Iranian Journal of Fuzzy Systems

Volume 23, Number 4, (2026), pp. 55-67

Original paper

University of Sistan and Baluchestan

FAUNet: A fuzzy-attention U-Net for diffusion-based Persian text image super-resolution

M. Kushki [1], E. Rashedi [2], E. Shabaninia [3] and M. Kamandar [4]

[1,2,4] *Department of Telecommunications and Electronics Engineering, Faculty of Electrical and Computer Engineering, Graduate University of Advanced Technology, Kerman, Iran*

[3] *Department of Applied mathematics, Faculty of Modern Sciences and Technologies, Graduate University of Advanced Technology, Kerman, Iran*

m.koushki@student.kgut.ac.ir, e.rashedi@kgut.ac.ir, e.shabaninia@kgut.ac.ir, m.kamandar@kgut.ac.ir

Abstract

The accurate enhancement of text images is a critical challenge in computer vision, particularly for languages such as Persian that exhibit complex writing structures, cursive connections, and fine-grained diacritical marks. Traditional super-resolution approaches often fail to preserve these delicate textual details. Here, diffusion model is adopted for text image super-resolution. The U-Net framework of this method is enhanced by incorporating fuzzy logic and attention mechanism (named FAUNet) to address mentioned problems. At the bottleneck of the network, a fuzzy layer is employed to softly model uncertainties and boundary variations, while a spatial channel attention block adaptively emphasizes crucial regions of the image. Together, these components strengthen the network's capacity to capture structural dependencies and semantic details essential for text clarity. The proposed model is rigorously evaluated on two large-scale Persian text datasets: IR-LPR that comprising vehicle license plate images, and IDPL-PFOD2 that is a dataset of printed Persian text. Experimental results show that FAUNet outperforms state-of-the-art methods achieving improvements in PSNR, SSIM, and MS-SSIM metrics. These improvements not only contribute to higher visual quality but also hold strong potential for downstream applications such as optical character recognition (OCR), license plate recognition, and digital document restoration in low-quality imaging conditions.

Keywords: Super-resolution, deep learning, diffusion model, fuzzy layer, attention mechanism.

1 Introduction

With the growing demand for accurate and reliable extraction of textual information in intelligent systems, text detection and recognition in images have become essential tasks in computer vision. These technologies are widely applied in domains such as vehicle license plate recognition [2], road signs recognition [10], and improving the accuracy of optical character recognition (OCR) in scanned documents [1].

Super-resolution (SR) is an advanced image reconstruction technique that transforms low-resolution (LR) images into high-resolution (HR) counterparts. Enhancing image resolution restores the fine details of character dots and connections in Persian script and sharpens character boundaries. This process reduces frequent recognition errors between visually similar letters, such as "ب" vs. "ن" or "ح" vs. "خ", thereby lowering the overall error rate. At the same time, human readers can interpret text more quickly and with greater accuracy.

Early attempts at image super-resolution using deep learning primarily employed convolutional neural networks to learn mappings from LR to HR images [7]. However, these approaches often struggle to preserve fine texture details,

Corresponding Author: E. Rashedi
Received: October 2025; Revised: February 2026; Accepted: May 2026.
https://doi.org/10.22111/ijfs.2026.53720.9515

especially under high magnification. To address this, Generative Adversarial Networks (GANs), such as SRGAN [22], were introduced, achieving sharper reconstructions. More recently, Denoising Diffusion Probabilistic Models (DDPMs) [14] have gained considerable attention, frequently outperforming GAN-based approaches. The U-Net architecture [33] serves as a fundamental backbone in DDPMs. In these models, the forward process involves gradually corrupting the input data by adding noise across multiple time steps, while the reverse process iteratively removes noise to reconstruct the original data.

For instance, it has been successfully applied to facial image correction and to removing transparency artifacts in documents [9]. The FDM model further demonstrated that incorporating fuzzy logic improves the reconstruction of document text [42]. Moreover, fuzzy systems have proven effective in other complex tasks such as adaptive control of nonlinear networked systems with packet dropouts [17], tracking under DoS attacks [18], and consensus control in multi-agent fractional-order systems [34]. Although these works lie outside the domain of image processing, they highlight the broader effectiveness of fuzzy reasoning in handling uncertainty and structural ambiguity, supporting its integration in vision-based applications such as text image SR.

In this paper, FAUNet is introduced that is a Fuzzy-Attention U-Net which improves reconstruction accuracy in diffusion-based SR. This approach augments the U-Net architecture with both a fuzzy block and an attention mechanism, enabling more effective feature extraction and precise image reconstruction. At the U-Net bottleneck, where data is represented in compressed form, the combination of fuzzy logic and attention allows the model to better identify and amplify salient features while also capturing long-range spatial and channel-wise dependencies. This integration enhances the model's ability to reconstruct structural and contextual details, yielding more faithful reconstructions of text images.

Incorporating fuzzy layers into U-Net diffusion models also resolves ambiguities at character boundaries where clarity is essential for distinguishing visually similar letters. We evaluate the effectiveness of the proposed method through extensive experiments and compare the results with state-of-the-art approaches. So, the main contributions of this work are as follows:

- Proposing FAUNet, a U-shaped architecture that integrates a fuzzy block and an attention mechanism into the bottleneck.
- The fuzzy block models the ambiguities and boundary variations, while the attention block focuses on salient features with high precision.
- Extensive experiments demonstrate that FAUNet outperforms baseline U-Net and other benchmark methods, achieving superior reconstruction of fine details and complex structures.

The structure of the paper is as follows. Section 2 provides a review of related research on SR. Section 3 details the proposed methodology. Section 4 presents a comprehensive evaluation of the experimental results from both quantitative and qualitative perspectives. Finally, Section 5 offers further discussions and concludes the paper.

2 Related works

Deep-learning-based methods have become the standard approach for image SR [39]. Early methods focused on convolutional neural networks (CNNs), which learn a direct end-to-end mapping from LR to HR images. Examples include FSRCNN [8] and ESPCNN [35], which improved speed and reconstruction quality over traditional interpolation. However, CNN-based methods often struggle to preserve fine-grained texture and structural details, especially under high magnification.

To address these shortcomings, GANs were introduced, such as SRGAN [22] which use an adversarial loss to generate sharper, more realistic textures. A GAN consists of a generator and a discriminator trained in competition: the generator aims to fool the discriminator by producing realistic HR images, while the discriminator tries to distinguish real from generated samples [12]. While GAN-based methods significantly improve perceptual quality, they often suffer from instability during training and may hallucinate unrealistic details [11].

Diffusion Models (DMs) [28] have recently emerged as a promising alternative to GANs, offering stable training and high-quality outputs. In these models, a forward process gradually adds Gaussian noise to data, and a learned reverse process denoises the image step by step. SRDiff [23] was the first model to apply diffusion to single-image SR, demonstrating the generation of diverse and realistic outputs. Following this, models like DiT-SR [6] and LFSRDiff [5] extended the framework to improve efficiency and address domain-specific needs, such as light field SR. YODA [27] proposed a dynamic diffusion approach that applies denoising selectively to detail-rich regions, increasing efficiency.

Other works such as Stable Diffusion [32] and latent-wise diffusion models [25, 43] perform diffusion in latent space, significantly reducing computational costs while preserving generation quality.

While diffusion models achieve impressive results, they may still struggle with structural ambiguities, particularly in fine-detailed or degraded regions. Recent research has explored integrating fuzzy logic into diffusion models to better handle uncertainty. For instance, the fuzzy-conditioned diffusion framework [9] enables per-pixel control over the diffusion prior, allowing highly precise and interpretable corrections in facial restoration. Similarly, the Fuzzy Diffusion Model (FDM) [42] applies fuzzy reasoning during the reverse diffusion process to eliminate bleed-through in document images, demonstrating superior visual clarity and robustness.

These studies highlight the potential of fuzzy systems to model uncertainty and improve reconstruction in ambiguous image regions a characteristic highly relevant to cursive text images such as Persian.

Despite the significant progress in super-resolution research, studies specifically targeting Persian text or license plate images remain scarce, and most existing works rely on classical or multi-frame reconstruction techniques rather than recent deep-learning-based approaches. For instance, prior studies such as [19], "Super-resolution of license plates using frames of low-resolution video" [26], and "Multi-frame super-resolution for improving vehicle license plate recognition" [37] have primarily used non-deep methods and evaluated on small, non-public datasets. Therefore, the application of diffusion-based SR and fuzzy-enhanced architectures to Persian text remains an open and underexplored research area which this work aims to address. Table 1 presents a summary of the reviewed studies.

Despite the progress in SR and diffusion models, no prior work has specifically addressed the reconstruction of Persian text using fuzzy-enhanced diffusion. Our proposed method, FAUNet, builds upon the DDPM framework and integrates both fuzzy logic and attention mechanisms into a U-Net backbone. Unlike prior fuzzy diffusion models focused on faces or scanned documents, FAUNet targets structural ambiguities unique to cursive Persian script, such as connected strokes, diacritical marks, and indistinct boundaries. This integration enables more precise restoration of fine textual details and improves both perceptual quality and recognition accuracy.

Table 1: Overview of the related studies addressed in this work.

Method	Application Domain	Model	Attention	Fuzzy	Year
FSRCNN [8]	Generic Image SR	CNN	×	×	2016
ESPCNN [35]	Generic Image SR	CNN	×	×	2016
SRGAN [22]	Image SR	GAN	×	×	2017
Real-ESRGAN [41]	Real-world SR	GAN	×	×	2021
SRDiff [23]	Single-image SR	Diffusion	×	×	2022
FDM [42]	Document clean up	Diffusion + fuzzy	×	✓	2024
DiT-SR [6]	Multi-scale SR	Diffusion + Transformer	✓	×	2025
LFSRDiff [5]	Light Field SR	Diffusion + UNet	×	×	2025
YODA [27]	Text SR	Diffusion + Attention	✓	×	2025

3 Methodology

3.1 Super resolution based on diffusion models

The overall structure of the model for training and inference is similar to [21]. A diffusion model is a generative method for data synthesis that gradually reconstructs data using a Markov chain. Sohl–Dickstein and colleagues introduced diffusion probabilistic models that reconstruct data by learning to add noise and reverse a gradual, multi-step process [36]. It is a Markov chain trained via variational inference to generate samples that match the data distribution. A diffusion model comprises two processes: the forward process and the backward process [23] as illustrated in Figure 1.

Forward process: In the forward process, starting from the initial low-resolution image y_0 that has been upscaled using bicubic interpolation, Gaussian noise with variance β_t is gradually added to each input image over multiple steps. The noise intensity varies at each step, resulting in different levels of corruption being applied throughout the process. During this process, the training images are progressively corrupted, and at step t, y_t is generated by adding noise to y_{t-1} and follows the formula below eventually becoming pure Gaussian noise [28].

$$q(y_t|y_{t-1}) = \mathcal{N}\left(y_t; \mu_t = \sqrt{1-\beta_t}y_{t-1}, \beta_t I\right). \tag{1}$$

Backward process: In the backward process, images are gradually reconstructed from Gaussian noise. At each step, a neural network typically based on a U-Net [33] estimates and removes noise while preserving image dimensions

[39]. We are interested in learning a parametric approximation through a stochastic iterative refinement process that maps a source image x to a target image y. The process reverses the diffusion steps to recover the original data distribution and can be modeled as p_θ using a convolutional neural network as defined by the formula below.

$$p_\theta(y_{t-1} \mid y_t, x) = \mathcal{N}(y_{t-1}; \mu_\theta(y_t, x, t), \beta_t I). \tag{2}$$

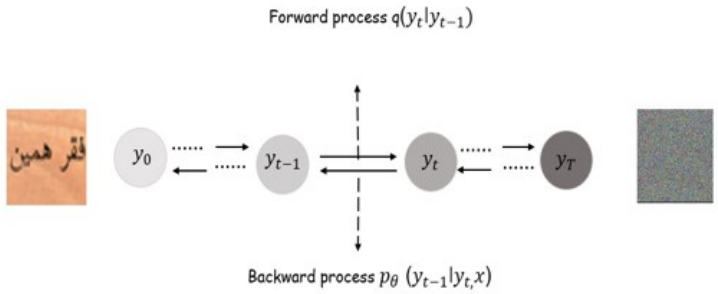

Figure 1: The DDPM diffusion model, which gradually adds noise to images in the forward process and gradually removes it in the backward process.

For the generative model to reconstruct the high-resolution image, it must have access to information from the original image. Therefore, at each time step t, the reverse process is conditioned based on the initial value of x that has been upscaled using bicubic interpolation.

In this paper, inspired by the SR3 [33] framework which is based on conditional DDPM [14], we propose a super-resolution model based on a FAUNet architecture. The U-shaped design facilitates multi-scale feature extraction, while the incorporation of fuzzy logic and attention mechanisms enhances the diffusion model's ability to accurately reconstruct fine structural details in Persian text images.

3.2 FAUNet: A fuzzy-attention U-Net

To date, numerous studies on diffusion models have been primarily implemented using the U-Net architecture [33]. The overall architecture of the proposed U-Net consists of a down-sampling path (encoder), a bottleneck strengthened with ResNet–attention and fuzzy modules, and an up-sampling path (decoder). The overall architecture of the proposed model is illustrated in Figure 2.

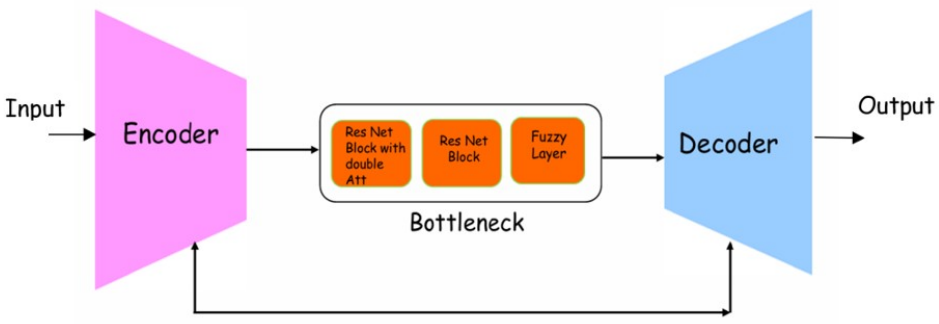

Figure 2: The overall architecture of the proposed U-Net model, consisting of a down-sampling path, a bottleneck with ResNet-attention blocks and a fuzzy layer, and an up-sampling path.

Contracting path

In the U-Net architecture, the down sampling path is constructed using ResNet Blocks [4]. In the experimental results, each block integrates a 3×3 convolution, a residual connection, and the Swish activation function, which are applied at designated resolution levels to enhance feature representation and support hierarchical learning. At the end of the

contracting path, the smallest spatial resolution with the greatest channel depth is achieved, ready to be passed to the bottleneck stage.

Bottleneck – double attention + fuzzy layer

To enhance the model's capability at the bottleneck stage, the proposed U-Net employs two ResNet blocks with channel–spatial attention, enabling the network to focus on key features. In addition, a fuzzy logic layer is incorporated to softly model uncertainties and boundary variations.

Attention mechanisms allow a model to selectively prioritize and adjust the importance of features, enabling it to focus on the most relevant parts of the input. This process allows the model to process critical information using limited computational resources [30]. By concentrating on different regions of the input at each step, the model can learn complex dependencies among features and improve the quality, accuracy, and interpretability of its outputs [15].

Through this mechanism, the model can focus on key features, assign flexible weights to different input components, and provide a more precise interpretation of the text [13]. After completing the down sampling path and reaching the smallest spatial resolution, the output of this stage is used as the input to the attention module. To prepare this data, three separate feature maps are generated for the Query (Q), Key (K), and Value (V) components. These maps are obtained by applying linear projections to the resulting feature map, ensuring that the channel dimensions and spatial arrangement are aligned with the requirements of the attention mechanism. The Q map represents the query vectors, the K map provides the comparison keys, and the V map contains the value vectors that will be weighted in the output. This process ensures that the features extracted from the previous stages are transformed into a format that allows the attention mechanism to effectively model intra-feature and inter-feature relationships, thereby enhancing image sharpness and preserving fine structural details.

In the attention mechanism [29], the cosine similarity between the Query (Q) and the Key (K) is first computed to generate the attention map. This map is then normalized and applied to weight and integrate the elements of the input sequence, yielding an attention-based representation as formulated in Eq. (3):

$$F_{attr} = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d}}\right). \tag{3}$$

Here, Q, k, and v represent the Query, Key, and Value, respectively, while d acts as the scaling factor. Feeding this output into subsequent processing stages enables the model to more effectively extract task-relevant information from the input data, thereby improving its overall efficiency and performance. The self-attention mechanism, by evaluating global relationships among features, ensures coherence in the generation process while enhancing the semantic depth and consistency of the output.

Attention mechanisms play a critical role in enabling the model to focus on the key pixels of Persian letters, reducing noise, and enhancing important details to improve image quality. Building upon this, to further increase accuracy and flexibility in feature processing and to better model uncertainties, a *fuzzy logic layer* has been integrated into the *bottleneck* of the U-Net architecture. Instead of producing simple scalar values, this layer generates a weighted distribution of fuzzy numbers, significantly enhancing the clarity and stability of Persian text images.

The logical fuzzy layer is placed at the U-Net's bottleneck and, for each spatial location in the feature map, generates a weighted distribution of "fuzzy numbers" instead of a single scalar value. This fuzzy design at the bottleneck enhances reconstruction quality, particularly in preserving fine details.

Following the ResNet attention module, the output feature map F_{attr} is processed by a fuzzy logic layer to model uncertainty and local variability in the attended features. Instead of keeping a single deterministic value per spatial location, the fuzzy layer assigns degrees of membership to multiple possible states, allowing the network to retain subtle details such as Persian letter edges, dots, and diacritics. This operation produces a fuzzy-enhanced feature map that is normalized before entering the U-Net upsampling path, ensuring stable feature distributions and sharper reconstructions.

In the fuzzy layer, the features are fuzzified, and for this purpose, the proposed model includes a layer with the architecture illustrated in Figure 3.

In the first step, a convolution is applied to perform local feature mixing and slight smoothing on the feature maps. In the second step, to enable parallel computation of membership degrees across all fuzzy sets, the input is expanded along a new fuzzy dimension. In this process, the data are reshaped so that each value is repeated n times, corresponding to the n predefined Gaussian membership functions. This mechanism allows simultaneous element-wise evaluation of each value with respect to all fuzzy states within a single tensor operation.

Next, Mathematically, for each spatial position (i, j) in F_{attr} the membership degree to fuzzy set A_K (μ_{A_K}) is computed using a Gaussian function with *Trainable* parameters of *Centers* C_k, and variances σ_K. Five Gaussian

membership functions define overlapping fuzzy states for each input feature value as:

$$\mu_{A_K}(F_{attr}) = \exp\left(-\frac{(F_{attr} - C_K)^2}{2\sigma_K^2}\right). \tag{4}$$

In the following step, the *weighted aggregation (defuzzification)* process is performed, where C_K is the center of the k-th fuzzy set, and σ_K controls its spread. We choose experimentally $n = 5$ fuzzy sets enough to capture nuanced variations without excessive computation. Final aggregation is the final fuzzy output is obtained by weighted aggregation and normalization over n sets:

$$F_{fuzzy}(i,j) = \frac{\sum_{k=1}^{n} w_k \mu_{A_K}(F_{attr})}{\sum_{k=1}^{n} \mu_{A_K}(F_{attr})}. \tag{5}$$

Here, w_k denotes the importance weight of each fuzzy state that are trainable parameters. After the *weighted aggregation (defuzzification)* stage, where the fuzzy outputs are combined through weighted aggregation and normalization to form the final feature map, the distribution of feature values may become unstable or non-uniform. To address this issue, a sequence consisting of Batch Normalization (BN1), followed by a convolution layer, and then a second normalization (BN2) is applied. This process enhances statistical stability, ensures uniform feature distribution, and improves spatial coherence in the output feature map.

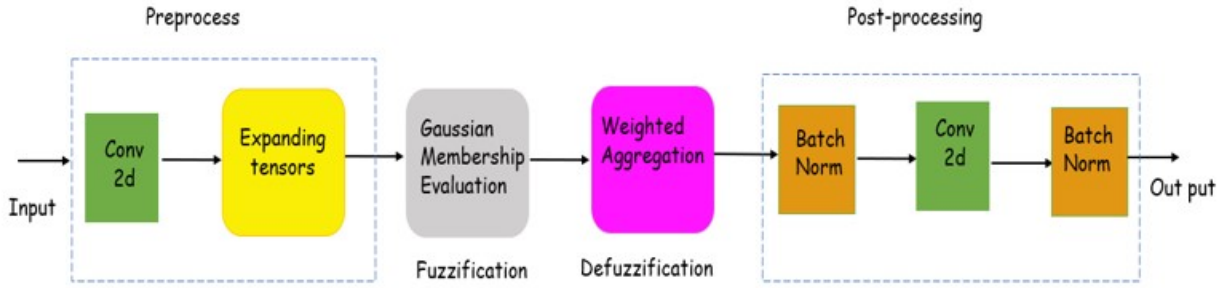

Figure 3: Overall architecture of the proposed fuzzy layer, showing the sequential process from preprocessing to the final fuzzy output.

Expanding path

The upsampling path mirrors downsampling, restoring spatial resolution via Upsample operations and skip-connected and ResNet blocks [4], producing F_{attup} at matching scales. Final convolutional mapping outputs the HR reconstruction.

In the *FAUNet* architecture, the *bottleneck* consists of ResNet with double Attention layer, followed by a *fuzzy layer* with Gaussian membership functions. This configuration enables the network to refine edge structures and preserve fine textual components throughout the diffusion steps.

4 Experiments

In this section, the dataset used, the model details, and the evaluation metrics are first described; then, the obtained results are reported and analyzed.

4.1 Datasets

In this paper, two different datasets are used to evaluate the performance of the proposed model. Each of these datasets has its own specific characteristics and challenges, and they have been selected to enable a comprehensive assessment of the model under diverse conditions.

Dataset A: To assess the performance of the proposed model, we evaluated it on real-world 8× super-resolution SR. The Iranian license-plate image dataset was released in 2022 [31]. This database contains 19,937 car images. The IR-LPR dataset was collected from 2018 to 2022 and was photographed and edited by at least 55 students from Malek Ashtar University of Technology. Most of the images were captured using mobile phone cameras. The dataset is

partitioned into day and night subsets. This dataset contains 28,600 samples of the Dummy dataset. Sample images from this dataset are shown in Figure 4.

Figure 4: Sample images from the IR-LPR dataset [31].

Dataset B: In addition, the IDPL-PFOD2 dataset was released in 2023 [3]. The IDPL-PFOD2 dataset is a large-scale dataset for printed Persian optical character recognition, containing 2,003,541 images. This dataset offers comprehensive coverage of Persian's cursive script and various writing styles, and can serve as a valuable auxiliary resource in the development of Persian OCR systems. Similar to dataset A, 19,937 samples were taken from dataset B. Figure 5 presents samples from the IDPL-PFOD2 dataset, showcasing variations in background, font, and style.

Figure 5: Sample images from the IDPL-PFOD2 dataset.

4.2 Implementation details

In this section, the model configuration and the details of its training process are examined. In accordance with the SR3 [33] model, the proposed architecture operates with a down sampling factor of $\times 8$. We trained the proposed model for 1,000,000 iterations with a batch size of 64 using a single NVIDIA Tesla V100 GPU. Adam [20] was used as the optimizer, and the learning rate was set to 3×10^{-6}. The LR images are $64 \times 64 \times 3$ in size, and the HR images are $512 \times 512 \times 3$ in size. A cosine noise schedule was used during the diffusion process.

4.3 Evaluation metrics

Image quality is a multidimensional concept that encompasses characteristics such as sharpness, contrast, and noise level. To accurately evaluate super-resolution models, we use the following three metrics:

- **Peak Signal-to-Noise Ratio (PSNR)**: This metric is based on the mean squared error (MSE) and expresses the ratio of the square of the maximum pixel value to the reconstruction error in decibels. The higher the PSNR, the closer the generated image is to the reference image [38].
- **SSIM Index**: Rather than computing simple pixel-wise differences, this metric quantifies perceptual similarity by assessing three components of the images: luminance, contrast, and structure [38].
- **MS-SSIM**: This multi-scale SSIM variant evaluates image quality at multiple resolution levels. By aggregating the results from each scale, it enhances the precision of the assessment across diverse conditions [40].

4.4 Comparison with state-of-the-arts

Quantitative Results: We evaluated our proposed model against several advanced methods. In this section, we compare our proposed method on dataset A with several state-of-the-art approaches Bicubic [16], SRGAN [22], Real-ESRGAN [41], and SwinIR [24] which we implemented ourselves and the results in Table 2.

Table 2: Quantitative performance metrics of various methods for image super-resolution on dataset A.

Method	Year	PSNR	SSIM	MS-SSIM
Bicubic	2003	21.9840	0.7250	0.8173
SRGAN	2017	29.0342	0.8020	0.8572
Real-ESRGAN	2021	24.7799	0.7443	0.8411
SwinIR	2021	20.2625	0.7075	0.8061
proposed method	2025	**31.7561**	**0.9111**	**0.9521**

In the PSNR metric, which measures the numerical similarity between the reconstructed image and the original image, the proposed model has demonstrated remarkable performance by achieving a score of 31.7561. The proposed model demonstrated notable performance by achieving a PSNR value of 31.7561. This value is 9.7721 higher than Bicubic, 2.7219 higher than SRGAN, 6.9762 higher than Real-ESRGAN, and even 11.4936 higher than SwinIR. These differences indicate more accurate reconstruction and a significant reduction in noise by the proposed model.

In the SSIM metric, which assesses the structural similarity between the reconstructed and original images in terms of luminance, contrast, and structure, the proposed model achieved a value of 0.9111, outperforming all other methods, which is higher than Bicubic, SRGAN, Real-ESRGAN, SwinIR by 0.1861, 0.1091, 0.1668, and 0.2036 units, respectively. This difference indicates more accurate reconstruction and noise reduction in the proposed model. This superiority demonstrates that the proposed model has performed remarkably well in preserving the fine structures of Persian characters and in maintaining precise textual boundaries.

In the MS-SSIM metric, which evaluates structural similarity across multiple scales, the proposed model achieved a value of 0.9521, demonstrating the highest performance among all compared methods, which is higher than Bicubic, SRGAN, Real-ESRGAN, SwinIR by 0.1348, 0.0949, 0.111, and 0.146 units, respectively.

Similarly, the evaluations on dataset B have also been conducted, and the obtained results are presented in Table 3.

Table 3: Quantitative performance metrics of various methods for image super-resolution on dataset B.

Method	Year	PSNR	SSIM	MS-SSIM
Bicubic	2003	27.4868	0.7253	0.8024
SRGAN	2017	28.4021	0.8011	0.8421
Real-ESRGAN	2021	25.1147	0.7012	0.7871
SwinIR	2021	20.8810	0.6952	0.8721
proposed method	2025	**31.1660**	**0.9021**	**0.9432**

In the PSNR metric, which measures the closeness of the reconstructed image to the original image, a value of 31.166 was achieved, which is higher than Bicubic, SRGAN, Real-ESRGAN and SwinIR by 3.6792, 2.7639, 6.0513, and 10.285 units, respectively.

In the SSIM metric, which evaluates structural similarity, a value of 0.9021 was achieved, which is higher than Bicubic, SRGAN, Real-ESRGAN, and SwinIR by 0.2068, 0.2369, 0.2309, and 0.131 units, respectively. This superiority demonstrates the model's strong capability in preserving the structure of Persian characters and maintaining precise text boundaries.

In the MS-SSIM metric, which evaluates the structural quality of the image at different scales, a value of 0.9432 was achieved, which is higher than Bicubic, SRGAN, Real-ESRGAN and SwinIR by 0.1408, 0.1011, 0.1561, and 0.0711 units, respectively.

Qualitative Results: Although quantitative metrics such as *PSNR*, *SSIM*, and *MS-SSIM* provide an accurate assessment of model performance, they do not always fully align with human visual perception. To more precisely evaluate the visual quality of images generated by the Bicubic, SRGAN, Real-ESRGAN, SwinIR and our proposed method, a qualitative evaluation was conducted. Figures 6 and 7 present the results evaluated on datasets A and B, respectively. Figure 8 shows that FAUNet effectively achieves a balance between detail sharpness and image naturalness, while maintaining high consistency with the low-resolution LR image.

Figure 6: Qualitative results of Dataset A.

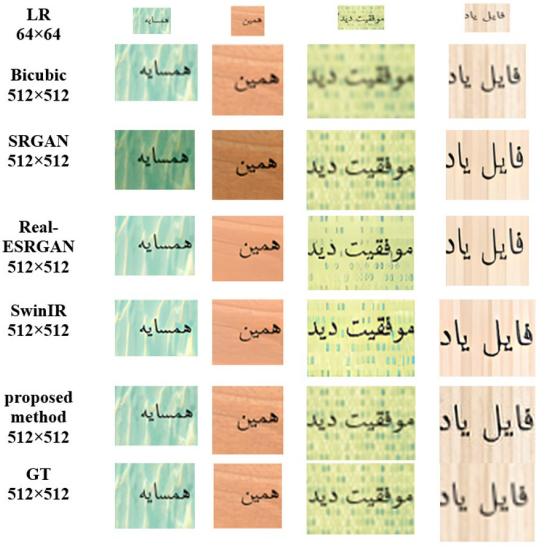

Figure 7: Qualitative results of Dataset B.

Figure 8: Visual results of FAUNet (×8). The model generates more natural and realistic textures, showing better consistency with the ground truth.

The qualitative results clearly demonstrate that the proposed FAUNet outperforms existing super-resolution models in reconstructing Persian text images, especially in terms of structural clarity and fine-detail preservation.

In contrast, FAUNet leverages *fuzzy logic* in the bottleneck and a *channel–spatial attention mechanism* to accurately reconstruct delicate features such as dots, diacritics, character links, and curves. Unlike other models that tend to over-sharpen or over-smooth the text, FAUNet maintains a visually balanced reconstruction and avoids artificial distortions. These improvements are particularly evident in noisy or low-quality Persian text images and are consistent with the quantitative gains observed in PSNR, SSIM, and MS-SSIM metrics.

Computational complexity: Table 4 presents the number of parameters and the approximate training time of IR-LPR dataset for different super-resolution models. The proposed FAUNet has a larger number of parameters compared to the baseline U-Net and other super-resolution methods such as Bicubic, SRGAN, Real-ESRGAN, and SwinIR. This increase is primarily due to the high capacity of the model architecture, while the addition of new modules such as the fuzzy layer and the channel–spatial attention block in the bottleneck further contributes to the expansion of the parameter space.

Although this leads to higher memory usage and computational requirements during training and inference, the increase is compensated by the improvement in reconstruction quality. In practice, the fuzzy layer introduces only a limited number of parameters since it employs Gaussian membership functions, and the attention block is designed to be lightweight. Therefore, although FAUNet contains more parameters than other models, it achieves better performance in terms of PSNR, SSIM, and MS-SSIM, indicating that the increased complexity directly results in more accurate reconstruction of fine details in Persian text images.

Although the training time of the FAUNet model is longer than that of other methods, this increase is mainly due to the iterative nature of diffusion models and the incorporation of additional modules such as the fuzzy layer and the attention block in the proposed architecture, which collectively contribute to improved performance in reconstructing Persian text images.

Table 4: Number of parameters and training time for different models.

Model	#Parameters	Training time
Bicubic	0	No training required
SRGAN	16.85 M	44h
Real-ESRGAN	16.73 M	40h
Swin IR	11.8 M	24h
proposed method	625 M	110h

5 Conclusion

In this study, we introduce FAUNet an innovative model developed using Diffusion Models to enhance the resolution of Persian text images. In this architecture, the U-Net is enhanced by integrating a fuzzy logic layer to effectively model uncertainties and boundary variations, alongside an attention mechanism that selectively salient focus on image features. The attention mechanism, located in the bottleneck of the network, helps the model preserve edges, fine letter details, and connection structures in Persian script, thereby enhancing text clarity and readability, while the fuzzy layer, also placed in the same section, reconstructs the output with more details. For evaluation, two datasets were used, where IR-LPR contains images of Iranian license plates and IDPL-PFOD2 is a large collection of printed Persian texts. The results show that FAUNet has outperformed advanced methods such as Bicubic, SRGAN, Real-ESRGAN, and SwinIR. Overall, by combining fuzzy logic and the attention mechanism within the framework of U-Net and diffusion models, this model has significantly enhanced the quality of Persian text image reconstruction.

Future works

Additionally, future work may explore the integration of differentiable neural architecture search (NAS) to automatically optimize the structure of FAUNet for text-specific super-resolution. Recent methods such as Differentiable Architecture Search with Attention Mechanisms for GANs [44] and Self-Adaptive Weighting Based on Dual-Attention for differentiable NAS [45] suggest promising strategies for improving generative architectures. Applying such techniques potentially with text-aware or language-conditioned modules could lead to more adaptive and efficient architectures tailored to complex textual patterns and scripts like Persian.

References

[1] A. Afkari-Fahandari, F. Asadi-Zeydabadi, E. Shabaninia, H. Nezamabadi-Pour, *Enhancing Farsi text recognition via iteratively using a language model*, in 2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP), (2024), 1-6. https://doi.org/10.1109/AISP61396.2024.10475269

[2] S. AlHalawani, B. Benjdira, A. Ammar, A. Koubaa, A. M. Ali, *DiffPlate: A diffusion model for super-resolution of license plate images*, Electronics, **13**(13) (2024), 2670. https://doi.org/10.3390/electronics13132670

[3] F. Asadi-Zeydabadi, A. Afkari-Fahandari, A. Faraji, E. Shabaninia, H. Nezamabadi-Pour, *IDPL-PFOD2: A new large-scale dataset for printed Farsi optical character recognition*, arXiv preprint, arXiv:2312.01177, (2023). https://doi.org/10.48550/arXiv.2312.01177

[4] A. Brock, J. Donahue, K. Simonyan, *Large scale GAN training for high fidelity natural image synthesis*, in Proceedings of the 7th International Conference on Learning Representations (ICLR), 2019. https://doi.org/10.48550/arXiv.1809.11096

[5] W. Chao, J. Zhao, F. Duan, G. Wang, et al., *LFSRDiff: Light field image super-resolution via diffusion models*, in ICASSP 2025–2025, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2025), 1-5. https://doi.org/10.1109/ICASSP49660.2025.10889642

[6] K. Cheng, et al., *Effective diffusion transformer architecture for image super-resolution*, in Proceedings of the AAAI Conference on Artificial Intelligence, **39**(3) (2025), 2455-2463. https://doi.org/10.1609/aaai.v39i3.32247

[7] C. Dong, C. C. Loy, K. He, X. Tang, *Image super-resolution using deep convolutional networks*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **38**(2) (2016), 295-307. https://doi.org/10.1109/TPAMI.2015.2439281

[8] C. Dong, C. C. Loy, X. Tang, *Accelerating the super-resolution convolutional neural network*, in European Conference on Computer Vision, (2016), 391-407. https://doi.org/10.1007/978-3-319-46475-6_25

[9] M. El Helou, *Fuzzy-conditioned diffusion and diffusion projection attention applied to facial image correction*, in 2023, IEEE International Conference on Image Processing (ICIP), (2023), 236-240. https://doi.org/10.1109/ICIP49359.2023.10223103

[10] C. Y. Fang, C. S. Fuh, P. Yen, S. Cherng, S. W. Chen, *An automatic road sign recognition system based on a computational model of human recognition processing*, Computer Vision and Image Understanding, **96**(2) (2004), 237-268. https://doi.org/10.1016/j.cviu.2004.02.007

[11] S. Frolov, T. Hinz, F. Raue, J. Hees, A. Dengel, *Adversarial text-to-image synthesis: A review*, Neural Networks, **144** (2021), 187-209. https://doi.org/10.1016/j.neunet.2021.07.019

[12] I. J. Goodfellow, et al., *Generative adversarial nets*, Advances in Neural Information Processing Systems, **27** (2014), 2672-2680. https://doi.org/10.48550/arXiv.1406.2661

[13] M. H. Guo, et al., *Attention mechanisms in computer vision: A survey*, Computational Visual Media, **8**(3) (2022), 331-368. https://doi.org/10.1007/s41095-022-0271-y

[14] J. Ho, A. Jain, P. Abbeel, *Denoising diffusion probabilistic models*, Advances in Neural Information Processing Systems, **33** (2020), 6840-6851. https://doi.org/10.48550/arXiv.2006.11239

[15] L. Hua, et al., *Attention in diffusion model: A survey*, arXiv preprint, arXiv:2504.03738, (2025). https://doi.org/10.48550/arXiv.2504.03738

[16] R. Keys, *Cubic convolution interpolation for digital image processing*, IEEE Transactions on Acoustics, Speech, and Signal Processing, **29**(6) (1981), 1153-1160. https://doi.org/10.1109/TASSP.1981.1163711

[17] T. R. Khalifa, X. Yu, X. Zhong, Z. Wu, *Adaptive general type-2 fuzzy model-based control for nonlinear networked systems with packet dropouts*, ISA Transactions, **159** (2025), 257-277. https://doi.org/10.1016/j.isatra.2025.02.009

[18] T. R. Khalifa, X. Yu, X. Zhong, Z. Wu, *Indirect adaptive interval type-3 fuzzy tracking control for nonlinear discrete-time networked control systems with DoS attacks*, IEEE Transactions on Cybernetics, **55**(10) (2025), 4967-4980. https://doi.org/10.1109/TCYB.2025.3591555

[19] E. Khodadadi, H. R. Kanan, *Which super-resolution algorithm is proper for Farsi text image sequences*, in 2015 2nd, International Conference on Pattern Recognition and Image Analysis (IPRIA), (2015), 1-4. https://doi.org/10.1109/PRIA.2015.7161617

[20] D. P. Kingma, J. Ba, *Adam: A method for stochastic optimization*, in Proceedings of the 3rd International Conference on Learning Representations (ICLR), (2015). https://doi.org/10.48550/arXiv.1412.6980

[21] M. Kushki, E. Rashedi, E. Shabaninia, M. Kamandar, *Enhancing low-resolution Persian license plates via diffusion models*, Journal of Computing and Security, under revision, 2026

[22] C. Ledig, et al., *Photo-realistic single image super-resolution using a generative adversarial network*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2017), 4681-4690. https://doi.org/10.1109/CVPR.2017.19

[23] H. Li, et al., *SRDiff: Single image super-resolution with diffusion probabilistic models*, Neurocomputing, **479** (2022), 47-59. https://doi.org/10.1016/j.neucom.2022.01.029

[24] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, *SwinIR: Image restoration using swin transformer*, in Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), (2021), 1833-1844. https://doi.org/10.1109/ICCVW54120.2021.00210

[25] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, T. B. Schön, *Refusion: Enabling large-size realistic image restoration with latent-space diffusion models*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), (2023), 1680-1691. https://doi.org/10.1109/CVPRW59228.2023.00169

[26] K. Mehregan, A. Ahmadyfard, H. Khosravi, *Super-resolution of license-plates using frames of low-resolution video*, in 2019 5th, Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS), (2019), 1-6. https://doi.org/10.1109/ICSPIS48872.2019.9066104

[27] B. B. Moser, S. Frolov, F. Raue, S. Palacio, A. Dengel, *Dynamic attention-guided diffusion for image super-resolution*, in 2025, IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), (2025), 451-460. https://doi.org/10.1109/WACV61041.2025.00054

[28] B. B. Moser, A. Shanbhag, F. Raue, S. Frolov, S. Palacio, A. Dengel, *Diffusion models, image super-resolution and everything: A survey*, IEEE Transactions on Neural Networks and Learning Systems, early access, **36**(7) (2025). https://doi.org/10.1109/TNNLS.2024.3476671

[29] J. Nam, H. Kim, D. Lee, S. Jin, S. Kim, S. Chang, *DreamMatcher: Appearance matching self-attention for semantically-consistent text-to-image personalization*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2024), 8100-8110. https://doi.org/10.1109/CVPR52733.2024.00774

[30] Z. Niu, G. Zhong, H. Yu, *A review on the attention mechanism of deep learning*, Neurocomputing, **452** (2021), 48-62. https://doi.org/10.1016/j.neucom.2021.03.091

[31] M. Rahmani, M. Sabaghian, S. M. Moghadami, M. M. Talaie, M. Naghibi, M. A. Keyvanrad, *IR-LPR: A large scale Iranian license plate recognition dataset*, in 2022 12th, International Conference on Computer and Knowledge Engineering (ICCKE), (2022), 53-58. https://doi.org/10.1109/ICCKE57176.2022.9960129

[32] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, *High-resolution image synthesis with latent diffusion models*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2022), 10684-10695. https://doi.org/10.1109/CVPR52688.2022.01042

[33] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, M. Norouzi, *Image super-resolution via iterative refinement*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **45**(4) (2023), 4713-4726. https://doi.org/10.1109/TPAMI.2022.3204461

[34] A. Sharafian, A. Ali, I. Ullah, T. R. Khalifa, X. Bai, L. Qiu, *Fuzzy adaptive control for consensus tracking in multiagent systems with incommensurate fractional-order dynamics: Application to power systems*, Information Sciences, **689** (2025), Article 121455. https://doi.org/10.1016/j.ins.2024.121455

[35] W. Shi, et al., *Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2016), 1874-1883. https://doi.org/10.1109/CVPR.2016.207

[36] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, *Deep unsupervised learning using nonequilibrium thermodynamics*, in Proceedings of the 32nd International Conference on Machine Learning (ICML), (2015), 2256-2265. https://doi.org/10.48550/arXiv.1503.03585

[37] A. Torkian, P. Moallem, *Multi-frame super resolution for improving vehicle licence plate recognition*, Signal and Data Processing, **16**(2) (2016), 61-76. https://doi.org/10.29252/jsdp.16.2.61

[38] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, *Image quality assessment: From error visibility to structural similarity*, IEEE Transactions on Image Processing, **13**(4) (2004), 600-612. https://doi.org/10.1109/TIP.2003.819861

[39] Z. Wang, J. Chen, S. C. Hoi, *Deep learning for image super-resolution: A survey*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **43**(10) (2021), 3365-3387. https://doi.org/10.1109/TPAMI.2020.2982166

[40] Z. Wang, E. P. Simoncelli, A. C. Bovik, *Multiscale structural similarity for image quality assessment*, in The Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, **2** (2003), 1398-1402. https://doi.org/10.1109/ACSSC.2003.1292216

[41] X. Wang, L. Xie, C. Dong, Y. Shan, *Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data*, in Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), (2021), 1905-1914. https://doi.org/10.1109/ICCVW54120.2021.00217

[42] Y. Wang, J. Xu, Z. Liang, Q. Chong, X. Cheng, *FDM: Document image seen-through removal via fuzzy diffusion models*, Pattern Recognition Letters, **184** (2024), 183-189. https://doi.org/10.1016/j.patrec.2024.06.015

[43] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, L. Zhang, *EDiffSR: An efficient diffusion probabilistic model for remote sensing image super-resolution*, IEEE Transactions on Geoscience and Remote Sensing, **62** (2024), 1-14. https://doi.org/10.1109/TGRS.2023.3341437

[44] Y. Xue, K. Chen, F. Neri, *Differentiable architecture search with attention mechanisms for generative adversarial networks*, IEEE Transactions on Emerging Topics in Computational Intelligence, **8**(4) (2024), 3141-3151. https://doi.org/10.1109/TETCI.2024.3369998

[45] Y. Xue, X. Han, Z. Wang, *Self-adaptive weight based on dual-attention for differentiable neural architecture search*, IEEE Transactions on Industrial Informatics, **20**(4) (2024), 6394-6403. https://doi.org/10.1109/TII.2023.3348843