PET-DSCFORMER: EFFICIENT DEGRADATION-AWARE TRANSFORMER FOR REAL-WORLD IMAGE DENOISING

Anonymous authors
Paper under double-blind review

ABSTRACT

Image denoising is a fundamental task aiming to recover clean images from real-world degradations while preserving structural fidelity and fine-grained details. Recent CNN- and Transformer-based restoration methods have achieved remarkable progress by improving local representation learning and global context modeling. However, existing approaches still face three major limitations: dense Transformer attention is computationally expensive for high-resolution restoration, conventional skip fusion may propagate unreliable degraded features into the decoder, and most methods predict restoration residuals in a deterministic manner without explicitly modeling the uncertainty and complementarity of multi-scale residual estimates. To address these limitations, we propose PET-DSCFormer, a posterior error transport restoration framework for real-world image denoising and general image restoration. Extensive experiments on multiple public real-world denoising datasets, including medical imaging and natural image benchmarks, demonstrate that PET-DSCFormer achieves competitive or superior restoration performance compared with state-of-the-art methods. Moreover, experiments on GoPro deblurring further validate the generalization capability of the proposed framework beyond denoising. The code will be publicly available on GitHub upon acceptance.

1 INTRODUCTION

Image denoising aims to recover a clean image from a noisy observation and remains a fundamental problem in low-level vision, computational photography, and medical imaging. Its difficulty lies not only in suppressing pixel-wise perturbations, but also in recovering weak edges, thin structures, fine textures, and long-range dependencies that are easily corrupted by spatially non-uniform noise. Existing denoising methods can be broadly categorized into several paradigms. Supervised CNN- or Transformer-based restorers learn powerful mappings from paired data, but may overfit the degradation statistics of the training set and degrade under shifted noise Zamir et al. (2021); Chen et al. (2022); Zamir et al. (2022); Zhao et al. (2023); Zhang et al. (2017; 2018); Guo et al. (2019); Anwar & Barnes (2019); Chen et al. (2021a); Liang et al. (2021); Wang et al. (2022); Tu et al. (2022); Chen et al. (2021b); Zamir et al. (2020). Self-supervised and zero-shot methods reduce the reliance on clean targets, yet often depend on restrictive noise assumptions, target-domain data, or costly test-time optimization Tu et al. (2025); Quan et al. (2025); Herbreteau & Meunier; Lehtinen et al. (2018); Krull et al. (2019); Batson & Royer (2019); Huang et al. (2021); Chen et al. (2023b;a). Transformer-based models improve global modeling, but dense attention or dense feature aggregation is still expensive for high-resolution restoration and usually treats all spatial locations as equally reliable Zamir et al. (2022); Zhao et al. (2023); Yang et al. (2022); Chen et al. (2021a); Liang et al. (2021); Wang et al. (2022); Tu et al. (2022). Recent foundation-model and diffusion-based approaches introduce strong priors, but their heavy pretrained models, iterative sampling, or hallucination tendency are not always suitable for efficiency-sensitive denoising where structural fidelity is more important than perceptual imagination Cheng et al. (2024); Kim et al. (2024; 2025); Yin et al. (2025); Tu et al. (2025); Quan et al. (2025); Herbreteau & Meunier; Garber & Tirer (2024); Ho et al. (2020); Kawar et al. (2022); Chung et al. (2022); Fei et al. (2023); Potlapalli et al. (2023); Conde et al. (2024). These limitations suggest that high-quality denoising requires more than a stronger backbone. A

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

practical restoration model should satisfy three requirements. First, it should model global structure efficiently while preventing severely degraded regions from dominating global statistics. Second, it should fuse encoder and decoder features according to their reliability, because shallow features contain high-resolution details but may also carry noise, whereas deep features provide stable structures but may lose local precision. Third, it should exploit the coarse-to-fine nature of U-shaped decoders. Coarse stages usually provide stable residual trends, while fine stages preserve edges and textures but may amplify residual noise. However, most existing networks still use heuristic skip fusion and a single deterministic residual head, leaving the uncertainty and complementarity of multi-scale residual estimates underexplored.

Motivated by these observations, we propose PET-DSCFormer, a degradation-aware posterior restoration framework for efficient and structure-preserving image denoising. For global modeling, we introduce Degradation-aware Structural Coreset Attention (DSCA). Instead of computing attention densely over all spatial locations, DSCA predicts a degradation confidence map and softly aggregates dense features into compact structural coreset tokens. Attention is then computed on these tokens and broadcast back to the dense feature map. Since the number of coresets is much smaller than the number of pixels, DSCA reduces the complexity of global modeling. More importantly, degradation-aware weighting encourages global context to be estimated from reliable structures rather than from corrupted responses. To make decoding reliable, PET-DSCFormer further introduces Precision-aware Skip Fusion (PSF). PSF interprets encoder and decoder features as two noisy observations of the same latent clean feature and predicts their precisions adaptively. The fused representation is obtained by posterior precision weighting, so the model preserves shallow details when they are reliable and relies more on deep structural evidence when shallow features are contaminated. Finally, we develop a posterior noise calibration mechanism composed of Posterior Error Transport (PET), a High-frequency Detail Residual Expert (HDRE), and Mixture Posterior Calibration (MPC). PET predicts residual distributions at multiple decoder scales and fuses them through closed-form Gaussian posterior estimation, transporting stable coarse residuals toward fine scales. HDRE supplements conservative high-frequency residuals with input-aware local cues and spatial gating, while MPC predicts multiple correction candidates with adaptive posterior weights and constrains calibration to a high-frequency subspace. Together, these modules improve local structural fidelity without disturbing the recovered low-frequency content.

Our contributions are summarized as follows:

1) We propose Degradation-aware Structural Coreset Attention, which performs global context modeling over compact degradation-weighted structural coresets, reducing attention cost while suppressing unreliable degraded regions.

2) We introduce Precision-aware Skip Fusion, a probabilistic skip fusion module that combines encoder details and decoder structures according to predicted precisions, preventing noisy shallow features from being blindly propagated.

3) We present a posterior noise calibration mechanism consisting of PET, HDRE, and MPC, which formulates multi-scale residual recovery as posterior inference, restores high-frequency structures conservatively, and performs mixture-based local calibration for ambiguous texture and edge regions.

2 Related Works

CNN-based restorers remain competitive because of their locality, efficiency, and stable optimization Buades et al. (2005); Dabov et al. (2007); Zhang et al. (2017; 2018); Guo et al. (2019); Anwar & Barnes (2019); Chen et al. (2021b); Zamir et al. (2020), ranging from the progressive MPRNet Zamir et al. (2021), activation-free NAFNet Chen et al. (2022), and frequency-modulated PoolNet Cui et al. (2025) to sensor- and deployment-oriented methods such as PMRID Wang et al. (2020), SplitterNet Flepp et al. (2024a), and CANS Ma et al. (2025). Transformer-based approaches further improve long-range modeling Chen et al. (2021a); Liang et al. (2021); Wang et al. (2022); Tu et al. (2022): Restormer Zamir et al. (2022) introduces efficient channel-wise attention, CODE Zhao et al. (2023) condenses pixels into superpixel-level representations, and SIST Yang et al. (2022) incorporates sinogram-specific priors. Generalizable and adaptive methods Cheng et al. (2024); Kim et al. (2024; 2025); Yin et al. (2025); Chen et al. (2023b;a), including CLIPDenoising Cheng et al. (2024),

LAN Kim et al. (2024), IDF Kim et al. (2025), and RNINet Yin et al. (2025), enhance robustness through pretrained features, test-time adaptation, dynamic filtering, or feature regularization. Meanwhile, self-supervised, zero-shot, and score-based methods Tu et al. (2025); Quan et al. (2025); Herbreteau & Meunier; Lehtinen et al. (2018); Krull et al. (2019); Batson & Royer (2019); Huang et al. (2021), such as C2S Tu et al. (2025), Implicit Neural Sampling Quan et al. (2025), and CENSURE Herbreteau & Meunier, reduce reliance on paired clean targets. Foundation-model, generative, and all-in-one frameworks Garber & Tirer (2024); Zheng et al. (2026); Tu et al. (2026); Rajagopalan et al. (2025); Rajagopalan & Patel (2025); Wu et al. (2025); Yang et al. (2025); Ho et al. (2020); Kawar et al. (2022); Chung et al. (2022); Fei et al. (2023); Potlapalli et al. (2023); Conde et al. (2024); Li et al. (2022); Guo et al. (2024a;b), including DDPG Garber & Tirer (2024), SubDAPS Zheng et al. (2026), TPGDiff Tu et al. (2026), RestoreVAR Rajagopalan et al. (2025), AWRaCLe Rajagopalan & Patel (2025), TUR Wu et al. (2025), and SCAN-PhysFed Yang et al. (2025), further exploit generative priors, contextual information, adaptive optimization, and physics-informed learning. Despite these advances, most methods still treat global aggregation, feature fusion, and residual prediction as deterministic operations. PET-DSCFormer instead introduces reliability and uncertainty modeling through degradation-aware structural attention, precision-aware skip fusion, and posterior residual calibration. A complete discussion of related work is provided in the Appendix.

3 METHODOLOGY

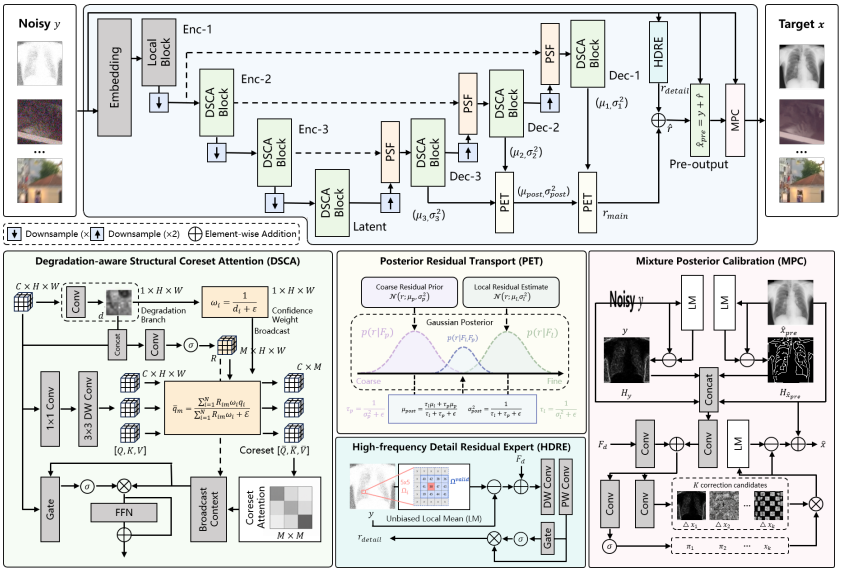

Figure 1: The pipeline of our proposed PET-DSCFormer for image denoising.

3.1 OVERALL ARCHITECTURE

Given a degraded image $y \in \mathbb{R}^{H\times W\times C_m}$, the goal of image restoration is to recover the clean image x. We formulate restoration as a residual estimation problem:

$$\widehat{x} = y + \widehat{r}, \tag{1}$$

where $\widehat{r}$ denotes the predicted restoration residual. The shallow embedding first maps the degraded input into feature space:

$$F_0 = \phi_{\text{emb}}(y), \tag{2}$$

where ϕ_{emb} is a 3×3 convolution. The encoder extracts hierarchical features:

$$F_1, F_2, F_3, F_4 = \text{Encoder}(F_0), \tag{3}$$

where F_1 is produced by lightweight local restoration blocks, and F_2, F_3, and F_4 are produced by DSCA transformer blocks. The decoder progressively upsamples the latent representation and fuses encoder features through PSF. At different decoder scales, residual distribution heads predict residual means and variances, which are fused by PET. Finally, an HDRE and an MPC block refine the output.

3.2 Degradation-aware Structural Coreset Attention

Given an input feature map $X \in \mathbb{R}^{C\times H\times W}$, we denote $N = H \times W$ as the number of spatial positions. For simplicity, the spatial index is written as $i \in \{1, \dots, N\}$. The feature vector at position i is denoted as $X_i \in \mathbb{R}^C$. We first generate query, key, and value features from X:

$$[Q, K, V] = \phi_{\mathrm{dw}}(\phi_{\mathrm{qkv}}(X)), \tag{4}$$

where ϕ_{qkv} is a shared 1×1 convolution that maps X to $3C$ channels, and ϕ_{dw} is a depth-wise 3×3 convolution that introduces local context before global aggregation while keeping the parameter cost low. The corresponding query, key, and value vectors at position i are denoted as $q_i, k_i, v_i \in \mathbb{R}^C$.

For degraded image restoration, different spatial locations are not equally reliable. Regions with stronger noise, blur, artifacts, or uncertain structures should contribute less to global structural statistics. Therefore, DSCA predicts a degradation map:

$$d = \phi_{\mathrm{deg}}(X), \tag{5}$$

where ϕ_{deg} is a lightweight convolutional branch, and $d \in \mathbb{R}^{1\times H\times W}$. The scalar d_i indicates the estimated degradation strength at position i. A smaller d_i means the feature is considered more reliable, while a larger d_i means the feature is less reliable. We convert the degradation estimate into a confidence weight:

$$\omega_i = \frac{1}{d_i + \epsilon}. \tag{6}$$

Thus, reliable regions receive larger weights during coreset aggregation, while strongly degraded regions are down-weighted.

Instead of using learnable spatial tokens or explicit prototype parameters, DSCA predicts the coreset assignment directly from the current image feature. First, we concatenate the feature map and the degradation map, then a lightweight 1×1 convolution maps the concatenated feature $[X, d] \in \mathbb{R}^{(C+1)\times H\times W}$ to a compact structural descriptor $Z = \phi_{\mathrm{desc}}([X, d])$, where ϕ_{desc} is simply a 1×1 convolution, and C_d is the descriptor dimension, such as 32. The descriptor $Z_i \in \mathbb{R}^{C_d}$ describes the local structural and degradation-aware state of position i. Next, another 1×1 convolution predicts the assignment logits:

$$A = \phi_{\mathrm{assign}}(Z), \tag{7}$$

where $A \in \mathbb{R}^{M\times H\times W}$ and M is the number of structural coreset tokens. The scalar A_{im} represents the assignment logit between spatial position i and the m-th coreset. The soft assignment is then computed as:

$$R_{im} = \frac{\exp(A_{im}/\tau)}{\sum_{j=1}^{M} \exp(A_{ij}/\tau)}, \tag{8}$$

where τ is a temperature parameter. After obtaining the assignment map R, DSCA aggregates dense Q, K, and V features into M compact coreset tokens. For the m-th coreset, the query token is:

$$\overline{q}_m = \frac{\sum_{i=1}^{N} R_{im}\omega_i q_i}{\sum_{i=1}^{N} R_{im}\omega_i + \epsilon}. \tag{9}$$

Similarly, the key token $\overline{k}_m$ and value token $\overline{v}_m$ are obtained. Here, R_{im} determines how much position i contributes to the m-th structural coreset, and ω_i controls the reliability of that contribution, which means the coreset token is a degradation-aware weighted structural summary. After aggregation, we perform attention only among the M coreset tokens:

$$\overline{G} = \mathrm{Softmax}\left(\frac{\overline{Q}\,\overline{K}^{\mathsf{T}}}{\sqrt{d_h}}\right)\overline{V}. \tag{10}$$

where $\overline{Q}$, $\overline{K}$, and $\overline{V}$ are the coreset query, key, and value tokens within each attention head. The attention cost becomes $\mathcal{O}(M^2C)$ since attention is computed among M tokens, instead of the dense spatial attention cost $\mathcal{O}(N^2C)$. Because $M \ll N$, this greatly reduces the cost of global context modeling.

The updated coreset context $\overline{G}$ is then broadcast back to dense spatial positions using the same assignment map:

$$g_i = \sum_{m=1}^{M} R_{im}\overline{g}_m. \tag{11}$$

This reconstructs a dense context feature map $G \in \mathbb{R}^{C\times H\times W}$. Finally, DSCA uses a gate to fuse the original feature and the coreset context and controls how much global structural context should be injected at each spatial location:

$$Y = \sigma(\phi_{\mathrm{gate}}([X, G])) \odot \phi_{\mathrm{out}}(G). \tag{12}$$

The full DSCA block follows a residual Transformer-style design, and the feed-forward network is a gated depth-wise convolutional feed-forward network, which further enhances local restoration capability while keeping the block lightweight.

3.3 Precision-aware Skip Fusion

In U-shaped restoration networks, skip connections are important because shallow encoder features preserve high-resolution edges, textures, and spatial details, while deeper decoder features contain stronger semantic and global structural information. A conventional U-Net usually concatenates encoder and decoder features directly:

$$F_{\mathrm{fuse}} = \mathrm{Conv}([F_e, F_d]), \tag{13}$$

where F_e denotes the encoder skip feature and F_d denotes the decoder feature. However, in image restoration, shallow encoder features may preserve useful details but also carry noise, artifacts, or blur. Direct concatenation may therefore propagate unreliable degraded information into the decoder. To address this issue, PSF treats encoder and decoder features as two noisy observations of the same latent clean feature Z:

$$F_e = Z + \varepsilon_e, \qquad F_d = Z + \varepsilon_d. \tag{14}$$

Here, ε_e and ε_d denote the estimation errors from the encoder and decoder branches, respectively. Assuming that both errors follow zero-mean Gaussian distributions:

$$\varepsilon_e \sim \mathcal{N}(0, \sigma_e^2), \qquad \varepsilon_d \sim \mathcal{N}(0, \sigma_d^2), \tag{15}$$

their corresponding precision values τ_e and τ_d are:

$$\tau_e = \frac{1}{\sigma_e^2 + \epsilon}, \qquad \tau_d = \frac{1}{\sigma_d^2 + \epsilon}. \tag{16}$$

In practice, τ_e and τ_d are adaptively predicted from the features using lightweight convolutional branches:

$$\tau_e = \phi_e(F_e), \qquad \tau_d = \phi_d(F_d), \tag{17}$$

where ϕ_e and ϕ_d consist of depth-wise convolution, 1×1 convolution, and Softplus activation to ensure positive precision values. To align the channel dimensions, the encoder and decoder features are first projected by 1×1 convolutions, obtaining $\overline{F}_e$ and $\overline{F}_d$. The fused feature is then obtained through precision-weighted posterior estimation:

$$F_{\text{fuse}} = \frac{\tau_e \overline{F}_e + \tau_d \overline{F}_d}{\tau_e + \tau_d + \epsilon}. \tag{18}$$

From a probabilistic perspective, this formula corresponds to the minimum mean square error estimate of two Gaussian observations. When the encoder feature is more reliable, τ_e becomes larger and the model preserves more shallow details. When the decoder feature is more reliable, τ_d becomes larger and the model relies more on deeper structural information.

3.4 Posterior Residual Transport and Calibration

Multi-scale Posterior Error Transport. Given the feature F_s at decoder scale s, the Residual Distribution Head predicts the residual mean and variance:

$$\mu_s = \phi_\mu^s(F_s), \tag{19}$$

$$\sigma_s^2 = \exp(\phi_\sigma^s(F_s)) + \epsilon. \tag{20}$$

where μ_s denotes the predicted residual mean at scale s, and σ_s^2 denotes its predicted uncertainty. Thus, the residual at scale s is modeled as:

$$r_s \sim \mathcal{N}(\mu_s, \sigma_s^2). \tag{21}$$

In the coarse-to-fine decoder, the coarse-scale residual can be viewed as the prior for the fine-scale residual. Let the local residual distribution at the current scale be $r_l \sim \mathcal{N}(\mu_l, \sigma_l^2)$, and the upsampled prior residual distribution from the previous coarser scale be $r_p \sim \mathcal{N}(\mu_p, \sigma_p^2)$. Here, μ_p and σ_p^2 are obtained by upsampling the residual mean and variance from the coarser scale. According to the product of Gaussian experts, the posterior distribution is:

$$p(r \mid F_l, F_p) \propto \mathcal{N}(r; \mu_l, \sigma_l^2)\mathcal{N}(r; \mu_p, \sigma_p^2). \tag{22}$$

This posterior is still a Gaussian distribution, and its closed-form solution is:

$$\mu_{\text{post}} = \frac{\tau_l \mu_l + \tau_p \mu_p}{\tau_l + \tau_p + \epsilon}, \qquad \sigma_{\text{post}}^2 = \frac{1}{\tau_l + \tau_p + \epsilon}, \tag{23}$$

where $\tau_l = 1/(\sigma_l^2 + \epsilon)$ and $\tau_p = 1/(\sigma_p^2 + \epsilon)$. Therefore, the posterior mean is a precision-weighted average of the local residual estimate and the coarse-scale prior residual estimate. If the current-scale residual is more reliable, τ_l becomes larger and the model relies more on the current scale. If the coarse-scale prior is more reliable, τ_p becomes larger and the model preserves more global structural information from the coarser scale. The final main residual is defined as $r_{\text{main}} = \mu_1$, where μ_1 is the local residual distribution at the top decoder (level 1) fused with the posterior residual from level 2, and so on.

High-frequency Detail Residual Expert. The unbiased local mean of the input image is computed as:

$$\text{LM}(y)_i = \frac{\sum_{j \in \Omega_i \cap \Omega^{\text{valid}}} y_j}{|\Omega_i \cap \Omega^{\text{valid}}|}, \tag{24}$$

where Ω_i denotes a local window centered at position i, and Ω^{valid} denotes the valid image region. Unlike ordinary average pooling, $\text{LM}(\cdot)$ does not include padded regions in the averaging process, which avoids local mean bias near image boundaries. The high-frequency component of the input image is defined as:

$$H_y = y - \text{LM}(y). \tag{25}$$

Then, the high-resolution decoder feature F_d and the input high-frequency component H_y are projected into the same feature space and then processed by depth-wise convolution and point-wise convolution for local detail modeling:

$$F_{\text{detail}} = \phi_{\text{detail}}(\phi_f(F_d) + \phi_h(H_y)). \tag{26}$$

To avoid enhancing high-frequency noise in smooth regions, the module further predicts a spatial gate $g_{\mathrm{detail}} = \sigma(\phi_g(F_{\mathrm{detail}}))$, and the detail residual is computed as:

$$r_{\mathrm{detail}} = g_{\mathrm{detail}} \odot \phi_o(F_{\mathrm{detail}}). \tag{27}$$

After obtaining the main residual r_{main}, the preliminary output residual is:

$$\widehat{r} = r_{\mathrm{main}} + r_{\mathrm{detail}}. \tag{28}$$

The preliminary restored image is then $\widehat{x}_{\mathrm{pre}} = y + \widehat{r}$.

In the implementation, the detail branch also contains a conservative local pooling refinement module. It first estimates the low-frequency component through local average pooling:

$$F_{\mathrm{low}} = \mathrm{LM}(F_{\mathrm{detail}}), \tag{29}$$

and extracts the feature-level high-frequency component $F_{\mathrm{hf}} = F_{\mathrm{detail}} - F_{\mathrm{low}}$. Then, lightweight convolutions refine F_{hf}, and a confidence gate $G_{\mathrm{hf}} = \sigma(\phi_{\mathrm{conf}}([F_{\mathrm{detail}}, |F_{\mathrm{hf}}|]))$ controls the injection strength:

$$F_{\mathrm{refine}} = F_{\mathrm{detail}} + \alpha\, G_{\mathrm{hf}} \odot \phi_{\mathrm{refine}}(F_{\mathrm{hf}}). \tag{30}$$

Here, α is a learnable scaling parameter initialized to zero, making this refinement module equivalent to an identity mapping at the beginning of training. This zero-initialized residual design prevents the newly introduced module from disrupting the original restoration output.

Mixture Posterior Calibration. Given the preliminary restored image $\widehat{x}_{\mathrm{pre}} = y + \widehat{r}$ and the input image y, their local high-frequency components are first computed:

$$H_{\widehat{x}_{\mathrm{pre}}} = \widehat{x}_{\mathrm{pre}} - \mathrm{LM}(\widehat{x}_{\mathrm{pre}}), \tag{31}$$

$$H_y = y - \mathrm{LM}(y). \tag{32}$$

Then, a structural calibration descriptor $D = [\widehat{x}_{\mathrm{pre}}, y, H_{\widehat{x}_{\mathrm{pre}}}, H_y]$ containing the preliminary restored image, the original input image, and their high-frequency structural information is constructed. It is then fused with the final decoder feature F_d:

$$F_c = \phi_f(F_d) + \phi_i(D), \tag{33}$$

where ϕ_f and ϕ_i are lightweight convolutional mappings. Based on the fused feature F_c, K correction candidates Δx_k and corresponding posterior weights π_k are predicted:

$$[\Delta x_1, \Delta x_2, \ldots, \Delta x_K] = \phi_{\mathrm{cand}}(F_c), \tag{34}$$

$$[\pi_1, \pi_2, \ldots, \pi_K] = \phi_{\mathrm{logits}}(F_c). \tag{35}$$

The posterior weights are normalized by softmax and therefore satisfy $\sum_{k=1}^{K} \pi_k = 1$. The mixture correction is:

$$\Delta x = \sum_{k=1}^{K} \pi_k \Delta x_k. \tag{36}$$

To avoid disrupting the low-frequency brightness structure already recovered by Posterior Error Transport, the mixture correction is projected into a local high-frequency subspace:

$$\Delta x_{\mathrm{hf}} = \Delta x - \mathrm{LM}(\Delta x). \tag{37}$$

This constraint encourages the calibration to mainly correct edges, textures, and local structures rather than arbitrarily changing the local image mean. The final output is:

$$\widehat{x} = \widehat{x}_{\mathrm{pre}} + \alpha \Delta x_{\mathrm{hf}}, \tag{38}$$

where α is a learnable scalar initialized to zero. As training progresses, the model automatically learns whether and to what extent the mixture posterior correction should be introduced.

4 EXPERIMENTS

4.1 DATASETS

All datasets used in this work are publicly available benchmarks collected from real-world imaging scenarios. We evaluate PET-DSCFormer on LDCT Moen et al. (2021) dataset, M4Raw Lyu et al. (2023), JSRT Shiraishi et al. (2000), SIDD Abdelhamed et al. (2018), PolyU Xu et al. (2018), MIDD Flepp et al. (2024b), LRD Zhang et al. (2023), PMRID Wang et al. (2020), and GoProNah et al. (2017), covering medical, natural, RAW, low-light, and motion-deblurring restoration tasks; detailed dataset descriptions and statistics are provided in the Appendix.

4.2 IMPLEMENTATION DETAILS

Detailed implementation, preprocessing, architectural, and training configurations are provided in the Appendix.

For evaluation, we reported Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). PSNR and SSIM were computed between the restored image and the clean reference in the normalized image intensity range. We used 10-fold cross-validation, with each metric reported as a mean value. All experiments were conducted using the predefined train, validation, and test splits, and the external validation sets.

4.3 COMPARISON WITH OTHER EXISTING WORKS

We present the experimental denoising results of the proposed methods, in comparison to a set of SOTA methods that include both foundational classics and recent top-conference publications. These benchmark methods are: NAFNet Chen et al. (2022), Restormer Zamir et al. (2022), CODE Zhao et al. (2023), CLIPDenoising Cheng et al. (2024), LAN Kim et al. (2024), SplitterNet Flepp et al. (2024a), TUR Wu et al. (2025), CANS Ma et al. (2025), IDF Kim et al. (2025), PoolNet Cui et al. (2025), RNINet Yin et al. (2025), and CENSURE Herbreteau & Meunier. To ensure a fair comparison, all baseline methods are evaluated from the experiments conducted using their official codes under the same settings.

table 1 shows that our method achieves the best PSNR on all six medical datasets and the best SSIM on most, with particularly strong performance on JSRT. The consistent gains across LDCT and M4Raw, together with fig. 8, demonstrate robust cross-modality generalization and improved anatomical structure preservation.

	LDCT-Chest		**LDCT-Abdomen**		**M4Raw-T1**		**M4Raw-T2**		**M4Raw-FLAIR**		**JSRT**	
	PSNR ↑	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**
NAFNet (ECCV'2022)	30.18	0.6447	46.04	0.9790	35.11	0.9270	34.50	0.9091	34.33	0.9071	27.22	0.8296
Restormer (CVPR'2022)	<u>30.32</u>	0.6443	46.05	0.9798	<u>35.17</u>	<u>0.9278</u>	<u>34.52</u>	<u>0.9091</u>	<u>34.38</u>	<u>0.9078</u>	27.76	0.8565
CODE (CVPR'2023)	29.98	0.6405	42.77	0.9656	34.83	0.9237	34.21	0.9049	34.00	0.9039	25.64	0.7952
CLIPDenoising (CVPR'2024)	29.99	0.6368	43.03	0.9746	35.08	0.9269	34.41	0.9078	34.29	0.9065	23.00	0.8389
LAN (CVPR'2024)	<u>30.32</u>	0.6430	45.86	0.9780	35.15	0.9271	34.50	0.9084	<u>34.38</u>	0.9072	26.02	0.8543
SplitterNet (CVPR'2024)	29.31	0.6328	44.21	0.9718	35.01	0.9258	34.30	0.9062	34.17	0.9057	21.56	0.5508
TUR (AAAI'2025)	30.16	0.6459	<u>46.06</u>	<u>0.9799</u>	35.16	0.9277	34.49	0.9090	34.37	0.9077	<u>27.87</u>	0.8617
CANS (CVPR'2025)	30.06	0.6385	42.93	0.9694	34.98	0.9254	34.31	0.9067	34.15	0.9052	25.99	0.7882
IDF (ICCV'2025)	29.20	0.5800	44.52	0.9737	34.67	0.9139	33.81	0.8919	33.82	0.8888	18.75	0.6972
PoolNet (TIP'2025)	30.13	**0.6565**	45.92	0.9799	35.14	0.9274	34.44	<u>0.9091</u>	34.25	0.9071	27.73	<u>0.8804</u>
RNINet (ICLR'2025)	29.18	0.6297	45.38	0.9788	34.83	0.9250	34.20	0.9057	34.22	0.9061	27.40	0.7467
CENSURE (ICLR'2026)	29.67	0.6319	43.60	0.9754	34.38	0.9035	33.49	0.8740	32.79	0.8411	24.69	0.7407
Ours	**30.41**	<u>0.6508</u>	**46.26**	**0.9804**	**35.21**	**0.9283**	**34.55**	**0.9093**	**34.42**	**0.9081**	**28.85**	**0.8861**

Table 1: Quantitative comparison (PSNR/SSIM) of different methods on various medical datasets. The best results are highlighted in bold and the second best is underlined.

table 2 shows that our method obtains the best PSNR on all natural image datasets and competitive or best SSIM, including strong results on PolyU. As illustrated in fig. 9, it better preserves local details while suppressing residual noise and over-smoothing.

table 3 reports the best overall deblurring performance, confirming that the proposed design generalizes beyond denoising. The examples in fig. 10 further show clearer boundaries and fewer residual artifacts.

	SIDD		SIDD-Val		PolyU		MIDD		LRD		PMRID	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
NAFNet (ECCV'2022)	39.19	0.9353	39.04	0.9212	34.69	0.9215	39.17	0.8853	34.83	0.8829	38.02	0.9145
Restormer (CVPR'2022)	39.06	0.9351	38.81	0.9207	35.48	0.9217	39.08	0.8850	34.67	0.8872	37.97	0.9199
CODE (CVPR'2023)	38.67	0.9327	38.64	0.9190	28.52	0.8616	38.61	0.8807	34.51	0.8835	33.79	0.8509
CLIPDenoising (CVPR'2024)	38.03	0.9272	37.74	0.9114	20.91	0.8724	38.34	0.8767	33.77	0.8737	32.71	0.8878
LAN (CVPR'2024)	39.03	0.9351	38.84	0.9207	35.64	0.9240	39.05	0.8849	34.68	0.8864	37.96	0.9212
SplitterNet (CVPR'2024)	37.30	0.9209	37.50	0.9094	34.53	0.9255	37.84	0.8726	33.85	0.8749	33.84	0.8383
TUR (AAAI'2025)	39.04	0.9222	39.13	0.9227	35.97	0.9468	38.74	0.8808	34.48	0.8858	38.70	0.9272
CANS (CVPR'2025)	38.80	0.9333	38.85	0.9202	30.02	0.8541	38.61	0.8800	34.49	0.8841	35.33	0.8956
IDF (ICCV'2025)	36.14	0.9096	36.14	0.8969	34.64	0.9490	37.49	0.8555	32.61	0.8554	35.21	0.8973
PoolNet (TIP'2025)	39.10	0.9347	38.97	0.9217	36.23	0.9470	39.04	0.8837	34.58	**0.8908**	37.82	0.9216
RNINet (ICLR'2025)	38.03	0.9292	38.02	0.9117	31.90	0.8650	38.24	0.8780	33.94	0.8691	33.63	0.7981
CENSURE (ICLR'2026)	38.16	0.9286	38.57	0.9167	28.62	0.7956	38.67	0.8810	33.90	0.8775	35.51	0.8896
Ours	**39.53**	**0.9374**	**39.18**	**0.9231**	**36.83**	**0.9567**	**39.22**	**0.8863**	**34.91**	0.8888	**38.88**	**0.9281**

Table 2: Quantitative comparison (PSNR/SSIM) of different methods on various natural image datasets. The best results are highlighted in bold and the second best is underlined.

	GoPro		GoPro-Val	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
NAFNet (ECCV'2022)	29.94	0.8806	29.20	0.8772
Restormer (CVPR'2022)	29.17	0.8657	28.69	0.8648
CODE (CVPR'2023)	27.21	0.8132	26.87	0.8134
CLIPDenoising (CVPR'2024)	27.69	0.8271	27.35	0.8277
LAN (CVPR'2024)	27.93	0.8359	27.54	0.8349
SplitterNet (CVPR'2024)	26.78	0.8028	26.54	0.8037
TUR (AAAI'2025)	29.12	0.8682	28.50	0.8663
CANS (CVPR'2025)	27.86	0.8305	27.26	0.8265
IDF (ICCV'2025)	26.53	0.7988	26.29	0.7984
PoolNet (TIP'2025)	29.87	0.8782	**29.52**	0.8818
RNINet (ICLR'2025)	28.49	0.8467	28.15	0.8481
CENSURE (ICLR'2026)	27.36	0.8144	27.09	0.8164
Ours	**29.98**	**0.8847**	29.50	**0.8856**

Table 3: Quantitative comparison on natural image deblurring datasets.

	Tech	#Parameter/M	GFLOPs	Inference GPU Memory/MB	Inference Time/s
CENSURE (ICLR'2026)	Divergence-Free	17.01	555.27	10335.9	439
TUR (AAAI'2025)	CNN	2.35	147.05	1147.5	130
IDF (ICCV'2025)	CNN	0.05	234.856	**346.3**	**13**
PoolNet (TIP'2025)	CNN	3.82	264.51	646.5	59
NAFNet (ECCV'2022)	U-Net	17.11	127.59	275.4	6
CLIPDenoising (CVPR'2024)	U-Net	34.52	178.09	415.9	**2**
SplitterNet (CVPR'2024)	U-Net	0.73	3.19	**190.0**	13
CANS (CVPR'2025)	U-Net	4.92	77.77	364.6	13
RNINet (ICLR'2025)	U-Net	11.57	303.43	391.3	4
Restormer (CVPR'2022)	Transformer	2.34	145.69	1147.5	130
CODE (CVPR'2023)	Transformer	1.25	68.94	1648.3	68
LAN (CVPR'2024)	Transformer	2.34	145.69	1166.0	134
Ours	Transformer	3.12	174.79	**677.1**	**60**

Table 4: Efficiency comparison (PSNR/SSIM) of different methods.

table 4 demonstrates that our method provides a favorable accuracy–efficiency trade-off with 3.12M parameters and 174.79 GFLOPs. It achieves higher restoration accuracy than Restormer and LAN while requiring less inference memory and time.

4.4 Ablation Study

We conduct ablation studies on LDCT-Chest and PMRID to validate the contribution of each major component. table 5 shows that DSCA, PSF, and Calibration progressively improve the baseline. Their combination achieves the best results on both LDCT-Chest and PMRID, confirming their complementary contributions.

DSC	PSF	Calibration	GFLOPs	LDCT-Chest		PMRID	
				PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
×, MDTA	×, Concatenation	×	133.37	30.30	0.6482	38.56	0.9260
✓	×, Concatenation	×	146.51	30.34	0.6503	38.70	0.9265
✓	✓	×	145.16	30.39	0.6498	38.78	0.9270
✓	✓	✓	174.79	**30.41**	**0.6508**	**38.88**	**0.9281**

Table 5: Quantitative comparison (PSNR/SSIM) of the ablation study in primary module on LDCT-Chest and PMRID datasets. The best results are highlighted in bold and the second best is underlined.

Due to space limitations, detailed ablation studies are provided in the Appendix. These experiments evaluate the key design choices of DSCA and PSF, the individual and combined contributions of PET, HF-DRE, and MPC, and the effects of residual scales, expert configurations, and high-frequency projection. Quantitative and qualitative analyses consistently validate the effectiveness and design rationale of each component.

5 Conclusion

In this work, we proposed PET-DSCFormer, a general image restoration framework that integrates degradation-aware structural coreset attention, precision-aware skip fusion, and posterior calibration for robust restoration. Extensive experiments across medical image denoising, natural image denoising, and deblurring demonstrate that PET-DSCFormer achieves consistently competitive or superior performance over recent state-of-the-art methods. Ablation studies and visual analyses further confirm that each component contributes complementary benefits, enabling reliable degradation modeling, robust feature fusion, and effective detail recovery.

REFERENCES

Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1692–1700, 2018.

Saeed Anwar and Nick Barnes. Real image denoising with feature attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3155–3164, 2019.

Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *International conference on machine learning*, pp. 524–533. PMLR, 2019.

Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 2, pp. 60–65. Ieee, 2005.

Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12299–12310, 2021a.

Hao Chen, Chenyuan Qu, Yu Zhang, Chen Chen, and Jianbo Jiao. Multi-view self-supervised disentanglement for general image denoising. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12281–12291, 2023a.

Haoyu Chen, Jinjin Gu, Yihao Liu, Salma Abdel Magid, Chao Dong, Qiong Wang, Hanspeter Pfister, and Lei Zhu. Masked image training for generalizable deep image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1692–1703, 2023b.

Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 182–192, 2021b.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pp. 17–33. Springer, 2022.

Jun Cheng, Dong Liang, and Shan Tan. Transfer clip for generalizable image denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 25974–25984, 2024.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

Marcos V Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. In *European Conference on Computer Vision*, pp. 1–21. Springer, 2024.

Yuning Cui, Wenqi Ren, and Alois Knoll. Exploring the potential of pooling techniques for universal image restoration. *IEEE Transactions on Image Processing*, 2025.

Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16 (8):2080–2095, 2007.

Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9935–9946, 2023.

Roman Flepp, Andrey Ignatov, Radu Timofte, and Luc Van Gool. Real-world mobile image denoising dataset with efficient baselines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22368–22377, 2024a.

Roman Flepp, Andrey Ignatov, Radu Timofte, and Luc Van Gool. Real-world mobile image denoising dataset with efficient baselines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22368–22377, 2024b.

Tomer Garber and Tom Tirer. Image restoration by denoising diffusion models with iteratively preconditioned guidance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 25245–25254, 2024.

Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *European conference on computer vision*, pp. 222–241. Springer, 2024a.

Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1712–1722, 2019.

Yu Guo, Yuan Gao, Yuxu Lu, Huilin Zhu, Ryan Wen Liu, and Shengfeng He. Onerestore: A universal restoration framework for composite degradation. In *European conference on computer vision*, pp. 255–272. Springer, 2024b.

Sébastien Herbreteau and Etienne Meunier. Divergence-free neural networks with application to image denoising. In *The Fourteenth International Conference on Learning Representations*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Tao Huang, Songjiang Li, Xu Jia, Huchuan Lu, and Jianzhuang Liu. Neighbor2neighbor: Self-supervised denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14781–14790, 2021.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in neural information processing systems*, 35:23593–23606, 2022.

Changjin Kim, Tae Hyun Kim, and Sungyong Baik. Lan: Learning to adapt noise for image denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 25193–25202, 2024.

Dongjin Kim, Jaekyun Ko, Muhammad Kashif Ali, and Tae Hyun Kim. Idf: Iterative dynamic filtering networks for generalizable image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12180–12190, 2025.

Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2129–2137, 2019.

Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*, 2018.

Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17452–17462, 2022.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833–1844, 2021.

Mengye Lyu, Lifeng Mei, Shoujin Huang, Sixing Liu, Yi Li, Kexin Yang, Yilong Liu, Yu Dong, Linzheng Dong, and Ed X Wu. M4raw: A multi-contrast, multi-repetition, multi-channel mri k-space dataset for low-field mri research. *Scientific Data*, 10(1):264, 2023.

Tengyu Ma, Long Ma, Ziye Li, Yuetong Wang, Jinyuan Liu, Chengpei Xu, and Risheng Liu. Rethinking reconstruction and denoising in the dark: New perspective, general architecture and beyond. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2323–2332, 2025.

Taylor R Moen, Baiyu Chen, David R Holmes III, Xinhui Duan, Zhicong Yu, Lifeng Yu, Shuai Leng, Joel G Fletcher, and Cynthia H McCollough. Low-dose ct image and projection dataset. *Medical physics*, 48(2):902–911, 2021.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3883–3891, 2017.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in neural information processing systems*, 36:71275–71293, 2023.

Yuhui Quan, Tianxiang Zheng, Zhiyuan Ma, and Hui Ji. Zero-shot blind-spot image denoising via implicit neural sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7502–7512, 2025.

Sudarshan Rajagopalan and Vishal M Patel. Awracle: All-weather image restoration using visual in-context learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6675–6683, 2025.

Sudarshan Rajagopalan, Kartik Narayan, and Vishal M Patel. Restorevar: Visual autoregressive generation for all-in-one image restoration. *arXiv preprint arXiv:2505.18047*, 2025.

Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American journal of roentgenology*, 174(1):71–74, 2000.

Jiachen Tu, Yaokun Shi, and Fan Lam. Score-based self-supervised mri denoising. *arXiv preprint arXiv:2505.05631*, 2025.

Yanjie Tu, Qingsen Yan, Axi Niu, and Jiacong Tang. Tpgdiff: Hierarchical triple-prior guided diffusion for image restoration. *arXiv preprint arXiv:2601.20306*, 2026.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5769–5780, 2022.

Yuzhi Wang, Haibin Huang, Qin Xu, Jiaming Liu, Yiqun Liu, and Jue Wang. Practical deep raw image denoising on mobile devices. In *European Conference on Computer Vision*, pp. 1–16. Springer, 2020.

Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17683–17693, 2022.

Gang Wu, Junjun Jiang, Yijun Wang, Kui Jiang, and Xianming Liu. Debiased all-in-one image restoration with task uncertainty regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8386–8394, 2025.

Jun Xu, Hui Li, Zhetong Liang, David Zhang, and Lei Zhang. Real-world noisy image denoising: A new benchmark. *arXiv preprint arXiv:1804.02603*, 2018.

Liutao Yang, Zhongnian Li, Rongjun Ge, Junyong Zhao, Haipeng Si, and Daoqiang Zhang. Low-dose ct denoising via sinogram inner-structure transformer. *IEEE transactions on medical imaging*, 42(4):910–921, 2022.

Ziyuan Yang, Yingyu Chen, Zhiwen Wang, Hongming Shan, Yang Chen, and Yi Zhang. Patient-level anatomy meets scanning-level physics: Personalized federated low-dose ct denoising empowered by large language model. In *Proceedings of the computer vision and pattern recognition conference*, pp. 5154–5163, 2025.

Zhengwei Yin, Hongjun Wang, Guixu Lin, Weihang Ran, and Yinqiang Zheng. Random is all you need: Random noise injection on feature statistics for generalizable deep image denoising. In *The Thirteenth International Conference on Learning Representations*, 2025.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *European conference on computer vision*, pp. 492–511. Springer, 2020.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14821–14831, 2021.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5728–5739, 2022.

Feng Zhang, Bin Xu, Zhiqiang Li, Xinran Liu, Qingbo Lu, Changxin Gao, and Nong Sang. Towards general low-light raw noise synthesis and modeling. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10820–10830, 2023.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26 (7):3142–3155, 2017.

Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018.

Haiyu Zhao, Yuanbiao Gou, Boyun Li, Dezhong Peng, Jiancheng Lv, and Xi Peng. Comprehensive and delicate: An efficient transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14122–14132, 2023.

Yang Zheng, Wen Li, and Zhaoqiang Liu. Image restoration via diffusion models with dynamic resolution. *arXiv preprint arXiv:2605.14267*, 2026.

A Appendix

A.1 LLM Usage

Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript. Specifically, we used an LLM to assist in refining the language, improving readability, and ensuring clarity in various sections of the paper. The model helped with tasks such as sentence rephrasing, grammar checking, and enhancing the overall flow of the text.

It is important to note that the LLM was not involved in the ideation, research methodology, or experimental design. All research concepts, ideas, and analyses were developed and conducted by the authors. The contributions of the LLM were solely focused on improving the linguistic quality of the paper, with no involvement in the scientific content or data analysis.

The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.

A.2 Theoretical Analysis

We further provide a theoretical analysis of the degradation-aware structural coreset attention, precision-aware skip fusion, posterior residual transport, and residual calibration modules. The analysis focuses on the robustness of degradation-aware aggregation, the statistical interpretation of precision-weighted fusion, and the stability of high-frequency residual refinement.

Degradation-aware structural coreset aggregation. For the m-th structural coreset, DSCA aggregates dense query features according to both the assignment probability R_{im} and the degradation-aware confidence weight ω_i:

$$\bar{q}_m = \frac{\sum_{i=1}^{N} R_{im}\omega_i q_i}{\sum_{i=1}^{N} R_{im}\omega_i + \epsilon}. \tag{39}$$

The confidence weight is defined as

$$\omega_i = \frac{1}{d_i + \epsilon}. \tag{40}$$

Under the intended assumption that the degradation estimate is non-negative, i.e., $d_i \geq 0$, we have $\omega_i > 0$. Moreover,

$$\frac{\partial \omega_i}{\partial d_i} = -\frac{1}{(d_i + \epsilon)^2} < 0. \tag{41}$$

Therefore, the contribution of a spatial location decreases monotonically as its estimated degradation strength increases.

To interpret Eq. (39), we omit the small numerical stabilization term ϵ in the denominator and define

$$a_{im} = R_{im}\omega_i. \tag{42}$$

Consider the weighted representation objective

$$\mathcal{L}_m(z) = \sum_{i=1}^{N} a_{im}\|q_i - z\|_2^2. \tag{43}$$

Taking the derivative with respect to z gives

$$\frac{\partial \mathcal{L}_m(z)}{\partial z} = 2\sum_{i=1}^{N} a_{im}(z - q_i). \tag{44}$$

Setting Eq. (44) to zero yields

$$z^* = \frac{\sum_{i=1}^{N} a_{im} q_i}{\sum_{i=1}^{N} a_{im}} = \frac{\sum_{i=1}^{N} R_{im}\omega_i q_i}{\sum_{i=1}^{N} R_{im}\omega_i}. \tag{45}$$

Thus, apart from the numerical stabilization term, the coreset query can be interpreted as the weighted least-squares representative of the query features assigned to the m-th structural coreset. The same argument applies to the aggregated key and value features.

Suppose the query feature at position i is perturbed from q_i to $q_i + \delta_i$. The induced perturbation of the coreset representation is

$$\Delta\bar{q}_m = \frac{\sum_{i=1}^{N} a_{im}\delta_i}{\sum_{i=1}^{N} a_{im}}. \tag{46}$$

By the triangle inequality,

$$\|\Delta\bar{q}_m\|_2 \leq \frac{\sum_{i=1}^{N} a_{im}\|\delta_i\|_2}{\sum_{i=1}^{N} a_{im}}. \tag{47}$$

Hence, a heavily degraded position with a small confidence weight has a correspondingly reduced influence on the aggregated structural representation.

For dense spatial self-attention, the pairwise interaction cost is

$$\mathcal{C}_{\text{dense}} = \mathcal{O}(N^2 C). \tag{48}$$

After compressing the dense features into M structural coreset tokens, self-attention among coreset tokens requires

$$\mathcal{C}_{\text{core}} = \mathcal{O}(M^2 C). \tag{49}$$

Considering dense-to-coreset aggregation and coreset-to-dense broadcasting, the overall global interaction cost is approximately

$$\mathcal{C}_{\mathrm{DSCA}} = \mathcal{O}(NMC + M^2C). \tag{50}$$

Since $M \ll N$, this is substantially lower than dense spatial attention.

Statistical interpretation of precision-aware skip fusion. After channel alignment, let $\bar{F}_e$ and $\bar{F}_d$ denote the projected encoder and decoder features. For theoretical analysis, we model them as two noisy observations of the same latent clean feature Z:

$$\bar{F}_e = Z + \varepsilon_e, \qquad \bar{F}_d = Z + \varepsilon_d. \tag{51}$$

Assume

$$\varepsilon_e \sim \mathcal{N}(0, \sigma_e^2), \qquad \varepsilon_d \sim \mathcal{N}(0, \sigma_d^2), \tag{52}$$

and that the two errors are conditionally independent. The likelihood is then

$$p(\bar{F}_e, \bar{F}_d \mid Z) \propto \exp\left[-\frac{(\bar{F}_e - Z)^2}{2\sigma_e^2} - \frac{(\bar{F}_d - Z)^2}{2\sigma_d^2}\right]. \tag{53}$$

Define the corresponding precisions as

$$\tau_e = \frac{1}{\sigma_e^2}, \qquad \tau_d = \frac{1}{\sigma_d^2}. \tag{54}$$

Then

$$p(\bar{F}_e, \bar{F}_d \mid Z) \propto \exp\left[-\frac{1}{2}\left(\tau_e(\bar{F}_e - Z)^2 + \tau_d(\bar{F}_d - Z)^2\right)\right]. \tag{55}$$

Completing the square with respect to Z gives

$$\tau_e(\bar{F}_e - Z)^2 + \tau_d(\bar{F}_d - Z)^2 = (\tau_e + \tau_d)(Z - \mu_f)^2 + C, \tag{56}$$

where C is independent of Z and

$$\mu_f = \frac{\tau_e\bar{F}_e + \tau_d\bar{F}_d}{\tau_e + \tau_d}. \tag{57}$$

Therefore,

$$p(Z \mid \bar{F}_e, \bar{F}_d) = \mathcal{N}\left(Z; \mu_f, \frac{1}{\tau_e + \tau_d}\right). \tag{58}$$

The posterior mean is exactly the precision-weighted fusion rule. With the numerical stabilization term used in implementation,

$$F_{\mathrm{fuse}} = \frac{\tau_e\bar{F}_e + \tau_d\bar{F}_d}{\tau_e + \tau_d + \epsilon}. \tag{59}$$

Ignoring the stabilization term for the statistical derivation, the posterior variance is

$$\sigma_f^2 = \frac{1}{\tau_e + \tau_d} = \frac{\sigma_e^2\sigma_d^2}{\sigma_e^2 + \sigma_d^2}. \tag{60}$$

Since $\sigma_e^2 > 0$ and $\sigma_d^2 > 0$,

$$\sigma_f^2 \le \min\{\sigma_e^2, \sigma_d^2\}. \tag{61}$$

The fusion mean can also be written as

$$\mu_f = \lambda_e\bar{F}_e + \lambda_d\bar{F}_d, \tag{62}$$

where

$$\lambda_e = \frac{\tau_e}{\tau_e + \tau_d}, \qquad \lambda_d = \frac{\tau_d}{\tau_e + \tau_d}, \qquad \lambda_e + \lambda_d = 1. \tag{63}$$

Thus, PSF performs an adaptive convex fusion between encoder and decoder features according to their estimated reliability.

Probabilistic derivation of posterior residual transport. At a decoder scale, PET models the local residual estimate and the transported coarse-scale residual estimate as

$$r_l \sim \mathcal{N}(\mu_l, \sigma_l^2), \qquad r_p \sim \mathcal{N}(\mu_p, \sigma_p^2). \tag{64}$$

Define

$$\tau_l = \frac{1}{\sigma_l^2}, \qquad \tau_p = \frac{1}{\sigma_p^2}. \tag{65}$$

Under the product-of-Gaussian-experts formulation,

$$p(r \mid F_l, F_p) \propto \mathcal{N}(r; \mu_l, \sigma_l^2)\mathcal{N}(r; \mu_p, \sigma_p^2). \tag{66}$$

Substituting the Gaussian densities gives

$$p(r \mid F_l, F_p) \propto \exp\left[-\frac{1}{2}\left(\tau_l(r-\mu_l)^2 + \tau_p(r-\mu_p)^2\right)\right]. \tag{67}$$

Completing the square gives

$$\tau_l(r-\mu_l)^2 + \tau_p(r-\mu_p)^2 = (\tau_l+\tau_p)(r-\mu_{\text{post}})^2 + C, \tag{68}$$

where

$$\mu_{\text{post}} = \frac{\tau_l\mu_l + \tau_p\mu_p}{\tau_l+\tau_p} \tag{69}$$

and

$$\sigma^2_{\text{post}} = \frac{1}{\tau_l+\tau_p}. \tag{70}$$

With numerical stabilization, the implementation uses

$$\mu_{\text{post}} = \frac{\tau_l\mu_l + \tau_p\mu_p}{\tau_l+\tau_p+\epsilon}, \qquad \sigma^2_{\text{post}} = \frac{1}{\tau_l+\tau_p+\epsilon}. \tag{71}$$

Ignoring ϵ for the statistical interpretation,

$$\mu_{\text{post}} = \lambda_l\mu_l + \lambda_p\mu_p, \tag{72}$$

where

$$\lambda_l = \frac{\tau_l}{\tau_l+\tau_p}, \qquad \lambda_p = \frac{\tau_p}{\tau_l+\tau_p}, \qquad \lambda_l+\lambda_p = 1. \tag{73}$$

Moreover,

$$\sigma^2_{\text{post}} \le \min\{\sigma_l^2, \sigma_p^2\}. \tag{74}$$

Equivalently, the posterior mean is the minimizer of the precision-weighted quadratic objective

$$\mu_{\text{post}} = \arg\min_r \left[\tau_l(r-\mu_l)^2 + \tau_p(r-\mu_p)^2\right]. \tag{75}$$

Thus, PET adaptively balances the current-scale residual estimate and the transported coarse-scale estimate according to their predicted uncertainties.

Boundary-consistent high-frequency detail estimation. The HDRE module estimates the local mean using only valid image pixels. The mathematically complete local mean is

$$LM(y)_i = \frac{\sum_{j\in\Omega_i\cap\Omega^{\text{valid}}} y_j}{|\Omega_i\cap\Omega^{\text{valid}}|}. \tag{76}$$

For a locally constant region, let

$$y_j = c, \qquad j \in \Omega_i \cap \Omega^{\text{valid}}. \tag{77}$$

Then

$$LM(y)_i = \frac{c\left|\Omega_i \cap \Omega^{\text{valid}}\right|}{\left|\Omega_i \cap \Omega^{\text{valid}}\right|} = c. \tag{78}$$

Therefore, the local high-frequency component

$$H_y = y - LM(y) \tag{79}$$

is zero in a locally constant region, including locations near image boundaries, because padded pixels are excluded from the normalization.

HDRE predicts a spatial confidence gate

$$g_{\text{detail}} = \sigma\left(\phi_g(F_{\text{detail}})\right) \tag{80}$$

and computes

$$r_{\text{detail}} = g_{\text{detail}} \odot \phi_o(F_{\text{detail}}). \tag{81}$$

Since the sigmoid function satisfies $0 < g_{\text{detail},i} < 1$, each element obeys

$$|r_{\text{detail},i}| \le \left|\phi_o(F_{\text{detail}})_i\right|. \tag{82}$$

Thus, the gate acts as a conservative modulation of the predicted detail residual.

For the additional local refinement branch,

$$F_{\text{refine}} = F_{\text{detail}} + \alpha G_{\text{hf}} \odot \phi_{\text{refine}}(F_{\text{hf}}), \tag{83}$$

where α is initialized to zero. Hence,

$$F_{\text{refine}}\big|_{\alpha=0} = F_{\text{detail}}. \tag{84}$$

The refinement branch therefore initially behaves as an identity mapping with respect to the original detail feature.

Stability of mixture posterior calibration. Let the calibration module predict K candidate corrections $\{\Delta x_k\}_{k=1}^K$ and corresponding logits $\{a_k\}_{k=1}^K$. The normalized posterior weights are

$$\pi_k = \frac{\exp(a_k)}{\sum_{j=1}^K \exp(a_j)}. \tag{85}$$

Therefore,

$$\pi_k \ge 0, \qquad \sum_{k=1}^K \pi_k = 1. \tag{86}$$

The mixture correction is

$$\Delta x = \sum_{k=1}^K \pi_k \Delta x_k. \tag{87}$$

For any vector norm, the triangle inequality gives

$$\|\Delta x\| \le \sum_{k=1}^K \pi_k \|\Delta x_k\| \le \max_k \|\Delta x_k\|. \tag{88}$$

Thus, the softmax mixture itself does not produce a correction whose norm exceeds the largest candidate correction.

Before being added to the preliminary restoration result, the correction is residualized with respect to its local mean:

$$\Delta x_{\text{hf}} = \Delta x - LM(\Delta x). \tag{89}$$

For a spatially constant correction $\Delta x = c\mathbf{1}$,

$$LM(c\mathbf{1}) = c\mathbf{1}, \tag{90}$$

and therefore

$$\Delta x_{\mathrm{hf}} = 0. \tag{91}$$

Hence, the calibration branch suppresses constant correction components and primarily retains local structural variations.

Finally, the calibrated output is

$$\hat{x} = \hat{x}_{\mathrm{pre}} + \alpha \Delta x_{\mathrm{hf}}, \tag{92}$$

where α is initialized to zero. Thus,

$$\hat{x}\big|_{\alpha=0} = \hat{x}_{\mathrm{pre}}. \tag{93}$$

This zero-initialized residual formulation leaves the preliminary restoration result unchanged at the beginning of training and allows the calibration term to be introduced gradually during optimization.

A.3 Related Works

Deep denoising and restoration architectures. CNN-based restorers remain important because of their locality, efficiency, and stable optimization Buades et al. (2005); Dabov et al. (2007); Zhang et al. (2017; 2018); Guo et al. (2019); Anwar & Barnes (2019); Chen et al. (2021b); Zamir et al. (2020). MPRNet Zamir et al. (2021) uses a multi-stage progressive design with supervised attention, high-resolution branches, and cross-stage feature fusion to recover details, but its staged structure increases complexity. NAFNet Chen et al. (2022) shows that strong restoration can be achieved by a minimalist activation-free block, yet long-range dependency modeling and reliability-aware fusion are not explicitly addressed. PoolNet Cui et al. (2025) uses average/max pooling to extract low- and high-frequency cues and modulates features efficiently, but its frequency decomposition is hand-crafted and lacks residual uncertainty modeling. Practical denoising works further consider sensor and deployment constraints. PMRID Wang et al. (2020) improves mobile RAW denoising through sensor-specific noise estimation and a k-Sigma transform; SplitterNet Flepp et al. (2024a) targets efficient high-resolution mobile denoising on smartphone GPUs and NPUs; CANS Ma et al. (2025) jointly studies RAW reconstruction and denoising in extremely dark scenes through chromaticity-and-noise perception and dual-head interaction. These methods are effective in their target settings, but they mainly emphasize data, sensors, or efficiency rather than general reliability-aware global reasoning and posterior residual calibration Buades et al. (2005); Dabov et al. (2007); Anwar & Barnes (2019); Zhang et al. (2018); Guo et al. (2019); Anwar & Barnes (2019); Chen et al. (2021b); Zamir et al. (2020).

Transformer-based restoration. Transformers enlarge the effective receptive field for restoration Chen et al. (2021a); Liang et al. (2021); Wang et al. (2022); Tu et al. (2022). Restormer Zamir et al. (2022) replaces spatial attention with channel-wise transposed attention and uses gated depth-wise feed-forward networks, achieving strong high-resolution restoration, but still aggregates dense degraded features without explicit spatial reliability estimation. CODE Zhao et al. (2023) condenses pixels into superpixel-level representations and transfers global information back to pixels, reducing cost while preserving global dependency; however, its condensed tokens are not explicitly weighted by degradation confidence. SIST Yang et al. (2022) introduces sinogram inner-structure Transformers and structure losses for low-dose CT denoising, demonstrating the value of domain physics, but its sinogram-specific design is less general. In contrast, DSCA constructs degradation-weighted structural coresets so that global context is estimated from reliable structures with lower complexity.

Generalizable and adaptive denoising Cheng et al. (2024); Kim et al. (2024; 2025); Yin et al. (2025); Chen et al. (2023b;a). CLIPDenoising Cheng et al. (2024) shows that multi-scale features from a frozen CLIP ResNet are relatively distortion-invariant and builds an asymmetric encoder-decoder with progressive feature augmentation, but it depends on an external pretrained encoder and does not model attention efficiency or residual uncertainty. LAN Kim et al. (2024) freezes a pretrained denoiser and learns an input noise offset for adaptation, but its performance depends on input-level correction at inference. IDF Kim et al. (2025) predicts iterative pixel-wise dynamic filters from noise-invariant features, global statistics, and local correlations, achieving compact generalization but remaining primarily local. RNINet Yin et al. (2025) injects random noise into feature statistics to improve robustness under noise shifts, yet this is a training-time regularization rather than an explicit reliability model.

Self-supervised, zero-shot, and score-based denoising reduce dependence on clean targets Tu et al. (2025); Quan et al. (2025); Herbreteau & Meunier; Lehtinen et al. (2018); Krull et al. (2019); Batson & Royer (2019); Huang et al. (2021). C2S Tu et al. (2025) formulates self-supervised MRI denoising through generalized denoising score matching and detail refinement, but remains tied to MRI noise modeling. Implicit Neural Sampling Quan et al. (2025) addresses correlated real-world noise by resampling noisy pixels with implicit neural representations for blind-spot prediction, although test-time optimization can be costly. CENSURE Herbreteau & Meunier designs divergence-free networks for SURE-based self-supervised denoising under unknown or varying noise levels, but focuses on unbiased risk estimation rather than architectural modeling of global context, skip reliability, and residual uncertainty.

Foundation-model, generative, and all-in-one restoration Garber & Tirer (2024); Zheng et al. (2026); Tu et al. (2026); Rajagopalan et al. (2025); Rajagopalan & Patel (2025); Wu et al. (2025); Yang et al. (2025); Ho et al. (2020); Kawar et al. (2022); Chung et al. (2022); Fei et al. (2023); Potlapalli et al. (2023); Conde et al. (2024); Li et al. (2022); Guo et al. (2024a;b). DDPG Garber & Tirer (2024) improves diffusion-prior restoration through iteratively preconditioned guidance; SubDAPS Zheng et al. (2026) accelerates diffusion posterior sampling in dynamic-resolution subspaces; TPGDiff Tu et al. (2026) injects degradation, structural, and semantic priors into diffusion restoration; RestoreVAR Rajagopalan et al. (2025) replaces iterative latent diffusion with scale-space visual autoregressive generation. These methods exploit strong generative priors [42-45], but often require pretrained generators, iterative sampling, or autoregressive decoding, and may prioritize perceptual plausibility over deterministic fidelity. AWRaCLe Rajagopalan & Patel (2025) uses CLIP-guided visual in-context learning for all-weather restoration, TUR Wu et al. (2025) dynamically regularizes task-specific losses in all-in-one restoration, and SCAN-PhysFed Yang et al. (2025) uses physics-informed prompts, hypernetworks, and federated learning for heterogeneous low-dose CT protocols. These directions broaden restoration with context, optimization, and privacy Potlapalli et al. (2023); Conde et al. (2024); Li et al. (2022); Guo et al. (2024a;b), but they rely on external priors, training strategies, or task-specific designs. PET-DSCFormer is complemntary: it improves the internal restoration process itself by injecting reliability and uncertainty into global attention, skip fusion, and coarse-to-fine residual calibration.

Overall, existing methods advance denoising from local CNN priors to efficient Transformers, adaptive denoisers, and generative all-in-one frameworks. Yet most treat feature fusion and residual prediction as deterministic operations despite the uncertainty and spatial non-uniformity of noise. PET-DSCFormer addresses this gap by introducing reliability at three critical levels: degradation-aware structural attention, precision-aware encoder-decoder fusion, and posterior residual calibration.

A.4 DATASETS

All datasets used in this work are publicly available benchmarks collected from real-world imaging scenarios.

For medical image denoising, we employ benchmarks derived from various imaging principles. Low-dose computed tomography reduces radiation exposure during CT acquisition, but it also introduces stronger noise and streak-like artifacts due to reduced photon flux. The LDCT Moen et al. (2021) dataset contains two subsets: LDCT-Chest and LDCT-Abdomen, including 16,710 and 8,159 paired noisy-clean CT images of size 512×512, respectively. MRI noise is mainly caused by acquisition electronics, thermal fluctuations, reconstruction processes, and modality-dependent signal variations. M4Raw Lyu et al. (2023) contains three modalities, namely M4Raw-T1, M4Raw-T2, and M4Raw-FLAIR. Each modality provides 11,232 paired MRI images of size 256×256. X-ray images may suffer from quantum noise, detector noise, and acquisition-dependent intensity fluctuations, which can reduce the visibility of subtle anatomical structures. JSRT Shiraishi et al. (2000) contains 247 paired X-ray images of size 256×256.

For natural image denoising, we adopt several widely used real-world benchmarks. SIDD Abdelhamed et al. (2018) contains 68,836 paired images of size 256×256, and SIDD-Val Abdelhamed et al. (2018) is used for external validation with 1,281 paired images of size 256×256. PolyU Xu et al. (2018) contains 100 real-world noisy images of size 512×512. We further use the Hynix_SL846 subset of MIDD Flepp et al. (2024b), which includes 23,040 paired images of size 512×512. To evaluate robustness under low-light conditions, we include LRD Zhang et al. (2023)

with 63,000 paired images of size 512×512. In addition, PMRID Wang et al. (2020) is used for camera RAW image restoration, containing 1,365 paired images of size 512×512.

Beyond denoising, GoPro Nah et al. (2017) contains dynamic blur caused by camera and object motion. We use 21,030 training paired images of size 256×256 and GoPro-Val as an external validation set, which contains 11,110 paired images of size 256×256.

A.5 Implementation Details

All experiments were implemented in Python using the PyTorch deep learning framework. The training and evaluation were performed on three NVIDIA RTX A6000 GPUs, each equipped with 48 GB of VRAM. For data preprocessing, medical grayscale images were trained with one input and one output channel, while natural RGB images used three input and three output channels. For CT images, the Hounsfield unit range was clipped to ([-1000, 1000]) before normalization. During training, random horizontal and vertical flipping were applied for data augmentation. For the network architecture, PET-DSCFormer used an embedding dimension of 24 and a four-level encoder-decoder structure with block numbers ([1,1,1,2]). The numbers of attention heads at the four levels were set to ([1,2,4,8]), respectively. The feed-forward expansion factor was set to 2.66, and four refinement blocks were used at the final restoration stage. For DSCA, the number of structural coreset tokens was set to 16, the descriptor dimension was set to 32, and the assignment temperature (τ) was set to 1.0. The model adopted the WithBias LayerNorm setting, and all inputs were automatically padded to multiples of 8 during forwarding and cropped back to their original spatial size after prediction. PET-DSCFormer was trained in a supervised paired restoration setting, where each degraded image was mapped to its corresponding clean target. The model was trained with a batch size of 2 for a maximum of 20 epochs. Early stopping was enabled with a patience of 5 epochs, and the best checkpoint was selected according to the validation loss. The AdamW optimizer was used with a weight decay of 1e-4 and default momentum parameters (0.9, 0.999). The learning rate was set to 3e-4. The training objective was the pixel-wise L1 reconstruction loss.

A.6 Detailed Ablation Study

table 6 evaluates the attention design in DSCA. Replacing MDTA with coreset attention already improves performance, suggesting that compact structural aggregation is beneficial. Further introducing the predicted degradation map gives the best results. In contrast, using a random degradation map or a spatially shuffled degradation map degrades performance, especially on PMRID. This verifies that the gain does not simply come from additional parameters or random weighting, but from spatially meaningful degradation-aware attention.

Attention	Degradation	GFLOPs	LDCT-Chest		PMRID	
			PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
MDTA	×	161.65	30.34	0.6495	38.02	0.9204
Coreset	×	166.53	<u>30.40</u>	<u>0.6504</u>	<u>38.82</u>	<u>0.9274</u>
Coreset	✓, degradation map (predicted)	174.79	**30.41**	**0.6508**	**38.88**	**0.9281**
Coreset	✓, degradation map (random)	174.79	30.36	0.6474	38.37	0.9244
Coreset	✓, degradation map (spatially shuffled)	174.79	30.39	0.6485	38.38	0.9240

Table 6: Quantitative comparison (PSNR/SSIM) of the ablation study in DSCA module on LDCT-Chest and PMRID datasets. The best results are highlighted in bold and the second best is underlined.

table 7 studies the number of coreset tokens. A moderate number of coresets provides the best balance between restoration accuracy and computational cost. In particular, using 16 coresets achieves the best LDCT-Chest result, while larger values only bring marginal or dataset-specific gains with increased GFLOPs. Therefore, 16 is used as the default setting.

table 8 evaluates the descriptor dimension in DSCA. The results remain relatively stable across different dimensions, while the default dimension of 32 provides a strong overall trade-off, achieving the best LDCT-Chest PSNR with moderate computational cost. This indicates that the proposed descriptor is not overly sensitive to a single hyperparameter setting.

Coresets number	GFLOPs	LDCT-Chest		PMRID	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
4	173.60	<u>30.40</u>	<u>0.6489</u>	<u>38.95</u>	0.9289
8	173.99	30.38	0.6464	38.86	0.9298
16	174.79	**30.41**	**0.6508**	38.88	0.9281
32	176.38	<u>30.40</u>	<u>0.6489</u>	**38.99**	<u>0.9308</u>
64	179.56	30.39	0.6467	38.88	**0.9309**

Table 7: Quantitative comparison (PSNR/SSIM) of the ablation study in coresets number of DSCA module on LDCT-Chest and PMRID datasets. The best results are highlighted in bold and the second best is underlined.

Descriptor dimension	GFLOPs	LDCT-Chest		PMRID	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
8	174.12	30.37	0.6398	**38.97**	**0.9318**
16	174.34	<u>30.40</u>	**0.6534**	38.83	0.9281
32	174.79	**30.41**	<u>0.6508</u>	38.88	0.9281
64	175.68	30.39	0.6467	<u>**38.96**</u>	<u>**0.9302**</u>

Table 8: Quantitative comparison (PSNR/SSIM) of the ablation study in descriptor dimension of DSCA module on LDCT-Chest and PMRID datasets. The best results are highlighted in bold and the second best is underlined.

table 9 compares concatenation, addition, scalar gate, and the proposed PSF. PSF achieves the best PSNR and SSIM on both LDCT-Chest and PMRID. Compared with simple concatenation or addition, PSF explicitly models the reliability of encoder and decoder features, leading to more robust feature fusion.

Skip fusion	LDCT-Chest		PMRID	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Concatenation	<u>30.39</u>	0.6479	38.77	0.9276
Addition	<u>30.39</u>	<u>0.6490</u>	38.74	0.9273
Scalar gate	30.38	0.6465	<u>38.76</u>	<u>0.9278</u>
PSF	**30.41**	**0.6508**	**38.88**	**0.9281**

Table 9: Quantitative comparison (PSNR/SSIM) of the ablation study in PSF module on LDCT-Chest and PMRID datasets. The best results are highlighted in bold and the second best is underlined.

fig. 2 visualizes the branch perturbation experiment. we manually inject the same perturbation into the encoder skip feature and then feed the perturbed features into models with different fusion strategies. For each strategy, the absolute reconstruction error between the restored output and the clean target is computed and visualized. Therefore, brighter regions in the figure indicate larger restoration errors caused by the corrupted skip information. The comparison shows that concatenation produces more evident structural errors because it directly passes the corrupted encoder feature to the decoder without reliability estimation. Scalar gate alleviates part of the error, but its single shared gate can-

not explicitly distinguish the reliability of encoder and decoder branches. In contrast, PSF produces a lower and more localized error response, especially around structural regions. This indicates that PSF can suppress unreliable skip information and rely more on the cleaner branch, supporting the effectiveness of precision-aware skip fusion.

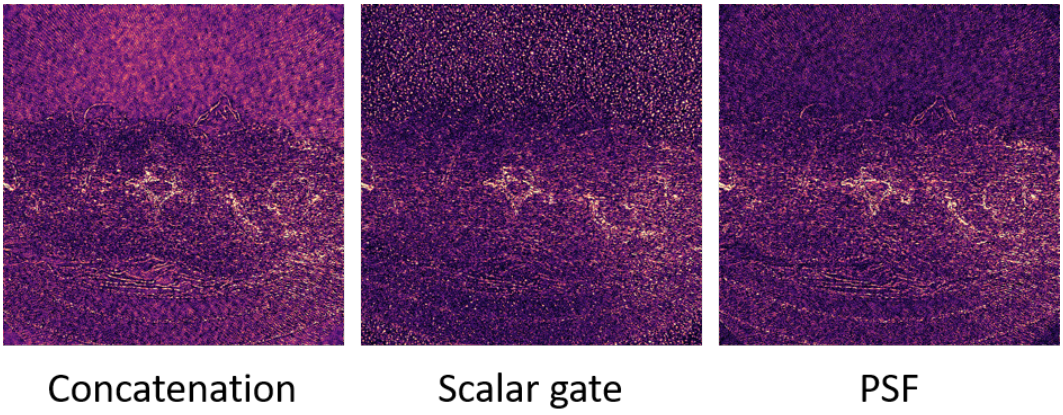

Figure 2: Reliability-aware skip fusion under branch perturbation.

table 10 evaluates the three calibration submodules. Using any single submodule improves the baseline to different degrees, while combining two submodules further improves performance. The full calibration design with PET, HF-DRE, and MPC achieves the best results on both LDCT-Chest and PMRID. This confirms that the three submodules address different aspects of restoration: multi-scale residual reliability, high-frequency detail recovery, and posterior correction.

PET	HF-DRE	MPC	LDCT-Chest		PMRID	
			PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
×	×	×	30.37	0.6433	38.78	0.9270
✓	×	×	<u>30.39</u>	0.6498	38.80	0.9272
×	✓	×	<u>30.39</u>	0.6473	38.79	0.9273
×	×	✓	<u>30.39</u>	0.6485	38.81	0.9273
✓	✓	×	<u>30.39</u>	<u>0.6506</u>	38.85	0.9275
✓	×	✓	<u>30.39</u>	0.6505	38.86	0.9278
×	✓	✓	<u>30.39</u>	<u>0.6506</u>	38.85	0.9276
✓	✓	✓	**30.41**	**0.6508**	**38.88**	**0.9281**

Table 10: Quantitative comparison (PSNR/SSIM) of the ablation study in Calibration module on LDCT-Chest and PMRID datasets. The best results are highlighted in bold and the second best is underlined.

table 11 compares different residual scales in PET. A single-scale residual prediction is less effective, especially on PMRID. Introducing posterior fusion across multiple decoder scales improves performance, and the three-scale design achieves the best results. This demonstrates that PET benefits from combining fine local residual evidence with coarse structural priors.

fig. 3 visualizes the local, prior, and posterior branches Level 1 of PET. The local estimate contains more fine-scale details but can be locally unstable, whereas the coarse prior provides more stable global structure. The posterior estimate integrates these two sources, preserving structural consistency while recovering sharper details. The restored posterior result is therefore more balanced than either local or prior estimation alone.

Residual scale	Posterior fusion	LDCT-Chest		PMRID	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Dec-1	×	<u>30.39</u>	0.6475	38.04	<u>0.9212</u>
Dec-2+Dec-1	✓	<u>30.39</u>	<u>0.6479</u>	<u>38.80</u>	0.9206
Dec-3+Dec-2+Dec-1	✓	**30.41**	**0.6508**	**38.88**	**0.9281**

Table 11: Quantitative comparison (PSNR/SSIM) of the ablation study in PET submodule of Calibration module on LDCT-Chest and PMRID datasets. The best results are highlighted in bold and the second best is underlined.

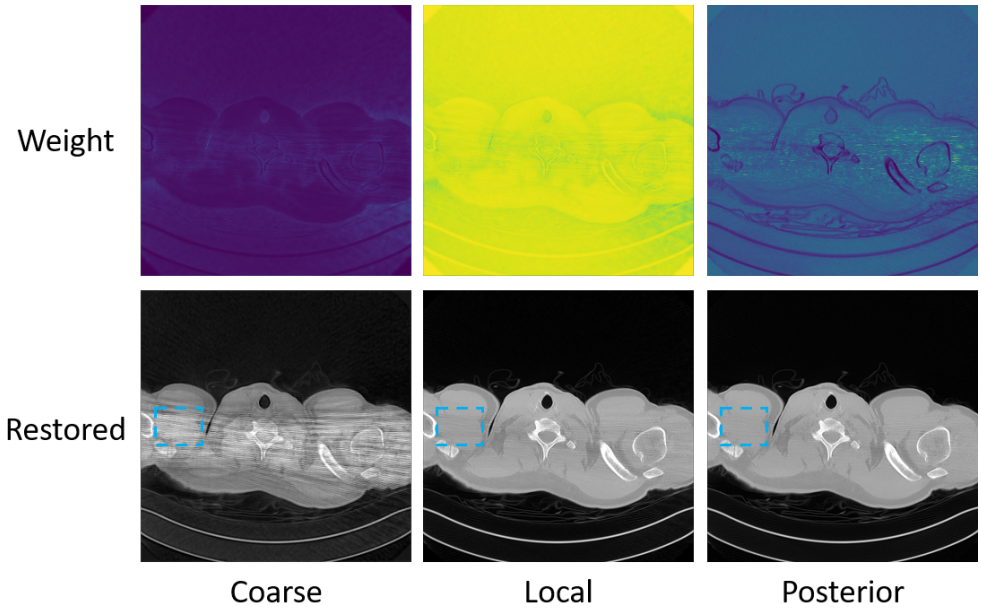

Figure 3: Posterior residual transport balances local evidence and coarse priors.

fig. 4 analyzes the HF-DRE module through gate and detail residual visualization. The detail residual is mainly activated around anatomical boundaries and high-frequency structures, rather than being uniformly distributed over the whole image. The ROI comparison shows that PET + HF-DRE restores sharper local details than PET alone, indicating that HF-DRE performs structure-aware detail compensation instead of simple global sharpening.

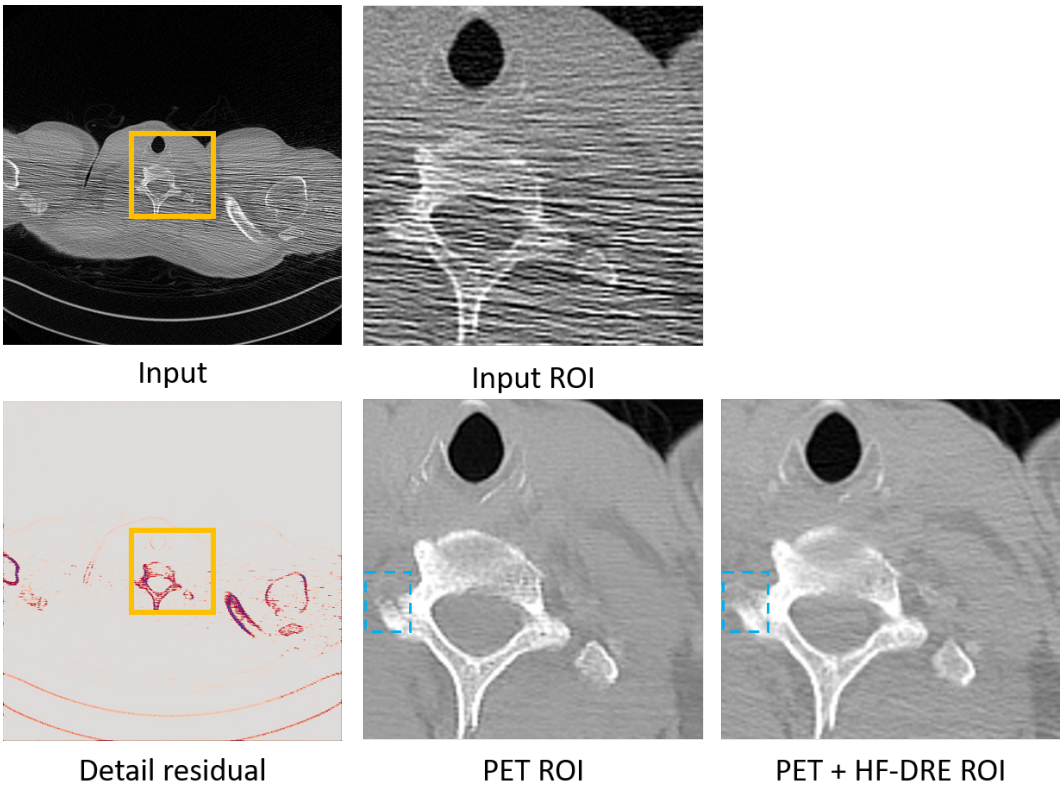

Figure 4: Structure-selective high-frequency detail injection.

fig. 5 further evaluates whether HF-DRE introduces additional noise in smooth regions. We select a smooth ROI and compare the intensity profiles of the input, PET output, PET + HF-DRE output, and target. The PET + HF-DRE curve follows the target more closely without introducing excessive oscillations, suggesting that HF-DRE enhances useful high-frequency details while avoiding obvious noise amplification in smooth areas.

table 12 studies the design choices in MPC. Using a single expert leads to clearly inferior performance, showing that a single correction pattern is insufficient. Increasing the number of experts improves flexibility, while high-frequency projection further improves the stability of correction. The three-expert setting with high-frequency projection provides a strong overall balance and is used as the default configuration.

fig. 6 visualizes the effect of high-frequency projection in MPC. Without projection, the correction may introduce low-frequency intensity shifts or over-smooth local structures. With projection, the correction is concentrated on local structural details, leading to a cleaner ROI and more stable restoration. This supports the design motivation that MPC should refine local residual errors without disturbing the low-frequency structure already recovered by previous modules.

Figure 10 presents the forced-expert counterfactual analysis. Different experts produce distinct correction patterns and are assigned different spatial weights, indicating that the experts do not collapse into identical outputs. The normal mixture combines these expert-specific corrections and provides a more balanced restoration. This demonstrates that MPC learns spatially adaptive correction modes for different local structures.

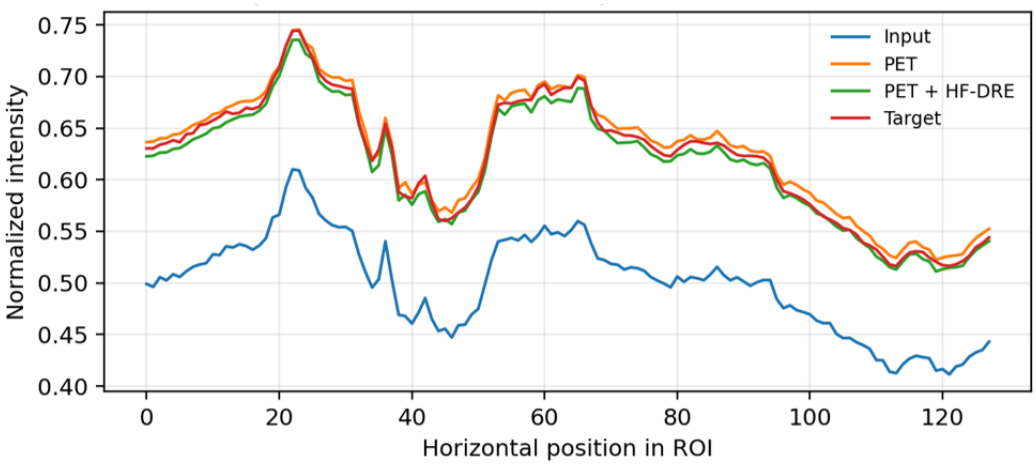

Figure 5: Smooth-region stability of HF-DRE.

High-frequency projection	nums_expert	LDCT-Chest		PMRID	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
✓	1	30.26	0.6468	38.79	0.9277
✓	2	<u>30.40</u>	0.6491	38.82	0.9281
×	3	30.39	0.6481	38.79	0.9273
✓	3	**30.41**	0.6508	<u>38.88</u>	0.9281
✓	4	30.39	<u>0.6509</u>	**39.05**	**0.9310**
✓	5	<u>30.40</u>	**0.6510**	38.75	0.9274
✓	6	<u>30.40</u>	0.6482	38.81	<u>0.9303</u>

Table 12: Quantitative comparison (PSNR/SSIM) of the ablation study in MPC submodule of Calibration module on LDCT-Chest and PMRID datasets. The best results are highlighted in bold and the second best is underlined.

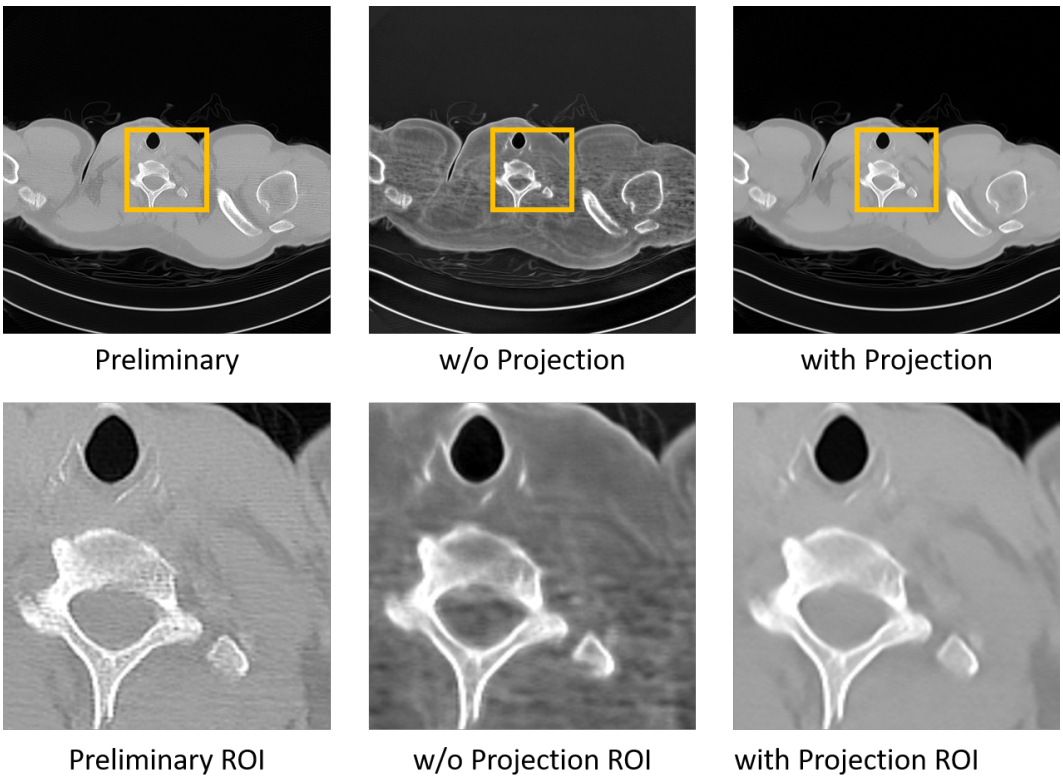

Figure 6: High-frequency projection prevents low-frequency drift.

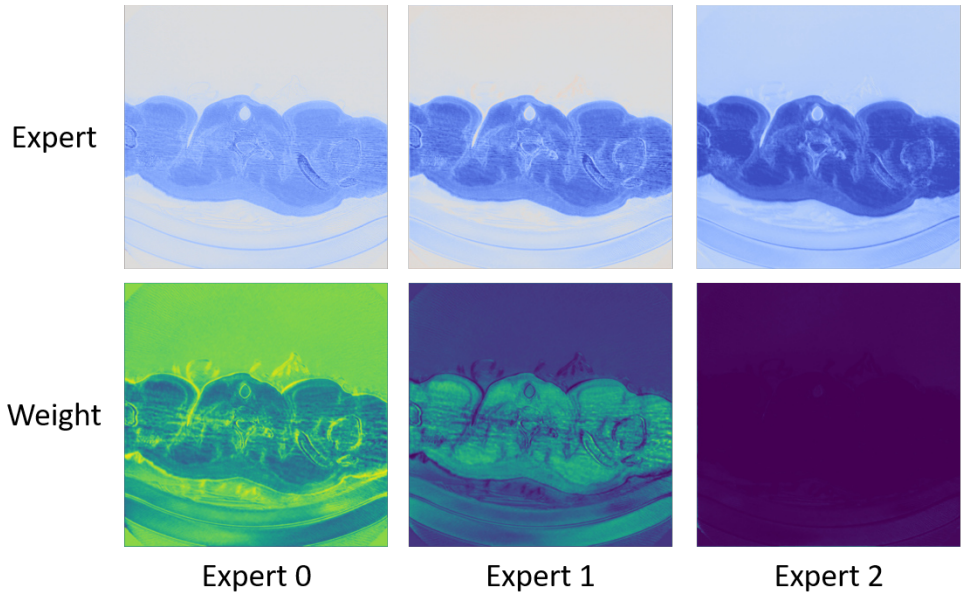

Figure 7: Expert counterfactual analysis of MPC.

A.7 VISUALIZATION

Complete qualitative results for all evaluated restoration settings are provided in the supplementary material . Specifically, fig. 8 shows the full medical image denoising comparisons across LDCT, M4Raw, and JSRT datasets. fig. 9 reports the complete natural image denoising results, and fig. 10 presents the full natural image deblurring comparisons. In these supplementary figures, challenging regions are highlighted with red and blue boxes to better compare structural fidelity and semantic consistency. The results further confirm that the proposed method consistently removes noise and artifacts while preserving fine structures across different image domains and degradation types.

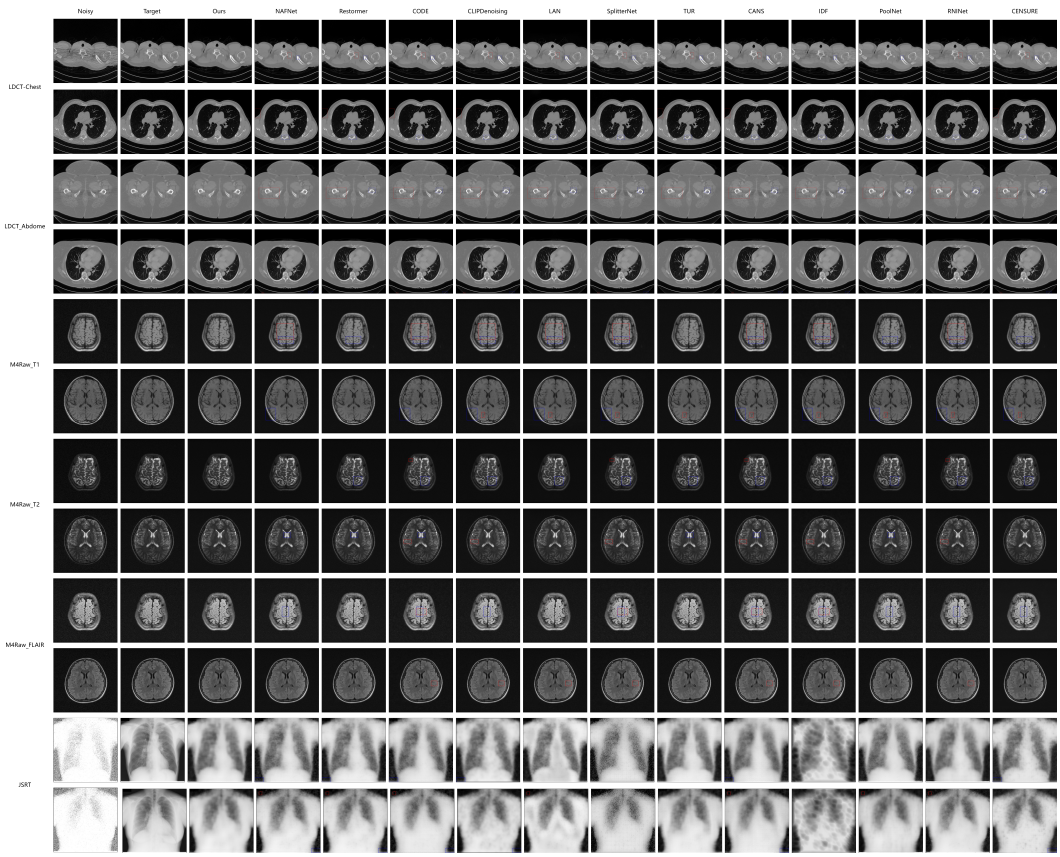

Figure 8: Complete medical image denoising results of various methods. Regions with poor structure and poor semantics are highlighted in red and blue boxes, respectively.



Figure 9: Complete natural image denoising results of various methods. Regions with poor structure and poor semantics are highlighted in red and blue boxes, respectively.

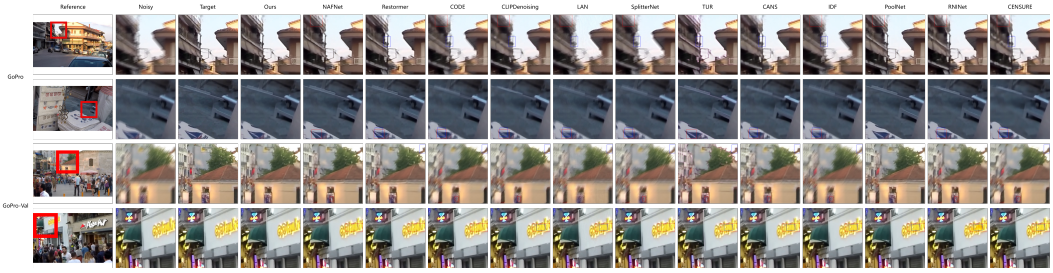

Figure 10: Complete natural image deblurring results of various methods. Regions with poor structure and poor semantics are highlighted in red and blue boxes, respectively.