Pattern Recognition 180 (2026) 114090

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Perceiving degradation prompts in residual diffusion model for real image denoising

Meiqin Liu [a,b], Xuan Long [a,b], Qi Tang [a,b], Chao Yao [c,*], Yao Zhao [a,b]

[a] Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
[b] Visual Intelligence +X International Cooperation Joint Laboratory of MOE, Beijing 100044, China
[c] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

ARTICLE INFO

Keywords:
Image denoising
Diffusion model
Sampling acceleration

ABSTRACT

Mitigating complex noise and restoring realistic textures are critical challenges in real-world image denoising. Existing diffusion model-based denoising methods leverage their generative modeling capabilities to iteratively denoise images, but suffer from the sampling inefficiency and error propagation. In this paper, we propose a noisy-intensity degradation prompts diffusion (NID-PD) framework. Specifically, a progressive state-coupled diffusion training strategy is designed to exploit and utilize the degradation priors for the bidirectional alignment between the forward and sampling processes. In this strategy, a noisy-intensity degradation learning stage is designed, followed by a degradation prompting stage. Furthermore, a degradation refinement unit is devised to adaptively incorporate and refine degradation priors, enabling a positive evolutionary process for both priors and network features. An adaptive sampling strategy is also designed within the residual diffusion model to accelerate the sampling process. Extensive experiments demonstrate that our method achieves superior real-world denoising performance on *SIDD*, *DND* and *Nam* datasets, even with only a single sampling step.

1. Introduction

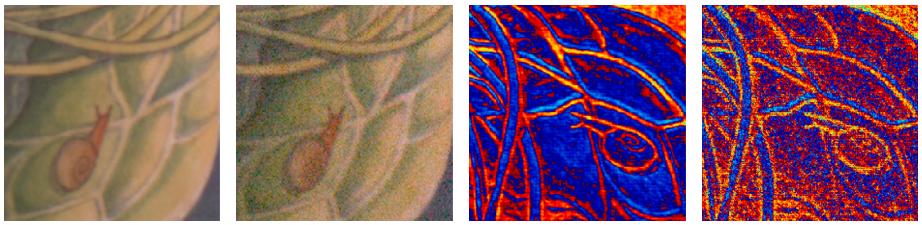
Image denoising is a core task in low-level vision, which aims to restore original signals from noise-corrupted observed images. The existing deep learning-based image denoising methods have achieved outstanding performance under synthetic image denoising (e.g., additive white Gaussian noise, AWGN). Nevertheless, in real-world scenarios, the noise distribution is unknown and heterogeneous. Real image denoising aims to remove such complex noise from real-world images while maintaining essential details with high fidelity, thereby improving visual quality for human perception and enabling practical applications such as medical imaging, remote sensing, and surveillance systems.

CNN-based image denoising methods [1] have been widely adopted for their effectiveness in capturing local features. These methods are particularly powerful for restoring fine details in images, yet their limited receptive fields restrict the ability to model global dependencies, which are crucial for high-quality denoising. Transformer-based methods [2] leverage self-attention mechanisms to capture long-range dependencies, thus overcoming the limitations of CNN-based methods. By fusing both local and global features, transformer-based image denoising methods achieve significant improvements. Leveraging the impressive generative capabilities of diffusion models (DM), numerous low-level vision tasks have witnessed remarkable achievements. These models have demonstrated the ability to produce high-quality results with intricate texture details, making them a highly effective tool for tasks such as image super-resolution, deblurring, etc.

Some researchers [3] have attempted to generate photorealistic images by leveraging the generative potential of DM in image denoising. The interpretability and robustness of DM motivate the exploration of optimal pathways to reverse the noise-additive process effectively. However, directly applying standard DMs to denoising tasks introduces challenges such as excessive sampling steps and inherent stochastic fluctuations. Meanwhile, most DM-based image denoising methods [4] typically utilize time embedding to represent the noise intensity of the input image during training, while neglecting specific noise characteristics such as spatial distribution and intensity variation. The denoising capability of these models is limited by insufficient noise-related information and the inherent variability of image content. As illustrated in Fig. 1(c), relying on coarse noise information (i.e., time embedding) alone results in inadequate noise priors, which leads to

* Corresponding author.
E-mail addresses: mqliu@bjtu.edu.cn (M. Liu), 22120321@bjtu.edu.cn (X. Long), qitang@bjtu.edu.cn (Q. Tang), yaochao@ustb.edu.cn (C. Yao), yzhao@bjtu.edu.cn (Y. Zhao).

https://doi.org/10.1016/j.patcog.2026.114090
Received 6 July 2025; Received in revised form 1 May 2026; Accepted 24 May 2026
Available online 30 May 2026



(a) Clean image (b) Noisy image (c) W/o DP (d) W DP

Fig. 1. Feature visualization in the NID-PD framework. (a) and (b) represent the clean image and corresponding noisy image, respectively. Intermediate feature visualizations in the NID-PD are illustrated as follows: (c) features relying solely on time embeddings, and (d) features enhanced by degradation embeddings. The warm-colored regions highlight the more significant areas of the features, attributed to the inclusion of degradation embeddings compared to the scenario without degradation information.

under-prediction and introduce inaccuracy and instability for denoising in subsequent sampling steps. Thus, we aim to extract degradation prior (DP) to characterize the noise representations more precisely in noisy images. By incorporating additional information about the texture and intensity of noise, our network enhances the ability to predict the most realistic counterparts.

In this paper, we propose a noisy-intensity degradation prompts diffusion (NID-PD) for real image denoising. NID-PD comprises two stages: a noisy-intensity degradation learning stage and a degradation prompting stage. These stages are designed to extract degradation priors from noisy observations and integrate them into a denoising network via prompt learning. Specifically, the intermediate noisy image is fed into the degradation-aware embedding module, where specific noise characteristics are encoded into compact degradation embeddings. Furthermore, a self-supervised learning strategy is introduced to constrain reliable degradation representations by noisy image synthesizer. Subsequently, a Transformer-based U-Net is adopted to perform the denoising process. The encoder introduces time embeddings to achieve coarse-grained denoising, while the decoder leverages the degradation embeddings to further distinguish noise from original content in a fine-grained manner. A degradation refinement unit (DRU) is introduced to effectively incorporate the degradation embeddings into the denoising network's feature representations. DRU module adapts the degradation embeddings to decoder features and progressively refines them. Additionally, NID-PD establishes a shorter pathway between noisy images and their corresponding clean counterparts, significantly accelerating the reverse sampling process.

In summary, our main contributions are summarized below:

- We present a residual diffusion model for real image denoising, exploiting degradation priors to capture local noise patterns. By integrating local noise characteristics with time-step global noise information, the design enables dynamic perception of image degradation intensity.
- We design a progressive state-coupled diffusion training strategy to extract degradation priors and incorporate them into the diffusion model for denoising. The strategy ensures the alignment of the intermediate noisy image at each step in both forward and sampling processes.
- We propose a degradation refinement unit based on prompt learning to facilitate the interaction between degradation priors and denoising features. The unit dynamically updates the priors to match the evolving features of the denoising network, ensuring adaptive guidance tailored to specific noisy input.

While our previous work [5] proposed the residual continuous diffusion model (RCDM) for efficient sampling, the current work introduces degradation prompts as a novel mechanism for guiding the denoising process. Specifically, the linear forward path of RCDM enables the feasibility of one-step sampling, while the newly proposed degradation prompts, progressive state-coupled training strategy, and degradation refinement unit ensure excellent denoising performance even with single-step inference, especially for complex real-world heterogeneous noise. The integration of degradation-aware embeddings with diffusion models represents a paradigm shift from purely generative approaches to prompt-guided restoration.

2. Related work

2.1. Real image denoising algorithms

Real image denoising algorithms can be classified into three main types: CNN-based, Transformer-based, and Diffusion Model-based methods. CNN-based methods have long been the cornerstone of image denoising, leveraging convolutional neural networks to effectively capture local features and achieve significant improvements over traditional techniques. Early works, such as DnCNN [1], demonstrated the effectiveness of CNNs in capturing local features and reducing noise. These methods employed key techniques, including residual learning, batch normalization, etc., to improve their performance. Building on these foundations, recent studies have proposed more sophisticated network designs. For instance, multi-scale approaches [6] and local attention mechanisms [7] have been incorporated to enhance feature representation and denoising capabilities. Moreover, architectures like UNet [8] and its derivatives leverage skip connections to effectively integrate low-level and high-level features, resulting in substantial improvements in denoising quality.

Transformer-based image denoising algorithms address the limitations of CNN-based methods by overcoming the constraint of local feature capture. Vision Transformers (ViTs), such as SwinIR [2] and Restormer [9], leveraged the self-attention mechanism to model long-range dependencies, offering distinct advantages over traditional CNNs. Techniques such as window partitioning and channel-wise self-attention are employed to reduce computational overhead. Additionally, X. Chen proposed DRSformer [10] with efficient sparse attention to focus on the most critical regions, further improving performance. Condformer [11] embeds noise priors into the latent space of a conditional transformer, realizing enhanced denoising via adaptive noise estimation. The transformer architecture is further optimized by Gautam et al. [12] for efficient image denoising with a lightweight design. Real image denoising is further advanced by Zhao et al. [13] via an adaptive dual-domain network designed for multi-scale spatial-frequency noise feature modeling. These advancements reflect the continuous evolution of image denoising methods, with recent approaches increasingly integrating the strengths of CNNs and Transformer-based architectures to achieve superior results.

2.2. Diffusion models for real image denoising

Diffusion models (DMs) have demonstrated significant potential in image restoration tasks, including image super-resolution, image

deblurring, etc., by iteratively refining noisy inputs through a stochastic process. By utilizing a parameterized Markov chain, DMs iteratively refine noisy inputs while optimizing the variational lower bound of the likelihood function. X. Lin devised DiffBir [14], which utilized pre-trained text-to-image DM to provide generative priors for various blind image restoration tasks. Y. Wang presented DDNM [15] to refine null-space contents during the reverse diffusion process, enabling diverse outputs that satisfy both data consistency and visual realism using pre-trained DM.

Recent studies have tailored DM specifically for real image denoising, introducing mechanisms to preserve image structures while effectively removing noise. C. Yang et al. [16] introduced a diffusion process with linear interpolation, effectively handling varying noise levels. L. Wang [17] proposed a coarse-to-fine training mechanism, enhancing restoration quality by constraining restoration outputs rather than noise, which improves performance during the fine-tuning stage. Unlike RDDM [18], which focuses on residual noise addition with stochastic processes, RCDM [5] adopts a deterministic continuous-time formulation that eliminates random noise injection. This design, combined with time embedding, enables more efficient sampling while maintaining high-quality reconstruction. DMID [19] utilizes adaptive embedding and ensembling strategies within diffusion models to enhance restoration quality. A task-adaptive diffusion framework with degradation-oriented adaptors [20] is proposed, which aligns task-specific degradation priors with the diffusion process for accurate image restoration. A degradation-calibrated cycle diffusion model [21] is designed, which achieves unified image restoration and enhancement through fine-grained degradation feature modeling. While DMs have demonstrated significant potential in image restoration tasks, their direct adaptation from image synthesis frameworks often overlooks the specific challenges of denoising, where preserving image structure is crucial. Consequently, recent efforts focus on developing DM variants that balance the thoroughness of the diffusion process with the efficiency needed for practical deployment in image denoising. For non-uniform degradation scenarios, a progressive refinement diffusion model [22] is further proposed to handle spatially varying degradation in real-world scenes. Several interpretable model-aided methods have also been proposed for complex noise removal, including STAR-Net [23], which exploits low-rank priors and deep unfolding for image denoising under complex degradation.

3. Method

We propose the noisy-intensity Degradation Prompts Diffusion (NID-PD) framework, which leverages a continuous-time Residual Diffusion Model with forward and reverse processes. The forward process generates intermediate noisy images for training, while the reverse process denoises the noisy inputs to sample the clean results.

3.1. Preliminaries

In this paper, we adopt diffusion models (DMs) to produce the denoised images. In traditional DMs, the input noisy image x_t is generated by adding random Gaussian noise to clean image x_0 with variance schedule $\{\beta_t \in (0,1)\}_{t=1}^{T}$ across a few time steps. At any arbitrary time, the marginal distribution of the intermediate noisy result $q(x_t|x_0)$ can be formulated as:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\cdot\mathbf{I}), \tag{1}$$

where $\mathcal{N}(\cdot)$ represents the Gaussian distribution and $\mathbf{I}$ represents the identity matrix. With the reparameterization trick, $\alpha_t = 1-\beta_t, \bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$. In the reverse process, from a sampled Gaussian random noise map x_T, DMs denoise x_T to x_0. The targeted distribution $p(x_{t-1}|x_t, x_0)$ can be formulated as:

$$p(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \frac{1}{\sqrt{\alpha_t}}\left(x_t - \epsilon\cdot\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\right), \sigma_t^2\cdot\mathbf{I}), \tag{2}$$

where $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\cdot\beta_t$ represents the variance of the added random noise. The denoising network accepts x_t and time t in DMs to predict ϵ_θ. The network parameters are optimized by minimizing the following objective, where $\epsilon \sim \mathcal{N}(0,\mathbf{I})$ is sampled and x_t is generated according to Eq. (1):

$$\mathcal{L} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I}),t}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \epsilon\sqrt{1-\bar{\alpha}_t}, t\right)\right\|_2^2\right], \tag{3}$$

where $\mathbb{E}$ represents the expectation operator.

Different from the above-mentioned DMs, we adopt a continuous-time residual diffusion model (RCDM) in our framework. The forward process adds residual noise $v(x_t,t) = x_1 - x_0$ into the clean image x_0 with constant speed to generate the intermediate noisy image x_t at time t until it reaches x_1. The marginal distribution $q(x_t|x_0)$ can be formulated as:

$$q(x_t|x_0) = \mathcal{N}(x_t; x_0 + v(x_t,t)\cdot t, \int_0^t \beta(s)ds\cdot\mathbf{I}), \tag{4}$$

where the time $t \sim U(0,1)$ represents the residual noise intensity in x_t. Furthermore, the random noise coefficient $\int_0^t \beta(s)ds = 0$, which indicates that the forward process eliminates the injection of random noise. A more direct pathway between x_0 and x_1 is created by this construction, resulting in a shorter and more efficient diffusion path compared to earlier methods. Consequently, in the reverse process, The targeted distribution $p(x_{t-\Delta t}|x_t, x_0, x_1)$ can be formulated as:

$$p(x_{t-\Delta t}|x_t, x_0, x_1) = \mathcal{N}(x_{t-\Delta t}; x_t - v(x_t,t)\cdot\Delta t, \beta(t)\cdot\mathbf{I}), \tag{5}$$

where Δ represents the discretization step of the variable t and $\Delta t \to 0$. In the reverse process, RCDM constructs a non-Markov chain, which allows the transition from x_t to $x_{t-n\times\Delta t}$. The transition distribution $p(x_{t-n\times\Delta t}|x_t, x_0, x_1)$ can be formulated as:

$$p(x_{t-n\times\Delta t}|x_t, x_0, x_1) = \mathcal{N}(x_{t-n\times\Delta t}; x_t - n\Delta t\cdot v(x_t,t), \mathbf{0}), n = 1, 2, 3, \ldots \tag{6}$$

where n represents the sampling factor. The distribution is concentrated at the mean, indicating a deterministic backward path without uncertainty. To intuitively illustrate the difference between stochastic diffusion (DDPM) and deterministic residual diffusion (RCDM), we present Fig. 2. (a) DDPM's forward process. The clean image x_0 is gradually corrupted by random Gaussian noise, eventually becoming pure noise X_T. The path is stochastic and non-linear. (b) RCDM's forward process. The clean image x_0 is transformed into a noisy image X_1 via deterministic residual interpolation $X_t = x_0 + t\times v$ with no random noise injection. The path is a straight line in the feature space. Consequently, the denoising network in RCDM accepts x_t and time t to predict v_θ. The network is optimized by a sampled $v(x_t,t)$ which generates x_t by Eq. (1):

$$\mathcal{L} = \mathbb{E}_{v,t}\left[\left\|v - v_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \epsilon\cdot(x_0 + v(x_t,t)\cdot t), t\right)\right\|_2^2\right], \tag{7}$$

where $\mathbb{E}$ represents the expectation operator.

3.2. Overview of NID-PD

3.2.1. Overall architecture

NID-PD is composed of a Transformer-based denoising U-Net and two pre-trained modules as illustrated in Fig. 3. The second training stage is denoted as the degradation prompting stage. In traditional DMs, the time t is usually encoded into the denoising U-Net by positional embedding, which provides the global noise intensity information to help the network handle various noisy inputs. In contrast, in NID-PD, degradation priors are incorporated to encapsulate local noise characteristics. The priors extracted by the pre-trained degradation embedding module (DEM), together with the time step, collaboratively supply detailed local and global noise information to guide the denoising process. Furthermore, the degradation priors are refined by the degradation refinement unit (DRU) to align with the evolving features

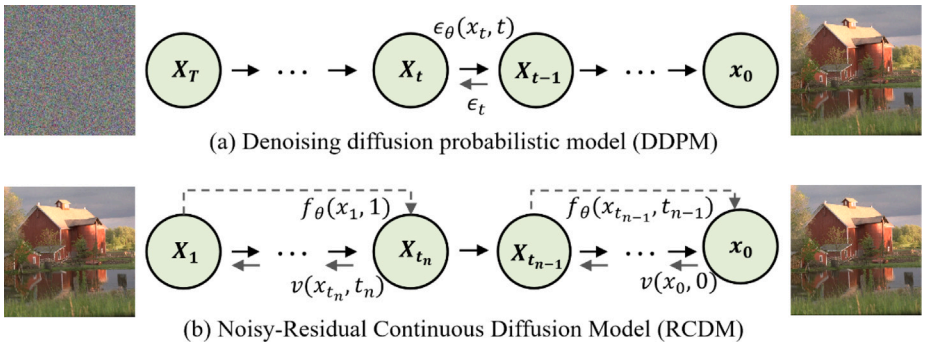


Fig. 2. Comparison of forward processes between DDPM and RCDM.

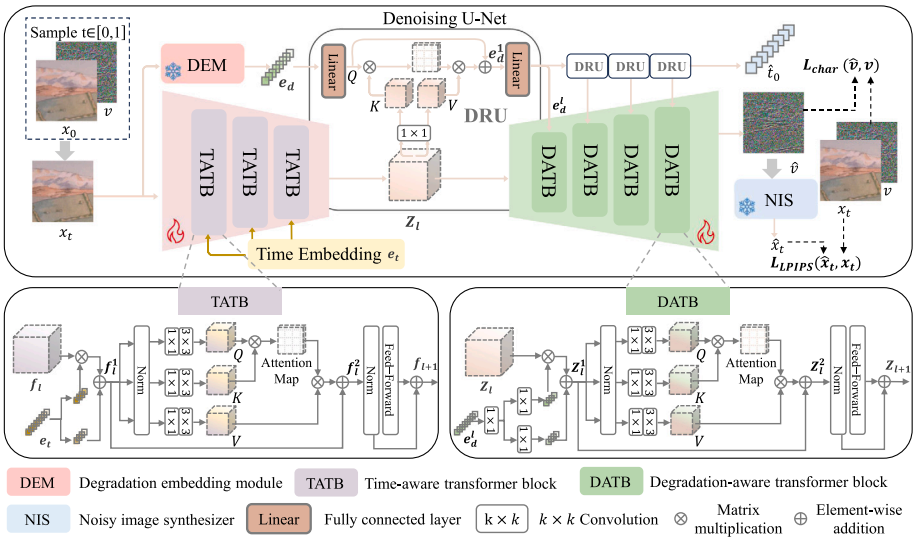


Fig. 3. The workflow of the degradation prompting stage in the NID-PD framework, with the DEM components pre-trained. The specific details of TATB and DATB modules are presented at the bottom.

in the decoder's input, thereby enabling more effective denoising. Finally, the pre-trained noisy image synthesizer is utilized to constrain DRU training for fine-grained noise representation learning. Specifically, NID-PD utilizes the continuous-time residual diffusion model (RCDM) to generalize intermediate noisy images and sample denoised images. Noise added to the clean image x_0 spreads into the noisy image x_1 with constant intensity, and the process does not introduce random noise. The input is an intermediate noisy images x_t at time t, which can be formulated as:

$$x_t = x_0 + t \times v, \tag{8}$$

where $t \in [0, 1]$ represents a continuous time variable indicating the noise intensity of the current intermediate noisy image x_t relative to the clean image x_0. $v = x_1 - x_0$ is the ground truth constant residual corresponding to the fixed slope of the linear path from x_0 to x_1 and it remains constant throughout the forward diffusion process (Eqs. (4)–(8)). In contrast, $v_\theta(x_t, t)$ or $\hat{v}$ denotes the network's predicted estimate of this residual given the current noisy state x_t and time step t used in the reverse sampling process (Eqs. (9)–(13)). The network learns to predict the constant ground truth $v$from observations at different degradation levels. The denoised image $\hat{x}_0$ can be formulated as:

$$\hat{x}_0 = x_t - (t - \hat{t}_0) \times \hat{v}, \tag{9}$$

where $\hat{t}_0$ and $\hat{v}$ represent the output of the denoising network.

Time-Aware Transformer Block. The Time-Aware Transformer Block (TATB) in NID-PD is designed to integrate time information as an encoder module of the denoising network, and the detailed structure is shown in Fig. 3. Each TATB module enables the denoising network to adapt to the noise distribution at different time steps of the forward process by introducing embedding features at specific time steps, enhancing the network's ability to handle different noise intensities.

Specifically, the time embedding e_t can be formulated as:

$$e_t = \text{Linear}(\text{GeLU}(\text{Linear}(\text{PE}(t)))), \tag{10}$$

where $\text{Linear}(\cdot)$ represents the linear layer, $\text{PE}(\cdot)$ represents the positional encoding operation, and $\text{GeLU}(\cdot)$ represents the activation function. Before integrating with the encoder features, the time embedding e_t is split along the channel dimension into modulated embedding pairs e_t^1 and e_t^2, expressed as $e_t = [e_t^1, e_t^2]$. These two parts of features are used together to modulate the input feature f_l of the TATB module to obtain the modulated feature f_l^1, which can be formulated as:

$$f_l^1 = e_t^1 * f_l + e_t^2, \tag{11}$$

where f_l represents the input features of the lth layer encoder. The modulated feature f_l^1 is normalized and passed through the 1×1 convolution $W_p^{(\cdot)}$ and the 3×3 convolution $W_d^{(\cdot)}$ to obtain the query, key and value, which can be formulated as $Q_t = W_p^Q W_d^Q \text{Norm}(f_l^1)$, $K_t = W_p^K W_d^K \text{Norm}(f_l^1)$, $V_t = W_p^V W_d^V \text{Norm}(f_l^1)$, where $\text{Norm}(\cdot)$ represents the layer normalization operation. Next, the output f_l^2 of multi-head attention can be formulated as:

$$f_l^2 = \text{MHA}(Q_t, K_t, V_t) + f_l^1, \tag{12}$$

where MHA(·) represents the multi-head attention operation, f_l^1 is the modulated feature representation before SA operation. Finally, TATB module generates the refined features f_{l+1}, which can be formulated as:

$$f_{l+1} = \mathrm{FFN}(\mathrm{Norm}(f_l^2)) + f_l^2, \tag{13}$$

where FFN(·) represents the feed-forward network, and Norm(·) represents the layer normalization [24]. By introducing a time embedding feature e_t and intra-layer modulation, the denoising network enables to perception of global noise intensity.

Degradation-Aware Transformer Block. The Degradation-Aware Transformer Block (DATB) in NID-PD is designed to integrate degradation information as a decoder module of the denoising network, shown in Fig. 3. The additional information embedded by the DATB module is the degradation priors, which encode both global and local degradation patterns in the input image. As a complement to the time embedding, degradation priors enhance the denoising network's ability to perceive semantically comprehensive noise information.

Specifically, the degradation embedding e_d^l is first processed through a convolutional layer, followed by two separate convolution operations, $\phi(\cdot)$ and $\varphi(\cdot)$, which generate scale and bias parameters, respectively. These parameters are then applied to the decoder features via element-wise multiplication and addition. Given the input feature Z_l of the lth DATB layer, the modulated feature Z_l^1 can be formulated as:

$$Z_l^1 = \phi(e_d^l) \cdot Z_l + \varphi(e_d^l), \tag{14}$$

where "·" represents the element-wise multiplication operation. Subsequently, the network utilizes a self-attention mechanism for feature fusion. The modulated features Z_l^1 are first normalized and passed through 1×1 convolution $W_p^{(\cdot)}$ and 3×3 convolution $W_d^{(\cdot)}$ to generate the query, key, and value, which can be formulated as: $Q_d = W_p^Q W_d^Q \,\mathrm{Norm}(Z_l^1)$, $K_d = W_p^K W_d^K \,\mathrm{Norm}(Z_l^1)$ and $V_d = W_p^V W_d^V \,\mathrm{Norm}(Z_l^1)$. The feature of multi-head self-attention mechanism is Z_l^2, which can be formulated as:

$$Z_l^2 = \mathrm{MHA}(Q_d, K_d, V_d) + Z_l^1. \tag{15}$$

Finally, the DATB module further enhances the feature representation by a feed-forward neural network to obtain Z_{l+1}, which can be formulated as:

$$Z_{l+1} = \mathrm{FFN}(\mathrm{Norm}(Z_l^2)) + Z_l^2, \tag{16}$$

where FFN(·) represents the feed-forward network, and Norm(·) represents the layer normalization [24]. By introducing the degradation embedding feature e_d and intra-layer modulation, the denoising network has the ability to perceive the global and local noise degradation information.

Degradation Refinement Unit. The Degradation Refinement Unit (DRU) is designed to adaptively adjust and refine the degradation embedding in response to evolving characteristics within the decoder, as depicted in Fig. 3. The inputs to the DRU module consist of the network features Z_l of the lth layer decoder and degradation embeddings e_d extracted by the pre-trained DEM module. The degradation embeddings are updated by cross-attention. Query Q_u, key K_u and value V_u can be formulated as:

$$Q_u = \mathrm{Linear}(e_d), K_u = \mathrm{Conv}(Z_l), V_u = \mathrm{Conv}(Z_l), \tag{17}$$

where Linear(·) represents the fully connected layer, Conv(·) represents the 1×1 convolution layer, and $Q_u \in \mathbb{R}^{1\times 8C}$, $K_u, V_u \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 8C}$. Subsequently, the query Q_u, key K_u and value V_u pass through the cross-attention layer and the degenerate embedding e_d is updated by decoder feature Z_l. The updated features e_d^1 can be formulated as:

$$e_d^1 = \mathrm{CA}(Q_u, K_u, V_u) + Q, \tag{18}$$

where CA(·) represents the cross-attention operation.

Through the aforementioned cross-attention operation, the degradation embedding representation is enriched and modulated by the decoder features. The degradation embedding is subsequently transformed by a standard fully connected layer into the representation e_d^l, which can be formulated as:

$$e_d^l = \mathrm{Linear}(e_d^1), \tag{19}$$

where Linear(·) represents the fully connected layer that aligns the channel dimension of the degradation embedding with the lth decoder.

3.2.2. The noisy-intensity degradation learning stage

In the noisy-intensity degradation learning stage, the NID-PD adopts a self-supervised training strategy based on the intermediate noisy images. This stage includes the degradation embedding module (DEM) and the noisy image synthesizer (NIS) in order to provide a reliable degradation representation for the degradation prompting stage, as shown in Fig. 4.

DEM module is designed with cascaded residual blocks to capture the degradation information. After these residual blocks, an adaptive average pooling layer is applied to distill the extracted features into a compact degradation embedding $e_d \in \mathbb{R}^{1\times c_d}$, where c_d represents the dimension of the degradation embedding. e_d can be formulated as:

$$e_d = S_\omega(x_t, e_t), \tag{20}$$

where $S_\omega(\cdot)$ represents the degradation-aware embedding module, e_t represents the time embedding to provide the global noise intensity.

Then, the predicted intermediate noisy image $\hat{x}_t$ at time t is reconstructed by the noisy image synthesizer (NIS). In the NIS, time embeddings e_t and degradation embeddings e_d are used as generators of the dynamic convolution kernel, which is applied to x_0 through parameterized convolution operations to simulate the degradation process. $\hat{x}_t$ can be formulated as:

$$\hat{x}_t = \xi_\omega(x_0, e_t, e_d), \tag{21}$$

where $\xi_\omega(\cdot)$ represents the noisy image synthesizer. The degradation embedding e_d is leveraged to modulate the clean image, ensuring that the predicted noisy image $\hat{x}_t$ is effectively influenced by the noise characteristics dictated by the degradation process. The seamless integration of the time embedding e_t and the degradation embedding e_d within the NIS enables precise reconstruction of the noisy image.

To maintain the consistency between the states at the same time t in both the forward and sampling processes, the first stage loss function L_{s1} can be formulated as:

$$\arg\min_\omega L_{s1}(x_t, \xi_\omega(x_0, e_t, e_d)), \tag{22}$$

where $\xi_\omega(\cdot)$ represents the NIS module. ω denotes the parameters to be learned in the noisy-intensity degradation learning stage. The optimal ω is attained only when x_t and the $\hat{x}_t$ predicted by the NIS module are the closest. Compared with generating clean images with complex textures, reconstructing the intermediate noisy image is more available to the model, which helps to simplify the training process and accelerate the network convergence.

3.2.3. Progressive state-coupled diffusion training strategy

A progressive state-coupled diffusion training strategy is designed to optimize the model through a two-step process. First, the model is trained in the noisy-intensity degradation learning stage to obtain a pre-trained DEM. The stage models the noise-related degradation process, aiming to capture and extract comprehensive information associated with noise degradation. The pre-trained DEM is capable of extracting degradation embedding, which serves as a support for the subsequent degradation prompting stage. The second stage is called the degradation prompting stage. In this stage, the noise degradation embeddings are enriched and refined by DRU, effectively guiding the denoising process. Time and degradation embeddings jointly support

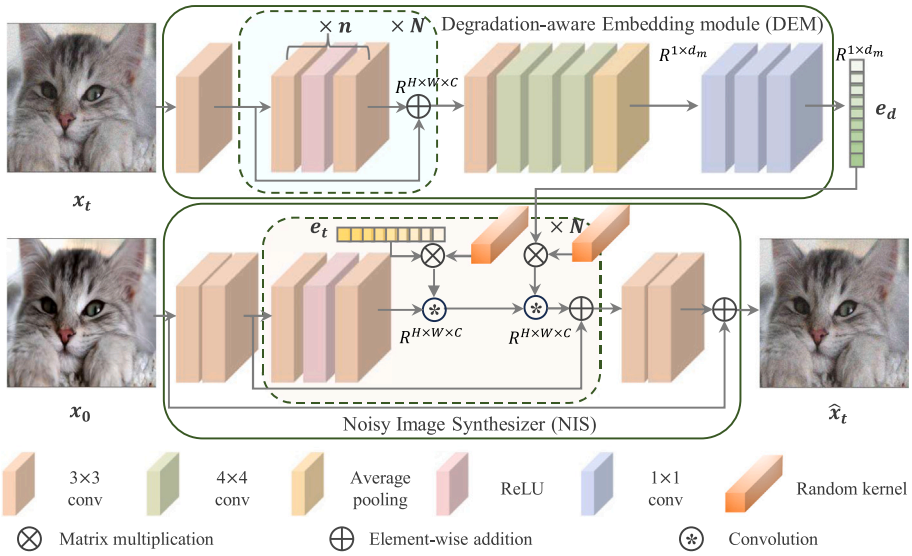


Fig. 4. A detailed architecture of the noisy-intensity degradation learning stage of NID-PD. Degradation embedding is extracted by the degradation-aware embedding module and is further utilized to modulate clean images into noisy counterparts in the noisy image synthesizer.

Algorithm 1 The Sampling and Inference Process

Input: Noisy image x_1, pre-trained DEM S_ω, pre-trained denoising network v_θ
Hyper-parameter: max step T_{max}, time precision t_ϵ

Let $T = 1, t = 1$, $\hat{x}_1 = x_1$
while $T < T_{max}$ and $t > t_\epsilon$ **do**
  $e_d = S_\omega(\hat{x}_t, t)$
  $\hat{v}, \hat{t}_0 = v_\theta(\hat{x}_t, t, e_d)$
  $\hat{x}_{\hat{t}_0} = x_t - (t - \hat{t}_0) \times \hat{v}$
  $T = T + 1, t = \hat{t}_0$
end while
return $\hat{x}_t$

the modeling of noise characteristics from both global and local perspectives, enhancing perceptual consistency in both the forward and sampling processes. This training strategy is designed to avoid training the DEM and denoising U-Net from scratch simultaneously, which may cause gradient instability and parameter divergence, ultimately harming denoising performance. By pre-training the DEM to generate informative degradation features and then incorporating them into the denoising network, this designed strategy improves both model stability and overall denoising effectiveness.

3.3. The sample process of NID-PD

We utilize the Residual Continuous Diffusion Model in the whole training and sampling process. The continuity of RCDM allows the adaption of our sampling process. Although our one-step sampling is a deterministic feed-forward process, the diffusion framework models noise degradation as a continuous probabilistic distribution transformation, which improves generalization to unseen real noise. It employs our progressive state-coupled training to align intermediate states between the forward and reverse processes for more stable prediction, and supports flexible switching between single-step and multi-step inference to fit different noise types, whereas a regression network can only perform fixed single forward inference. With Eq. (6), the reverse process of RCDM is defined based on the non-Markov chain to accelerate sampling further, and the predicted target distribution $p_\theta(x_{t-n\times dt}|x_t, x_1)$ can be formulated as:

$$p_\theta(x_{t-n\times\Delta t}|x_t, x_1) = \mathcal{N}(x_{t-\Delta t}; x_t - n\times v_\theta(x_t, t, e_d)\Delta t, \mathbf{0}), n = 1, 2, 3, \dots, \quad (23)$$

where $v_\theta(\cdot,\cdot,\cdot)$ represents the denoising network in the degradation prompting stage, x_1 represents the degraded image to be denoised, e_d represents the degradation embeddings extracted by DEM module.

Specifically, the sampling and inference process can be described in an algorithm 1. The maximum sampling steps T_{max} is a key hyper-parameter. T_{max} limits the sampling process for computational efficiency, while the network adjusts sampling time dynamically based on residual bias from previous predictions. Ideally, the network is capable of achieving one-step sampling, thereby optimizing efficiency. The performance of these parameters will be rigorously evaluated in the experimental section.

3.4. Loss function

The training objective employs a dual-phase loss formulation, where each stage addresses distinct optimization goals through carefully weighted perceptual and pixel-level constraints. The design ensures progressive learning of degradation characteristics while maintaining reconstruction fidelity.

Stage I: Degradation Embedding Consistency. Focused on preserving noise-degradation correspondence, the first-stage loss $\mathcal{L}_{s1}$ can be formulated as:

$$\mathcal{L}_{s1} = \underbrace{L_{\text{Char}}(x_t, \hat{x}_t)}_{\text{pixel accuracy}} + \lambda \underbrace{L_{\text{LPIPS}}(x_t, \hat{x}_t)}_{\text{perceptual quality}}, \quad (24)$$

where $\lambda \in \mathbb{R}^+$ represents hyper-parameter to balance reconstruction precision versus visual naturalness.

Stage II: Refined Denoising Optimization. To enhance temporal consistency in diffusion models, the second-stage loss $\mathcal{L}_{s1}$ can be formulated as:

$$\mathcal{L}_{s2} = \underbrace{L_{\text{Char}}(x_1 - x_0, \hat{v})}_{\text{velocity matching}} + \gamma \underbrace{L_{\text{LPIPS}}(x_t, \hat{x}_t)}_{\text{perceptual anchoring}}, \quad (25)$$

where $\gamma \in \mathbb{R}^+$ represents a hyper-parameter to control the preservation of perceptual features during denoising trajectory optimization.

Table 1
Quantitative results on real image denoising for *SIDD*, *DND* and *Nam* dataset. The best performance is displayed in bold.

Type	Methods	*SIDD* [25]		*DND* [26]		*Nam* [27]		Inference
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	Step
CNN	DnCNN [1]	23.66	0.583	32.43	0.790	37.49	0.927	–
	CBDNet [29]	30.78	0.801	38.06	0.942	40.02	0.969	–
	RIDNet [30]	38.71	0.951	39.26	0.953	–	–	–
	DANet+ [31]	39.47	0.957	39.58	0.955	–	–	–
	CycleISP [32]	39.52	0.957	39.56	0.956	40.42	0.992	–
	MIRNet [6]	39.72	0.959	39.88	0.956	39.81	0.988	–
	MPRNet [33]	39.71	0.958	39.80	0.954	39.81	0.990	–
	NBNet [34]	39.75	0.959	39.89	0.955	–	–	–
	ADFNet [35]	39.79	0.960	39.87	0.955	–	–	–
Transformer	UFormer [36]	39.77	0.959	39.96	0.956	39.83	0.991	–
	SwinIR [2]	39.77	0.958	40.01	**0.958**	–	–	–
	Restormer [9]	40.02	0.960	40.03	0.956	40.50	0.992	–
	Condformer [11]	**40.21**	0.961	**40.10**	0.956	42.04	0.978	–
DM	DDRM [37]	34.77	0.915	–	–	–	–	200
	LDM [38]	34.77	0.900	–	–	–	–	200
	RnG [39]	38.03	0.926	–	–	–	–	76
	C. Yang [16]	39.40	0.959	39.75	0.957	–	–	10
	C2F-DFT [17]	39.84	0.960	39.95	0.955	40.14	0.992	4
	DMID-p [19]	31.07	0.703	–	–	39.53	0.9567	–
	DMID-d [19]	–	–	–	–	**42.59**	0.983	–
	DMID-4500 [19]	33.41	0.9130	–	–	–	–	–
	NID-PD (Ours)	39.96	**0.960**	40.08	0.957	40.55	**0.993**	1

4. Experiments

4.1. Datasets and evaluation metrics

We employ 320 high-resolution images of the *SIDD Medium* [25] dataset as the training dataset. For the test datasets, we adopt three datasets: 1280 patches from the *SIDD* validation, 1000 patches from the *DND* [26] dataset, and images with 11 static scenes from *Nam* [27] dataset. PSNR, SSIM, and LPIPS are employed to evaluate the objective results. The Charbonnier loss [28] L_{char} is used to optimize our model in two training stages, which can be formulated as:

$$L_{char} = \sqrt{\|x - x_{GT}\|^2 + \epsilon^2}, \tag{26}$$

where x represents the predicted output, and x_{GT} represents the corresponding ground-truth target. $\epsilon = 1 \times 10^{-3}$ represents a constant in all training stages.

4.2. Implementation details

In the training process, horizontal and vertical flips are randomly applied for data augmentation. During the noisy-intensity degradation learning stage, the networks are trained on 128×128 patches with a batch size of 16. The initial learning rate is 2×10^{-4}, and decreased to 1×10^{-6} in the cosine annealing strategy [40]. The hyper-parameter of the loss function $\lambda = 1 \times 10^{-4}$. During the degradation prompting stage, training continues with the same learning rate. The networks are trained on 128×128 patches with a batch size of 8. The hyper-parameter of the loss function $\gamma = 1 \times 10^{-2}$. The learning rate is gradually reduced from 1×10^{-4} to 1×10^{-6} in the cosine annealing strategy [40]. We use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ throughout all training stages. The experiments are conducted using PyTorch, with the NID-PD model trained on two NVIDIA GeForce RTX 3090.

4.3. Comparison with competitive methods

The quantitative comparison of NID-PD with other competitive denoising methods is presented in Table 1. Among the six CNN-based methods, ADFNet [35] achieves the best performance, with a PSNR of 39.79 dB and SSIM of 0.960 on the *SIDD* dataset, and 39.87 dB and 0.955 on the *DND* dataset, respectively. Restormer [9] tops the PSNR on *SIDD* at 40.02 dB, 0.23 dB higher than ADFNet [35], verifying the value of long-range dependency modeling for image denoising. Against Transformer-based methods, our NID-PD outperforms SwinIR [2] by 0.07 dB in PSNR on *DND* (40.08 dB) and achieves state-of-the-art results on *Nam* with a PSNR of 40.55 dB and SSIM of 0.993, 0.05 dB and 0.001 higher than Restormer [9], respectively. These findings confirm that DM-based models excel at preserving structural information and recovering high-frequency details. Among DM-based methods, DDRM [37] and LDM [38] deliver limited performance (34.77 dB PSNR) due to 200 sampling steps. C2F-DFT [17] cuts steps to 4 and reaches 39.84 dB on *SIDD*, but still lags NID-PD by 0.13 dB on *DND*. Notably, NID-PD uses only one sampling step yet outperforms C2F-DFT [17] by 0.12 dB on *SIDD*, 0.13 dB on *DND* and 0.41 dB on *Nam*. NID-PD outperforms DMID variants [19] in efficiency, matches or exceeds Condformer [11] in SSIM, and uniquely integrates dynamic prior adjustment. Compared to the fixed prior of Condformer [11] and the high-latency pipeline of DMID [19], NID-PD retains single-step speed while offering stronger robustness in complex real-noise settings. These results validate the effectiveness of degradation embeddings in diffusion models.

Table 2
Comparisons of perceptual similarity (LPIPS ↓) on *SIDD* and *Nam* datasets.

Methods	Parameters (M)	*SIDD* [25]	*Nam* [27]
DnCNN [1]	0.7	0.3553	0.0816
MIRNet [6]	31.8	0.1320	0.0403
MPRNet [33]	20.1	0.1315	0.0556
CycleISP [32]	–	0.1361	0.0488
UFormer [36]	20.6	0.1299	0.0437
Restormer [9]	26.11	0.1276	0.0393
PDASR [41]	26.8	0.1298	–
DDRM [37]	–	0.2879	–
C2F-DFT [17]	25.36	0.1305	0.0405
NID-PD (Ours)	48.67	**0.1252**	**0.0377**

The comparison results of LPIPS for the NID-PD are shown in Table 2. NID-PD achieves the lowest LPIPS values on both the *SIDD* and *Nam* datasets. Specifically, on the *SIDD* dataset, it obtains an LPIPS of 0.1252, which is 0.0063 lower than the CNN-based method MPRNet [33], 0.0024 lower than the Transformer-based method Restormer

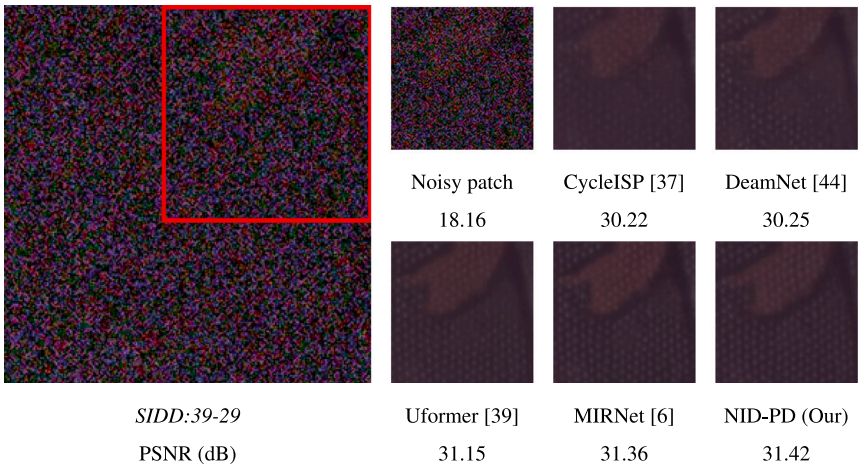


Fig. 5. Visual comparisons of real image denoising on *SIDD* dataset.

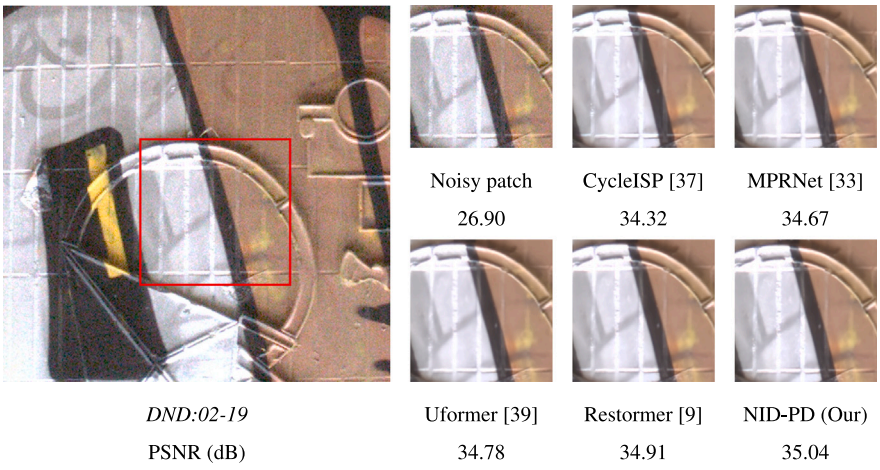


Fig. 6. Visual comparisons of real image denoising on *DND* dataset.

Table 3
Comparisons of perceptual similarity (LPIPS ↓), parameters and runtime on *SIDD* dataset.

Methods	LPIPS	Parameters (M)	Runtime (s)
Restormer [9]	0.1276	26.11	0.1968
NAFNet [43]	0.1271	29.16	**0.0821**
KBNet [42]	**0.1198**	141.97	0.2332
C2F-DFT [17]	0.1305	25.36	0.5213
NID-PD (Ours)	0.1252	48.67	0.1325

[9], and 0.0053 lower than the DM-based method C2F-DFT [17]. On the *Nam* dataset, NID-PD achieves an LPIPS of 0.0377, outperforming MIRNet [6] by 0.0026, Restormer [9] by 0.0016, and C2F-DFT [17] by 0.0028. These results demonstrate that the introduction of degradation embeddings can effectively improve the perceptual quality of denoised images in complex real-world scenarios.

The comparison results of LPIPS, parameters, and runtime on *SIDD* dataset are shown in Table 3. KBNet [42] achieves the best perceptual quality on the *SIDD* dataset with an LPIPS of 0.1198. However, it has a high model complexity with 141.9M parameters. Compared to C2F-DFT [17], NID-PD reduces runtime on *SIDD* by 74.58% and also lowers the LPIPS by 0.0053. These results demonstrate that NID-PD achieves the best perceptual quality with minimal runtime, effectively balancing perceptual performance and computational efficiency.

The qualitative results of different methods on *SIDD*, *DND* and *Nam* datasets are shown in Figs. 5–7. On the "*SIDD:39-29*" image, CycleISP [32] and DeamNet [44] exhibit severe loss of structural and textural details in their denoised images, along with noticeable artifacts. MIRNet [6] and MPRNet [33] are able to recover most of the structures and textures, but still suffer from detail loss in edge regions. In comparison, Uformer [36] preserves the structural integrity and restores most edges with more complete texture details, but results in blurring and errors in complex texture areas. In contrast, NID-PD successfully reconstructs the sharp edges and fine textures while avoiding artifacts and over-smoothing. On the "*DND:02-19*" image, MIRNet [6] and Restormer [9] display a color blending problem in the yellow region at the bottom right. On the "*Nam:11*" image, C2F-DFT [17] produces overly blurred textures for distant trees. By contrast, NID-PD effectively removes noise and reconstructs realistic edges and textures without introducing blurring or artifacts. These results demonstrate the effectiveness of NID-PD in enhancing perceptual visual quality.

4.4. Adaptive sampling performance

To evaluate the effectiveness of the designed adaptive sampling strategy, we conduct several experiments on the *Nam* dataset, as shown in Table 4. t_ϵ represents the time precision to control the granularity of sampling. T_{max} represents the maximum number of sampling steps. T_{avg} represents the average sampling steps to reflect the actual sampling efficiency. We analyze the PSNR, SSIM, and LPIPS performance of the C2F-DFT [17] method under different sampling steps. C2F-DFT [17] performs reasonably well only when the sampling step is set to 3.

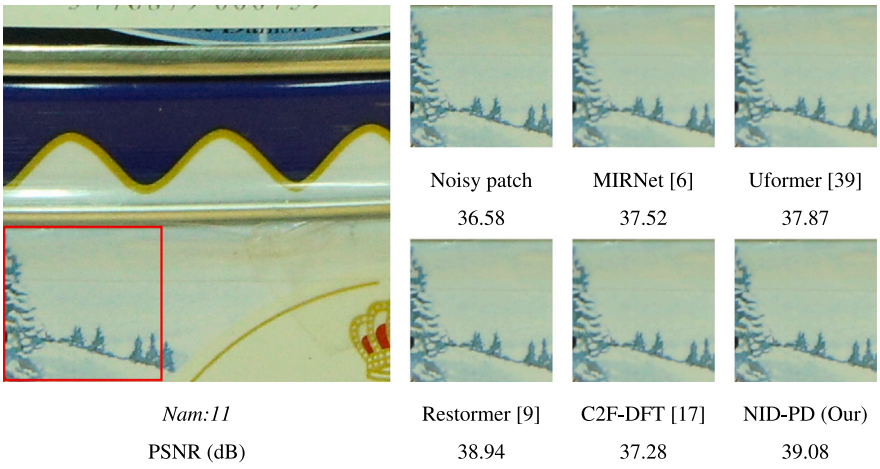


Fig. 7. Visual comparisons of real image denoising on *Nam* dataset.

Table 4
Quantitative ablation results of sampling process between C2F-DFT [17] and our NID-PD on *Nam* dataset.

Methods	T_{max}	t_ϵ	T_{avg}	PSNR (dB) ↑	SSIM↑	LPIPS↓
C2F-DFT [17]	4	–	4	38.66	0.990	0.0359
	3	–	3	40.14	0.992	0.0405
	2	–	2	30.83	0.956	0.1697
NID-PD (Ours)	4	1e−6	4	40.65	0.993	0.0312
	3	**1e-6**	**3**	**40.65**	**0.993**	**0.0310**
	2	1e−4	1	40.55	0.992	0.0412
	1	1e−6	1	40.55	0.993	0.0377

Smaller steps provide insufficient information for accurate noise estimation, while larger steps cause cumulative errors due to prediction inaccuracies. The results demonstrate that C2F-DFT [17] suffers from error accumulation during the sampling process. Our NID-PD achieves the best performance at $T_{max} = 3$ and $t_\epsilon = 1 \times 10^{-6}$, with a PSNR of 40.65 dB, SSIM of 0.993, and LPIPS of 0.0310. Although the overall performance slightly decreases under other parameter configurations, the performance remains comparable. Even under extreme settings (e.g., $T_{max} = 1$), the performance remains stable. These results validate the effectiveness of NID-PD in reducing cumulative errors.

4.5. Ablation study

4.5.1. Effects of the degradation embeddings

To validate the effectiveness of degradation embeddings (DE), we conduct an ablation study on the *SIDD* and *DND* datasets, as shown in Table 5. In Model A, both the encoder and decoder of the denoising network are TATB modules. In Model B, the encoder of the denoising network is comprised of TATB modules and the decoder is comprised of DATB modules. Specifically, compared to Model A, Model B improves the PSNR of 0.21 dB on *SIDD* and 0.08 dB on *DND* datasets.

The qualitative ablation results of degradation embedding on *SIDD* dataset are shown in Fig. 8. On the "*SIDD:01-15*" and "*SIDD:05-03*" images, the denoised image of Model A still remains noisy and blurred edge. The denoised image of model B not only removes the noise effectively, but also has clearer details. The above experimental results demonstrate the effectiveness of degenerate embedding on the denoising network to remove noise and restore texture structure.

4.5.2. Effects of the degradation refinement unit

To validate the effectiveness of the degradation refinement unit (DRU), we conduct the ablation study on the *SIDD* and *DND* dataset, as shown in Table 5. In Model B, the degradation embedding remains the same during denoising, while Model C uses DRU to update it for guiding the network. Compared to Model B, Model C improves the PSNR of 0.11 dB and the SSIM of 0.001, and decreases the LPIPS of 0.0056 on *SIDD*, and the PSNR of 0.09 dB and the SSIM of 0.001 on *DND* dataset. The qualitative ablation results of degradation embedding on *SIDD* dataset are shown in Fig. 8. On the"*SIDD:01-15*" and "*SIDD:05-03*" images, the denoised image of Model C has better human visual perception. The experimental results demonstrate that the DRU effectively facilitates the integration of degradation embeddings with the denoising network, validating the effectiveness of dynamically updated degradation embeddings.

Table 5
Quantitative ablation results of degradation embeddings and the degradation refinement unit on *SIDD* and *DND* datasets.

Model	DE	DRU	*SIDD* [25]			*DND* [26]	
			PSNR ↑	SSIM ↑	LPIPS ↓	PSNR↑	SSIM ↑
Model A			39.64	0.958	0.1361	39.91	0.955
Model B	✓		39.85	0.959	0.1308	39.99	0.956
Model C	✓	✓	**39.96**	**0.960**	**0.1252**	**40.08**	**0.957**

4.5.3. Effects of hyper-parameters λ and γ

To evaluate the LPIPS loss weights λ and γ in the noisy-intensity degradation learning and the degradation prompting stage, we conduct ablation studies on *SIDD* dataset, as shown in Fig. 9. The blue line represents the PSNR of the intermediate noisy image generated by the noisy-intensity degradation stage. A higher PSNR indicates that the extracted degradation embeddings more accurately reflect the true noise degradation patterns. The orange line represents the PSNR of the final denoised image in the degradation prompting stage, where higher PSNR values mean a better reconstruction of the NID-PD. As λ decreases, the blue line rises, suggesting improved feature extraction. In contrast, as γ decreases, the denoising performance of NID-PD initially improves and then declines. The highest reconstruction PSNR occurs when $\gamma = 1 \times 10^{-2}$. Based on these results, the optimal training configuration for the NID-PD is $\lambda = 1 \times 10^{-4}$ and $\gamma = 1 \times 10^{-2}$.

5. Conclusion

In this paper, we propose the noisy-intensity degradation prompts diffusion framework for real-world image denoising (NID-PD), integrating a progressive state-coupled diffusion training strategy with a residual diffusion model. The training scheme leverages intricate noise patterns to construct robust degradation embeddings through noise image reconstruction and incorporates these embeddings into the denoising network. By supplementing the noisy-intensity priors, the

Fig. 8. Qualitative ablation results of degradation embedding on *SIDD* dataset.

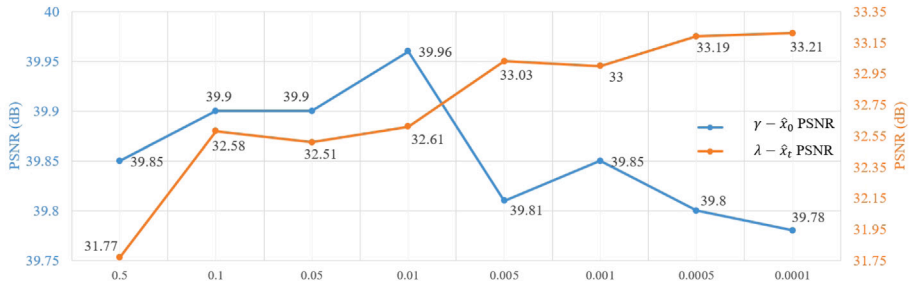


Fig. 9. Quantitative ablation results of hyper-parameters λ and γ on *SIDD* dataset.

model has the ability to distinguish diverse structures and non-uniform noise distributions. The progressive state-coupled diffusion training strategy aligns the intermediate states of the forward diffusion and reverse denoising processes, enhancing the coherence and accuracy of the reconstruction. Additionally, the NID-PD, combined with an adaptive sampling strategy, accelerates the denoising process by shortening the diffusion steps, achieving an effective balance between performance and computational efficiency. Despite the performance improvements from noise intensity priors, our framework faces challenges such as reliance on pre-learned priors and potential overfitting. Future research could focus on adaptive end-to-end methods for obtaining noise priors, improving training scalability, and enhancing generalization to new noise types.

CRediT authorship contribution statement

Meiqin Liu: Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition. **Xuan Long:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology. **Qi Tang:** Writing – review & editing, Methodology. **Chao Yao:** Writing – review & editing, Writing – original draft, Resources, Methodology. **Yao Zhao:** Writing – review & editing, Supervision, Resources, Project administration.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used ChatGPT to polish the manuscript and enhance its readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grant 62372036, Grant 62120106009, Grant 62332017 and Grant U22A2022.

Data availability

My training set and test set are public data sets, both of which are revealed in public websites.

References

[1] K. Zhang, W. Zuo, Y. Chen, D. Meng, L. Zhang, Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising, IEEE Trans. Image Process. 26 (7) (2017) 3142–3155.
[2] J. Liang, J. Cao, G. Sun, K. Zhang, L.V. Gool, R. Timofte, SwinIR: Image restoration using swin transformer, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2021, pp. 1833–1844.
[3] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, L.V. Gool, Diffir: efficient diffusion model for image restoration, in: In Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2023, pp. 13095–13105.
[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: In Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2022, pp. 10684–10695.
[5] X. Long, M. Liu, Q. Tang, C. Yao, J. Jin, Y. Zhao, Noisy-residual continuous diffusion models for real image denoising, in: In Proceedings of the IEEE International Conference on Multimedia and Expo, ICME, 2024, pp. 1–6.

[6] S. Zamir, A. Arora, S. Khan, M. Hayat, F. Khan, M.-H. Yang, L. Shao, Learning enriched features for real image restoration and enhancement, in: In Proceedings of the European Conference on Computer Vision, ECCV, 2020, pp. 492–511.
[7] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T.-S. Chua, Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning, in: In Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2017, pp. 5659–5667.
[8] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, J. Wu, UNet 3+: A full-scale connected unet for medical image segmentation, in: In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 1055–1059.
[9] S. Zamir, A. Arora, S. Khan, M. Hayat, F. Khan, M.-H. Yang, Restormer: Efficient transformer for high-resolution image restoration, in: Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2022, pp. 5728–5739.
[10] X. Chen, H. Li, M. Li, J. Pan, Learning a sparse transformer network for effective image deraining, in: In Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2023, pp. 5896–5905.
[11] Y. Huang, H. Huang, Beyond image prior: Embedding noise prior into latent space of conditional denoising transformer, Int. J. Comput. Vis. 133 (11) (2025) 7591–7611.
[12] A. Gautam, A. Pawar, A. Joshi, S. Tazi, S. Chaudhary, P. Hambarde, A. Dudhane, S. Vipparthi, S. Murala, Pureformer: Transformer-based image denoising, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2025, pp. 1–9.
[13] X. Zhao, C. Zhao, X. Hu, H. Zhang, Y. Tai, J. Yang, Learning multi-scale spatial-frequency features for image denoising, Pattern Recognit. 172 (2026) 112300.
[14] X. Lin, J. He, Z. Chen, Z. Lyu, B. Dai, F. Yu, Y. Qiao, W. Ouyang, C. Dong, DiffBIR: Toward blind image restoration with generative diffusion prior, in: In Proceedings of the European Conference on Computer Vision, ECCV, 2025, pp. 430–448.
[15] Y. Wang, J. Yu, J. Zhang, Zero-shot image restoration using denoising diffusion null-space model, in: In Proceedings of the International Conference on Learning Representations, ICLR, 2023, pp. 1–31.
[16] C. Yang, C. Wang, L. Liang, Z. Su, Real-world image denoising via efficient diffusion model with controllable noise generation, J. Electron. Imaging 33 (4) (2024) 043003.
[17] L. Wang, Q. Yang, C. Wang, W. Wang, Z. Su, Coarse-to-fine mechanisms mitigate diffusion limitations on image restoration, Comput. Vis. Image Underst. 248 (2024) 1–12.
[18] J. Liu, Q. Wang, H. Fan, Y. Wang, Y. Tang, L. Qu, Residual denoising diffusion models, in: In Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2024, pp. 2773–2783.
[19] T. Li, H. Feng, L. Wang, L. Zhu, Z. Xiong, H. Huang, Stimulating diffusion model for image denoising via adaptive embedding and ensembling, 2024,
[20] J. Wu, S. Pan, N. Li, B. Chen, B. An, Z. Wang, Y. Wang, S.-T. Xia, Universal image restoration via task-adaptive diffusion degradation oriented model, Pattern Recognit. 176 (2026) 113193.
[21] M. Xue, J. He, S. Palaiahnakote, M. Zhou, Unified image restoration and enhancement: Degradation calibrated cycle reconstruction diffusion model, Pattern Recognit. 171 (2026) 112073.
[22] T. Wang, K. Zhang, Y. Zhang, W. Luo, B. Stenger, T. Lu, T.-K. Kim, W. Liu, Lldiffusion: Learning degradation representations in diffusion models for low-light image enhancement, Pattern Recognit. 166 (2025) 111628.
[23] J. Liu, J. Jin, X. Xiu, J. Zhang, W. Liu, STAR-Net: an interpretable model-aided network for remote sensing image denoising, Pattern Recognit. 172 (2026) 112496.
[24] J. Ba, J. Kiros, G. Hinton, L. normalization, In Proceedings of the Advances in Neural Information Processing Systems, NIPS, 2016.
[25] A. Abdelhamed, S. Lin, M. Brown, A high-quality denoising dataset for smartphone cameras, in: In Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2018, pp. 1692–1700.
[26] T. Plotz, S. Roth, Benchmarking denoising algorithms with real photographs, in: In Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2017, pp. 1586–1595.
[27] S. Nam, Y. Hwang, Y. Matsushita, S. Kim, A holistic approach to cross-channel image noise modeling and its application to image denoising, in: In Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2016, pp. 1683–1691.
[28] P. Charbonnier, L. Blanc-Feraud, G. Aubert, M. Barlaud, Two deterministic half-quadratic regularization algorithms for computed imaging, in: In Proceedings of the IEEE International Conference on Image Processing (ICIP), Vol. 2, 1994, pp. 168–172.
[29] S. Guo, Z. Yan, K. Zhang, W. Zuo, L. Zhang, Toward convolutional blind denoising of real photographs, in: In Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2019, pp. 1712–1722.
[30] S. Anwar, N. Barnes, Real image denoising with feature attention, in: In Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2019, pp. 3155–3164.
[31] Z. Yue, Q. Zhao, L. Zhang, D. Meng, Dual adversarial network: Toward real-world noise removal and noise generation, in: In Proceedings of the European Conference on Computer Vision, ECCV, 2020, pp. 41–58.
[32] S. Zamir, A. Arora, S. Khan, M. Hayat, F. Khan, M.-H. Yang, L. Shao, CycleISP: Real image restoration via improved data synthesis, in: In Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2020, pp. 2696–2705.
[33] S. Zamir, A. Arora, S. Khan, M. Hayat, F. Khan, M.-H. Yang, L. Shao, Multi-stage progressive image restoration, in: In Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2021, pp. 14821–14831.
[34] S. Cheng, Y. Wang, H. Huang, D. Liu, H. Fan, S. Liu, NBNet: Noise basis learning for image denoising with subspace projection, in: In Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2021.
[35] H. Shen, Z.-Q. Zhao, W. Zhang, Adaptive dynamic filtering network for image denoising, in: In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Vol. 37, 2023, pp. 2227–2235.
[36] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, H. Li, Uformer: A general U-shaped transformer for image restoration, in: In Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2022, pp. 17683–17693.
[37] B. Kawar, M. Elad, S. Ermon, J. Song, Denoising diffusion restoration models, in: In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vol. 35, 2022, pp. 23593–23606.
[38] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: In Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2022, pp. 10684–10695.
[39] Y. Wang, L. Li, T. Xue, J. Gu, Reconstruct-and-generate diffusion model for detail-preserving image denoising, 2023, pp. 1–14, arXiv preprint arXiv:2309.10714.
[40] I. Loshchilov, F. Hutter, SGDR: Stochastic gradient descent with warm restarts, in: In Proceedings of the International Conference on Learning Representations, ICLR, 2017.
[41] Y. Zhang, B. Ji, J. Hao, A. Yao, Perception-distortion balanced ADMM optimization for single-image super-resolution, in: In Proceedings of the European Conference on Computer Vision, ECCV, 2022, pp. 108–125.
[42] Y. Zhang, D. Li, X. Shi, D. He, K. Song, X. Wang, H. Qin, H. Li, KBNet: Kernel basis network for image restoration, 2023, pp. 1–15, arXiv preprint arXiv:2303.02881.
[43] L. Chen, X. Chu, X. Zhang, J. Sun, Simple baselines for image restoration, in: In Proceedings of the European Conference on Computer Vision, ECCV, Springer, 2022, pp. 17–33.
[44] C. Ren, X. He, C. Wang, Z. Zhao, Adaptive consistency prior based deep network for image denoising, in: In Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2021, pp. 8596–8606.

Meiqin Liu received the M.E. degree and Ph.D. degree from Beijing Jiaotong University (BJTU), China, in 2007 and 2018, respectively. From 2014 to 2015, she was a Visiting Scholar at Simon Fraser University (SFU), Canada. She is currently a professor at the Institute of Information Science, BJTU. Her research interests include image/video compression and video processing.

Xuan Long received the B.S. degree from Hunan University of Science and Technology (HNUST), China, in 2022. She is currently pursuing the M.E. degree at the Institute of Information Science, Beijing Jiaotong University (BJTU), China. Her research interests include image denoising and diffusion model.

Qi Tang received the B.S. degree from Beijing Jiaotong University (BJTU), China, in 2021. He is currently pursuing the M.E. degree at the Institute of Information Science, Beijing Jiaotong University (BJTU), China. His research interests include super-resolution and diffusion model.

Chao Yao received the M.E. degree and Ph.D. degree from Beijing Jiaotong University (BJTU) in 2010 and 2016. From 2014 to 2015, he was a Visiting Ph.D. student at Swiss Federal Institute of Technology (EPFL). He is currently a professor with University of Science and Technology Beijing (USTB), researching video compression and human–computer interaction.

Yao Zhao received the B.S. degree from the Radio Engineering Department, Fuzhou University, Fuzhou, China, in 1989, the M.E. degree from the Radio Engineering Department, Southeast University, Nanjing, China, in 1992, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996. He became an Associate Professor with BJTU in 1998, where he became a Professor in 2001. From 2001 to 2002, he was a Senior Research Fellow with the Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands. He is currently the Director at the Institute of Information Science, BJTU. His current research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding. He is a Fellow of the IET and IEEE.