. Yang, Y.-W. Tai, and M.-H. Yang, "Dual convolutional neural networks for low-level vision," *Int. J. Comput. Vis.*, vol. 130, no. 6, pp. 1440–1458, Jun. 2022, doi: [10.1007/s11263-022-01583-y](https://doi.org/10.1007/s11263-022-01583-y).
- [34] J. Dong, J. Pan, J. S. Ren, L. Lin, J. Tang, and M.-H. Yang, "Learning spatially variant linear representation models for joint filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8355–8370, Nov. 2022, doi: [10.1109/TPAMI.2021.3102575](https://doi.org/10.1109/TPAMI.2021.3102575).
- [35] J. Dong, S. Roth, and B. Schiele, "DWDN: Deep Wiener deconvolution network for non-blind image deblurring," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9960–9976, Dec. 2022, doi: [10.1109/TPAMI.2021.3138787](https://doi.org/10.1109/TPAMI.2021.3138787).
- [36] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Learning face hallucination in the wild," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2015, pp. 3871–3877, doi: [10.1609/aaai.v29i1.9795](https://doi.org/10.1609/aaai.v29i1.9795).
- [37] H. Huang, R. He, Z. Sun, and T. Tan, "Wavelet-SRNet: A wavelet-based CNN for multi-scale face super resolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1698–1706, doi: [10.1109/ICCV.2017.187](https://doi.org/10.1109/ICCV.2017.187).
- [38] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "FSRNet: End-to-end learning face super-resolution with facial priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2492–2501, doi: [10.1109/CVPR.2018.00264](https://doi.org/10.1109/CVPR.2018.00264).
- [39] C. Ma, Z. Jiang, Y. Rao, J. Lu, and J. Zhou, "Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5568–5577, doi: [10.1109/CVPR42600.2020.00561](https://doi.org/10.1109/CVPR42600.2020.00561).
- [40] C. Wang, J. Jiang, Z. Zhong, and X. Liu, "Spatial-frequency mutual learning for face super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22356–22366, doi: [10.1109/cvpr52729.2023.02141](https://doi.org/10.1109/cvpr52729.2023.02141).
- [41] H. Dastmalchi and H. Aghaeinia, "Super-resolution of very low-resolution face images with a wavelet integrated, identity preserving, adversarial network," *Signal Process., Image Commun.*, vol. 107, Sep. 2022, Art. no. 116755, doi: [10.1016/j.image.2022.116755](https://doi.org/10.1016/j.image.2022.116755).
- [42] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114, doi: [10.1109/CVPR.2017.19](https://doi.org/10.1109/CVPR.2017.19).
- [43] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481, doi: [10.1109/CVPR.2018.00262](https://doi.org/10.1109/CVPR.2018.00262).
- [44] J. Xu, Y. Chae, B. Stenger, and A. Datta, "Dense ByNet: Residual dense network for image super resolution," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 71–75, doi: [10.1109/ICIP.2018.8451696](https://doi.org/10.1109/ICIP.2018.8451696).
- [45] X. T. Wang, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Comput. Vis. ECCV Workshops*, vol. 11133, 2019, pp. 63–79, doi: [10.1007/978-3-030-11021-5_5](https://doi.org/10.1007/978-3-030-11021-5_5).
- [46] *Image Super-Resolution Algorithm Based on RRDB Model | IEEE Journals & Magazine | IEEE Xplore*. Accessed: Feb. 10, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/9562515>
- [47] Y.-Z. Chen, T.-J. Liu, and K.-H. Liu, "Super-resolution of satellite images by two-dimensional RRDB and edge-enhancement generative adversarial network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 1825–1829, doi: [10.1109/ICASSP43922.2022.9747063](https://doi.org/10.1109/ICASSP43922.2022.9747063).
- [48] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4713–4726, Apr. 2023, doi: [10.1109/TPAMI.2022.3204461](https://doi.org/10.1109/TPAMI.2022.3204461).
- [49] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: [10.1145/3422622](https://doi.org/10.1145/3422622).
- [50] J. J. Yu, K. G. Derpanis, and M. A. Brubaker, "Wavelet flow: Fast training of high resolution normalizing flows," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6184–6196. Accessed: Jul. 10, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/hash/4491777b1aa8b5b32c2e8666db1e4a95-Abstract.html
- [51] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Proc. Adv. Neural Inf. Process. Syst.*, Red Hook, NY, USA: Curran Associates, 2018. Accessed: Jul. 10, 2024. [Online]. Available: <https://proceedings.neurips.cc/paper/2018/hash/d139db6a236200b21cc7f752979132d0-Abstract.html>
- [52] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.
- [53] A. van den Oord, N. Kalchbrenner, L. Espeholt, K. Kavukcuoglu, O. Vinyals, and A. Graves, "Conditional image generation with PixelCNN decoders," in *Proc. Adv. Neural Inf. Process. Syst.*, Red Hook, NY, USA: Curran Associates, 2016, pp. 1–11. Accessed: Jul. 10, 2024. [Online]. Available: <https://proceedings.neurips.cc/paper/2016/hash/b1301141feffabac455e1f90a7de2054-Abstract.html>
- [54] M. Noroozi, I. Hadji, B. Martinez, A. Bulat, and G. Tzimiropoulos, "You only need one step: Fast super-resolution with stable diffusion via scale distillation," 2024, *arXiv:2401.17258*.
- [55] R. Corvi, D. Cozzolino, G. Poggi, K. Nagano, and L. Verdoliva, "Intriguing properties of synthetic images: From generative adversarial networks to diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 973–982. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023W/WMF/html/Corvi_Intriguing_Properties_of_Synthetic_Images_From_Generative_Adversarial_Networks_to_CVPRW_2023_paper.html
- [56] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5, doi: [10.1109/ICASSP49357.2023.10095167](https://doi.org/10.1109/ICASSP49357.2023.10095167).
- [57] Y. Benny, T. Galanti, S. Benaim, and L. Wolf, "Evaluation metrics for conditional image generation," *Int. J. Comput. Vis.*, vol. 129, no. 5, pp. 1712–1731, May 2021, doi: [10.1007/s11263-020-01424-w](https://doi.org/10.1007/s11263-020-01424-w).
- [58] M. Hassanin, S. Anwar, I. Radwan, F. S. Khan, and A. Mian, "Visual attention methods in deep learning: An in-depth survey," *Inf. Fusion*, vol. 108, Aug. 2024, Art. no. 102417, doi: [10.1016/j.inffus.2024.102417](https://doi.org/10.1016/j.inffus.2024.102417).
- [59] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [60] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, "Transformer for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Sep. 2022, pp. 457–466. Accessed: Jul. 10, 2024. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022W/NTIRE/html/Lu_Transformer_for_Single_Image_Super-Resolution_CVPRW_2022_paper.html
- [61] W. You, S. Sun, and M. Iyyer, "Hard-coded Gaussian attention for neural machine translation," 2020, *arXiv:2005.00742*.
- [62] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: Face recognition with very deep neural networks," 2015, *arXiv:1502.00873*.
- [63] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 464–472, doi: [10.1109/ICCV.2017.58](https://doi.org/10.1109/ICCV.2017.58).
- [64] T. Chen, Z. Wang, G. Li, and L. Lin, "Recurrent attentional reinforcement learning for multi-label image recognition," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, pp. 6730–6737, doi: [10.1609/aaai.v32i1.12281](https://doi.org/10.1609/aaai.v32i1.12281).
- [65] G. Li, Y. Gan, H. Wu, N. Xiao, and L. Lin, "Cross-modal attentional context learning for RGB-D object detection," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1591–1601, Apr. 2019, doi: [10.1109/TIP.2018.2878956](https://doi.org/10.1109/TIP.2018.2878956).
- [66] 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, Feb. 2016, doi: [10.1109/TPAMI.2015.2439281](https://doi.org/10.1109/TPAMI.2015.2439281).
- [67] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1132–1140, doi: [10.1109/CVPRW.2017.151](https://doi.org/10.1109/CVPRW.2017.151).

- [68] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham, Switzerland: Springer, 2016, pp. 391–407, doi: [10.1007/978-3-319-46475-6_25](https://doi.org/10.1007/978-3-319-46475-6_25).
- [69] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2790–2798, doi: [10.1109/CVPR.2017.298](https://doi.org/10.1109/CVPR.2017.298).
- [70] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654, doi: [10.1109/CVPR.2016.182](https://doi.org/10.1109/CVPR.2016.182).
- [71] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017. Accessed: Jul. 10, 2024. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Lai_Deep_Laplacian_Pyramid_CVPR_2017_paper.html
- [72] Y. Tai, J. Yang, X. Liu, and C. Xu, "MemNet: A persistent memory network for image restoration," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4539–4547. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2017/html/Tai_MemNet_A_Persistent_ICCV_2017_paper.html
- [73] Y. R. Musunuri and O.-S. Kwon, "Deep residual dense network for single image super-resolution," *Electronics*, vol. 10, no. 5, p. 555, Feb. 2021, doi: [10.3390/electronics10050555](https://doi.org/10.3390/electronics10050555).
- [74] M. R. Ibrahim, R. Benavente, F. Lumbreras, and D. Ponsa, "3DRRDB: Super resolution of multiple remote sensing images using 3D residual in residual dense blocks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 322–331, doi: [10.1109/CVPRW56347.2022.00047](https://doi.org/10.1109/CVPRW56347.2022.00047).
- [75] P. Sharma, S. Coleman, P. Yogarajah, L. Taggart, and P. Samarasinghe, "Comparative analysis of super-resolution reconstructed images for micro-expression recognition," *Adv. Comput. Intell.*, vol. 2, no. 3, p. 24, May 2022, doi: [10.1007/s43674-022-00035-x](https://doi.org/10.1007/s43674-022-00035-x).
- [76] K. P. Gunasekaran, "Ultra sharp: Study of single image super resolution using residual dense network," in *Proc. IEEE 3rd Int. Conf. Comput. Commun. Artif. Intell. (CCAI)*, May 2020, pp. 261–266, doi: [10.13140/RG.2.2.25001.06246](https://doi.org/10.13140/RG.2.2.25001.06246).
- [77] T. N. Thanh, "Deep learning based approach implemented to image super-resolution," *J. Adv. Inf. Technol.*, vol. 11, no. 4, pp. 209–216, Nov. 2020. Accessed: Feb. 10, 2024. [Online]. Available: https://www.academia.edu/72768607/Deep_Learning_Based_Approach_Implemented_to_Image_Super_Resolution
- [78] X. Wang, K. Yu, C. Dong, and C. Change Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018. Accessed: Jul. 10, 2024. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Wang_Recovering_Realistic_Texture_CVPR_2018_paper.html
- [79] C. Tian, X. Zhang, J. C.-W. Lin, W. Zuo, Y. Zhang, and C.-W. Lin, "Generative adversarial networks for image super-resolution: A survey," 2022, *arXiv:2204.13620*.
- [80] Y. Liu, Q. Li, Q. Deng, Z. Sun, and M.-H. Yang, "GAN-based facial attribute manipulation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 14590–14610, Dec. 2023, doi: [10.1109/TPAMI.2023.3298868](https://doi.org/10.1109/TPAMI.2023.3298868).
- [81] A. Li, G. Li, L. Sun, and X. Wang, "FaceFormer: Scale-aware blind face restoration with transformers," 2022, *arXiv:2207.09790*.
- [82] B. Guo, X. Zhang, H. Wu, Y. Wang, Y. Zhang, and Y.-F. Wang, "LAR-SR: A local autoregressive model for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1899–1908, doi: [10.1109/cvpr52688.2022.00195](https://doi.org/10.1109/cvpr52688.2022.00195).
- [83] A. Lugmayr, M. Danelljan, L. Van Gool, and R. Timofte, "SRFlow: Learning the super-resolution space with normalizing flow," *Proc. Eur. Conf. Comput. Vis.*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Cham, Switzerland: Springer, 2020, pp. 715–732, doi: [10.1007/978-3-030-58558-7_42](https://doi.org/10.1007/978-3-030-58558-7_42).
- [84] Z. Lin, P. Garg, A. Banerjee, S. A. Magid, D. Sun, Y. Zhang, L. Van Gool, D. Wei, and H. Pfister, "Revisiting RCAN: Improved training for image super-resolution," 2022, *arXiv:2201.11279*.
- [85] C. Chen, D. Gong, H. Wang, Z. Li, and K. K. Wong, "Learning spatial attention for face super-resolution," *IEEE Trans. Image Process.*, vol. 30, pp. 1219–1231, 2021, doi: [10.1109/TIP.2020.3043093](https://doi.org/10.1109/TIP.2020.3043093).
- [86] K. Zeng, Z. Wang, T. Lu, J. Chen, J. Wang, and Z. Xiong, "Self-attention learning network for face super-resolution," *Neural Netw.*, vol. 160, pp. 164–174, Mar. 2023, doi: [10.1016/j.neunet.2023.01.006](https://doi.org/10.1016/j.neunet.2023.01.006).
- [87] S. Gao and X. Zhuang, "Bayesian image super-resolution with deep modeling of image statistics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1405–1423, Feb. 2023, doi: [10.1109/TPAMI.2022.3163307](https://doi.org/10.1109/TPAMI.2022.3163307).
- [88] T. Ru and Z. Zhu, "Deep clustering efficient learning network for motion recognition based on self-attention mechanism," *Appl. Sci.*, vol. 13, no. 5, p. 2996, Feb. 2023, doi: [10.3390/app13052996](https://doi.org/10.3390/app13052996).
- [89] R. A. Howard, *Dynamic Programming and Markov Processes*. in *Dynamic Programming and Markov Processes*. Oxford, U.K.: Wiley, 1960, p. 136.
- [90] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.
- [91] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 1833–1844. Accessed: Jul. 10, 2024. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021W/AIM/html/Liang_SwinIR_Image_Restoration_Using_Swin_Transformer_ICCVW_2021_paper.html
- [92] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738. Accessed: Jul. 10, 2024. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2015/html/Liu_Deep_Learning_Face_ICCV_2015_paper.html
- [93] O. Jesorsky, K. J. Kirchberg, and R. W. Frischholz, "Robust face detection using the Hausdorff distance," in *Audio- and Video-Based Biometric Person Authentication* (Lecture Notes in Computer Science), J. Bigun and F. Smeraldi, Eds., Berlin, Germany: Springer, 2001, pp. 90–95, doi: [10.1007/3-540-45344-X_14](https://doi.org/10.1007/3-540-45344-X_14).
- [94] G. B. Huang, V. Jain, and E. Learned-Miller, "Unsupervised joint alignment of complex images," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8, doi: [10.1109/ICCV.2007.4408858](https://doi.org/10.1109/ICCV.2007.4408858).
- [95] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 365–372, doi: [10.1109/ICCV.2009.5459250](https://doi.org/10.1109/ICCV.2009.5459250).
- [96] E. Guha and V. Lakshman, "On the diminishing returns of width for continual learning," 2024, *arXiv:2403.06398*.



EMRE ALTINKAYA received the B.S. and M.S. degrees from the Electrical and Electronics Engineering Faculty, Sakarya University, Sakarya, Türkiye, in 2016 and 2018, respectively. He is currently pursuing the Ph.D. degree. In 2019, he was a Lecturer in hybrid and electric vehicles technology program with the Vocational School, Bilecik Şeyh Edebali University, and continues to work in this department. His primary research interests include image processing, artificial intelligence, and electric vehicles.



BURHAN BARAKLI (Member, IEEE) received the B.S. degree in electrical and electronic engineering and the M.S. degree in semiconductors from Sakarya University, Sakarya, Türkiye, in 2005 and 2007, respectively, and the Ph.D. degree in electronics from the Institute of Science, Sakarya University, in 2014, with a dissertation on reversible video watermarking. From 2005 to 2008, he was a Control Engineer with Vur-Kontrol Ltd., Istanbul, Türkiye. Since 2015, he has been an Assistant Professor with the Department of Electrical and Electronic Engineering, Sakarya University. He is also a founder of Bayt ARGE Ltd. Company, Sakarya, in 2014. At Bayt ARGE, he developed products focused on visual product tracking and automatic reading systems, until 2020. His contributions to these fields are primarily applied to developing sophisticated algorithms for embedded control systems in industrial applications. His research interests include machine learning, image and video processing, deep learning, control theory, and embedded systems.