CoT-SD: Chain-of-Thought Semantic Denoising

Yanlin Jiang[1,2]
SegelJiang@gmail.com
Yuchen Liu[1,2]
lyc0118@emails.bjut.edu.cn
Mingren Liu[3]
liumingren.lmr@alibaba-inc.com

[1] School of Information Science and Technology, Beijing University of Technology, China
[2] Beijing Artificial Intelligence Institute
[3] Alibaba Cloud, China

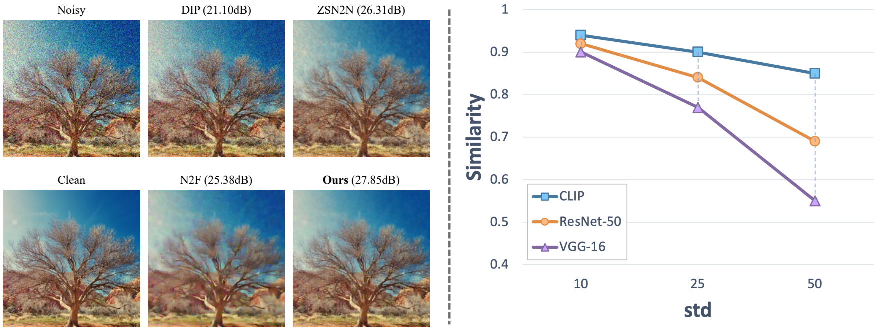

Figure 1: (a) **Visual comparison of zero-shot denoisers.** The methods compared include DIP [21], ZSN2N [17], and N2F [11]. Our CoT-SD achieves the best denoising result. (b) The cosine similarity between the features extracted from the original and noisy images by three pre-trained networks: CLIP, ResNet-50, VGG-16.

Abstract

Recent advances in zero-shot image denoising have largely been driven by self-supervised learning and generative models. However, these methods often struggle to preserve fine-grained structural details in real-world noisy images. Inspired by the human ability for slow thinking in problem-solving, which involves breaking down tasks and addressing them step by step, this paper proposes Chain-of-Thought Semantic Denoising (CoT-SD), a novel zero-shot denoising framework that leverages multi-step self-reflective reasoning and CLIP-based semantic alignment to progressively enhance image quality. Unlike conventional single-step denoising approaches, CoT-SD formulates denoising as an iterative reasoning process in which both the encoder and the decoder progressively decompose the problem, refine their understanding, and iteratively generate cleaner images. The CLIP-guided decoder ensures that denoised outputs align with high-level image semantics, enabling deeper image understanding for reflection, thereby improving the quality of the denoised images. Experimental results demonstrate that our CoT-SD outperforms other state-of-the-art (SOTA) dataset-free denoisers.

1 Introduction

Image denoising seeks to remove noise while preserving structural and textural details to enhance image quality [12, 16, 33]. Recent progress has shifted the focus from traditional statistical methods (e.g., Gaussian and Median filtering) to learning-based approaches, which are typically categorized into supervised [13, 27, 28, 29, 30], self-supervised [7, 9, 10, 18, 23, 24], and zero-shot [11, 17, 19, 20, 21] methods. These CNN-based methods model the complex distribution of noise to remove it effectively while retaining image integrity.

The implementation of learning-based denoisers resembles human fast thinking [5], which involves making intuitive decisions based on prior experience, such as directly mapping 1+1 to 2. Supervised denoisers learn from large-scale noisy-clean image pairs to construct direct mappings from noisy inputs to clean outputs. In contrast, self-supervised and zero-shot methods exploit the statistical properties of noise itself, enabling fast mapping strategies as exemplified by the Noise2Noise theory (explained in Section 2). These approaches, while efficient, often suffer from poor generalization when confronted with complex and irregular real-world noise [4, 15, 22]. Supervised methods are vulnerable to unseen noise distributions, and although self-supervised and zero-shot methods improve adaptability, they still struggle to match diverse noise characteristics accurately. To address this, we propose Chain of Thought Semantic Denoising (CoT-SD), a zero-shot method designed to flexibly handle unfamiliar and complex noise distributions by extending beyond fast-thinking paradigms.

CoT-SD is inspired by the human problem-solving mechanism rooted in slow thinking. Following the Chain of Thought paradigm [6, 14, 25], we structure the denoising process into four reasoning-driven stages: problem observation, task decomposition, step-by-step inference, and reflective refinement. First, a low-resolution smooth image is obtained through blurring and downsampling preprocessing, which serves as prior knowledge for the initial observation stage. The process of reconstructing the denoised image gradually generates the underlying structure and colors through iterative self-attention mechanisms, followed by a progressive refinement of the details. During this process, the semantic features extracted by CLIP are multimodally aligned with the image information generated at each step in a reflection space, forming memory tokens that guide the network in optimizing image reconstruction. Experimental results on several real-world image datasets show that CoT-SD outperforms other recent dataset-free methods in denoising performance.

Our main contributions are summarized as follows:

- The network adopts a Chain of Thought structure, using a multi-stage self-attention mechanism to progressively generate and refine the image, from basic structure to global details. This stepwise generation of the denoised image enhances the hierarchical quality and stability of the denoising process.

- The semantic features extracted by CLIP are aligned with the features of the generated image through a multimodal reflection space, guiding image generation and ensuring visual and semantic consistency in the reconstructed image.

- Through the multimodal alignment and stepwise generation mechanism, this method performs excellently in zero-shot denoising tasks, with the quality of the reconstructed denoised image outperforming the state-of-the-art dataset-free methods.

2 Related Work

Supervised methods Supervised denoisers achieve end-to-end mappings from noisy to clean images using deep networks trained on large-scale noisy-clean image pairs. CNN-based methods such as FFDNet [30] and DnCNN [29] extract image priors to model noise: FFDNet incorporates noise level maps and downsampling for handling multiple noise levels, while DnCNN adopts residual learning for blind Gaussian denoising. Generative models like Diff-Retinex [27] employ Transformer-based Retinex decomposition to extract illumination and reflectance, followed by a diffusion model to reconstruct natural lighting and suppress noise and color bias. Despite their strong performance in controlled and some real-world settings, these methods still face challenges such as dependency on data and noise assumptions, limited generalization, and high computational demands in complex real-world noise conditions.

Self-supervised methods Self-supervised denoisers rely solely on noisy images for training, offering a practical solution when clean data is unavailable. The seminal Noise2Noise (N2N) [10] trains on multiple independent noisy observations under the assumption of zero-mean noise, achieving performance comparable to supervised methods without clean targets. However, collecting multiple noisy instances of the same scene is often impractical. To alleviate this, Neighbor2Neighbor (Ne2Ne) [7] generates pseudo-pairs using random neighbor subsampling under the assumption of independent noisy pixels. Noise2Void (N2V) [9] introduced blind-spot networks (BSNs) that mask the central pixel of each receptive field to prevent identity mapping. Sampling Difference As Perturbation (SDAP) [18] improves BSN performance by adding perturbations through Random Sub-samples Generation. Despite these advances, self-supervised methods still face limitations due to reliance on noise assumptions and challenges in data acquisition.

Zero-shot methods Zero-shot denoisers utilize the statistical properties of noise to restore images without requiring labeled data. Among them, DIP [21] leverages the low-level statistical priors of untrained networks to separate noise from signal. ZSN2N [17] and N2F [11], both rooted in the N2N framework, adopt different strategies. ZSN2N extends the zero-shot approach by leveraging fixed filters to a noisy image, generating two corresponding downsampled versions to create input-target pairs. N2F applies a checkerboard downsampling strategy to generate four sub-images from a single noisy input, exploiting post-downsampling noise independence and structural invariance for efficient training and denoising. However, both methods are limited in handling structured or complex real-world noise due to their reliance on noise independence.

3 Method

The overall architecture of CoT-SD is illustrated in Figure2. It consists of four key stages: (1) initial observation, where the model extracts coarse structural priors from the noisy input; (2) stepwise reconstruction, where the image is gradually restored via iterative attention mechanisms; (3) semantic reflection, which refines reconstruction through multimodal alignment in a reflection space guided by CLIP features and memory tokens; and (4) result inspection, where a refinement network and loss supervision ensure consistency with the input.

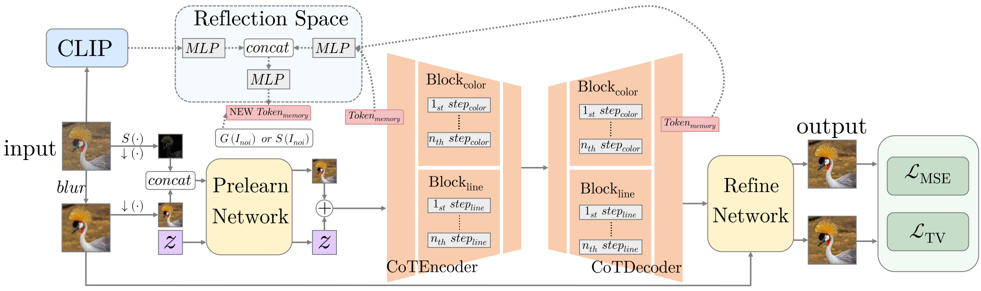

Figure 2: **The overall architecture of CoT-SD.**

3.1 Preliminary

Contrastive Language-Image Pretraining (CLIP) is a model trained through contrastive learning on a large-scale image-text pair dataset, enabling it to simultaneously understand the semantics of both images and text, and map them to a shared semantic space. The objective of CLIP training is to capture high-level semantic information of images, such as the main objects, scenes, and context within the image. Therefore, noise, which has a minimal impact on semantics, has a limited effect on the semantic features extracted by CLIP. To prove it, we used the pre-trained CLIP, VGG16, and ResNet50 to extract features from the original image and noise-corrupted images with noise levels $\sigma \in (10,\ 25,\ 50)$, respectively. The cosine similarity between the features extracted from the original and noisy images by these pre-trained networks is shown in Figure 1 (right). It is evident that the image semantic features collected by CLIP are less affected by noise. Therefore, we aim to use the CLIP semantic features to guide the network in reconstructing the clear image.

3.2 Observation Module

Similar to how humans solve problems, in the chain of thought, the network first needs to perform a "preliminary observation" of the input noisy image, thereby perceiving the basic contours and structure of the image. Therefore, at the beginning of the algorithm, we first apply two preprocessing steps to the given noisy image I_{noi}. The first preprocessing step involves applying a Gaussian filter $G(r)$ with a radius of r to the input noisy image, resulting in a smooth blurred image. Then, a downsampling operation is performed to reduce its size to one-quarter of the original image, thus obtaining a low-resolution but approximately clean image representation I_{down}:

$$\mathrm{I_{down}} = \mathrm{Downsample}(\mathrm{I_{noi}} * \mathrm{G(r)}) \tag{1}$$

The second preprocessing step applies the Sobel operator $S(\cdot)$ to extract the first-order gradient of I_{noi}, producing the edge map E. I_{down} and E represent the "clean low-resolution observation" and "local edge detail observation," respectively. After downsampling E to obtain E_{down}, they are concatenated to form the unified output o:

$$\mathrm{o} = \mathrm{Concat}(\mathrm{I_{down}}, \mathrm{E_{down}}) \tag{2}$$

Guessing latent coding z. Humans often generate hypotheses after observing a problem, in order to validate their understanding of the issue. To enable the network to preliminarily understand these observations through hypotheses, we introduce a learnable guessing latent coding z, which is continuously updated during the training process to represent the network's hypotheses and augmentations of the image content. To enable the network to preliminarily understand these observations through hypotheses, we introduce a learnable guessing latent coding z, which is continuously updated during the training process. Through this hypothesis and learning, the network's initial perception of the overall texture and lines of the image is reinforced. After passing through a "prelearn network" constructed by convolutional layers, the observations o and z are concatenated and fed into the multi-step CoTEncoder. z provides additional degrees of freedom, and after optimization, z learns how to complement the downsampled features of the image, thereby improving the model's performance.

3.3 Stepwise CoTEncoder and CoTDecoder

The denoising task involves large-scale color smoothing and fine texture restoration. Solving both simultaneously makes it difficult to balance the image's color and details. CoTEncoder and CoTDecoder decompose the denoising image reconstruction task into two sub-tasks: coarse color rendering of the image $Stage_{color}$ and fine detail filling $Stage_{line}$. For each of these stages, $Block_{color}$ and $Block_{line}$ are established, respectively. During $Stage_{color}$, only $Block_{color}$ is activated. When transitioning to $Stage_{line}$, $Block_{color}$ is frozen and $Block_{line}$ is activated. In $Block_{color}$ and $Block_{line}$, the network performs $steps_{color}$ and $steps_{line}$ iterations, respectively, to gradually complete the two stages. This division is different from traditional single-step denoising methods and allows for stepwise generation within the same training loop, where color foundation is established first, and detail lines are filled in afterward.

However, dividing the task into two stages and performing stepwise iterations presents two challenges. First, when the network transitions from $Stage_{color}$ to $Stage_{line}$, it is likely to lose its initial understanding of the image structure, leading to a disruption in the overall learning process. Second, it cannot be ensured that the network focuses on the underlying color reconstruction during $Stage_{color}$ and on the reconstruction of the line texture during $Stage_{line}$. Therefore, we introduce the 'memory Token' as a cross-stage memory bridge between the two stages and apply attention supervision.

The $Token_{memory}$ does not represent the specific pixel information of the image. $Token_{memory}$ stores the 'previous stage' understanding of the image through learnable parameters and the self-attention mechanism, such as global color estimation and local structure determination. After each step iteration, the $Token_{memory}$ will refresh with the updated image features and be aligned cross-modally with the CLIP semantic features in the 'reflection space' (Sec 3.6). For example, when the network completes the Color reconstruction at the t_{th} $step$, the $Token_{memory}$ will retain aggregated information such as the "current color mapping" and "brightness distribution," and perform information correction in the CLIP cross-modal semantic space. In the $(t+1)_{th}$ $step$, the network will concatenate the New $Token_{memory}$ with the Patch features and integrate them using new self-attention, fully utilizing the previously stored information for subsequent refinement or supplementation. The $Token_{memory}$ is passed back into the input sequence at each iteration step and is not cleared or reset during the stage transitions. Therefore, when stage transitions from $Stage_{color}$ to $Stage_{line}$, the large-scale color and luminance information accumulated in the $Token_{memory}$ is retained and can be fused with the features of $Stage_{line}$.

Attention supervision. When the network is in $Stage_{color}$, $X = X_{color} = I_{noi} * G(r)$ is sent to the reflection space, where the cross-attention mechanism aligns the NEW$Token_{memory}$ with

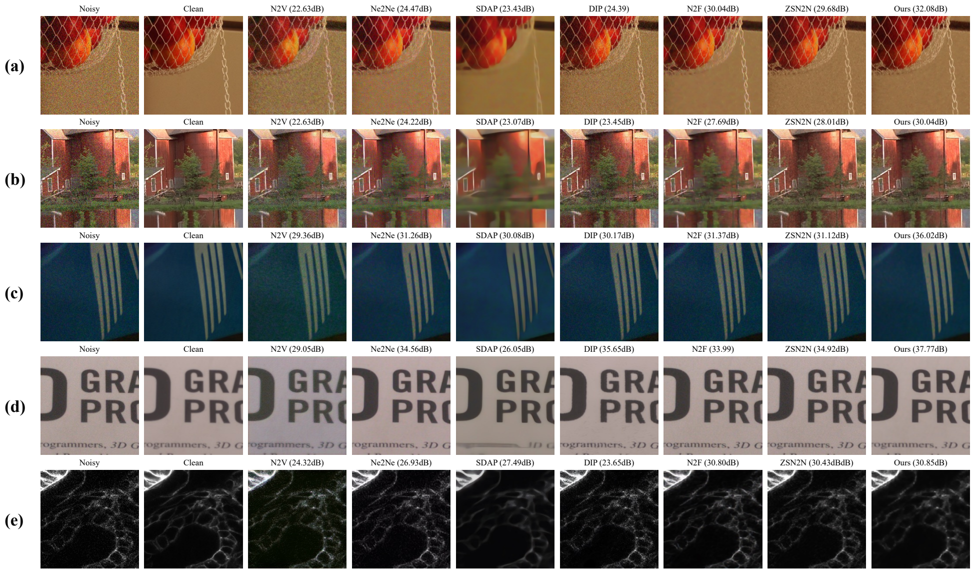

Figure 3: Visual quality comparison on **(a)** McMaster18 **(b)** Kodak24 **(c)** RENOIR, **(d)** SIDD Medium and **(e)** Fluorescence Microscopy Denoising datasets. Best viewed in color.

the low-frequency color features of the image. Similarly, when the network is in $Stage_{line}$, $X = X_{line} = S(I_{noi})$ is sent into the reflection space, where the cross-attention mechanism aligns the NEW$Token_{memory}$ with the high-frequency texture features of the image.

3.4 Reflection Space

Multi-modal alignment. Although the CLIP model can provide semantic information, this information cannot be directly used for $Token_{memory}$. This is because the image features in the network and the semantic features provided by CLIP originate from two different modalities, meaning that the network cannot directly interpret the semantic information from CLIP. To address this, we introduce several MLP models to map both the CLIP-provided semantic features and the image features from the $Token_{memory}$ into a shared dimensional "reflection space." After fusion in this space, another MLP model is used to map the fused information back into a new $Token_{memory}$.

MLPs Training. To training the alignment MLPs, we first extract the CLIP semantic features from the images in DIV2K [2] dataset, and simultaneously extract the image features. We employ loss functions that include cosine similarity loss.

3.5 Memory Token

Cross-stage memory bridge. In the network, the memory Token $Token_{memory}$ is input along with the image patch features into the Transformer Block, forming the following initial sequence:

$$\mathrm{Seq}_0 = [\mathrm{Patch}_1, \ldots, \mathrm{Patch}_\mathrm{n}, \mathrm{Token}_1, \ldots, \mathrm{Token}_\mathrm{m}] \tag{3}$$

Table 1: Quantitative comparison of CoT-SD and compared methods for synthetic noise with Gaussian noise ($\sigma \in (25, 50)$) and Poisson noise ($\lambda \in (25, 50)$). **Bold** text indicates the best performance, while underlined indicates the second best.

Datasets	Noise	Self-Supervised Methods			Zero-Shot Methods			
		N2V PSNR/SSIM	Ne2Ne PSNR/SSIM	SDAP PSNR/SSIM	DIP PSNR/SSIM	N2F PSNR/SSIM	ZSN2N PSNR/SSIM	Ours PSNR/SSIM
McMaster18	σ=25	22.62/0.486	24.82/0.554	23.59/0.653	24.40/0.547	29.03/0.799	28.62/0.770	**30.02/0.836**
	σ=50	19.44/0.328	19.63/0.337	22.20/0.545	19.35/0.325	24.35/0.646	23.74/0.588	**25.14/0.667**
	λ=25	22.13/0.490	23.80/0.582	23.42/0.658	21.61/0.564	28.55/0.835	28.08/0.800	**29.10/0.846**
	λ=50	23.87/0.576	26.31/0.682	23.69/0.683	26.67/0.719	30.39/0.875	30.16/0.854	**31.46/0.887**
Kodak24	σ=25	22.95/0.509	24.17/0.530	23.20/0.620	23.77/0.521	28.15/0.779	28.67/0.785	**29.96/0.815**
	σ=50	18.71/0.319	19.27/0.311	22.50/0.529	18.73/0.292	24.69/**0.633**	24.58/0.608	**24.95**/0.631
	λ=25	21.39/0.437	22.21/0.451	22.98/0.600	20.26/0.405	26.64/0.736	26.95/0.732	**28.81/0.802**
	λ=50	23.42/0.541	24.65/0.567	23.22/0.632	25.74/0.660	28.27/0.797	28.95/0.806	**30.23/0.848**

3.6 Inspection

The reconstructed img_{denoi} is then passed through a Refine Network $R(\cdot)$, constructed using convolutional layers, for inspection and detail enhancement. The role of the Refine Network is to provide high flexibility for the inspection task. To prevent the reconstructed image from deviating from the original, we apply the Refine Network to both img_{denoi} and a Gaussian-blurred image $I_{noi} * G(r)$, obtaining $R(img_{denoi})$ and $R(I_{noi} * G(r))$, respectively. We then calculate the MSE loss $\mathcal{L}$mse between these images and apply a Total Variation regularization loss $\mathcal{L}$TV to $R(img_{denoi})$ to prevent the Refine Network from introducing noise. The total loss function is as follows:

$$\mathcal{L}_{\text{mse}} = mse(R(img_{denoi}), R(I_{noi} * G(r))) \tag{4}$$

$$\mathcal{L}_{\text{TV}} = \sum_{i=1}^{H-1} \sum_{j=1}^{W-1} \left(|I_{i+1,j} - I_{i,j}| + |I_{i,j+1} - I_{i,j}| \right) \tag{5}$$

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{mse}} + (1 - \alpha) \mathcal{L}_{\text{TV}} \tag{6}$$

In $\mathcal{L}_{\text{TV}}$, i and j denote the pixel positions within the image, while W and H represent the width and height of the image, respectively.

4 Experiments

4.1 Implementation Details

Experimental details. We implement our method and conduct experiments on an Nvidia 4090 server. We use two metrics, peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM), to evaluate the denoising performance of the methods involved in the experiment. Higher PSNR and SSIM values indicate better denoising performance. Due to space limitations, detailed results and analysis of the ablation studies are provided in the supplementary material.
Compared Methods. We compare the denoising performance of our method with existing state-of-the-art (SOTA) self-supervised denoisers and zero-shot denoisers. The self-supervised denoisers include Noise2Void (N2V) [9], Neighbor2Neighbor (Ne2Ne) [7] and SDAP [18]. The zero-shot denoisers include Deep Image Prior (DIP) [21], Noise2Fast (N2F) [11] and Zero-Shot Noise2Noise (ZSN2N) [17].

Table 2: Comparison of self-supervised and zero-shot methods on real-world noise datasets.

Datasets\Methods	Self-Supervised Methods			Zero-Shot Methods			
	N2V PSNR/SSIM	Ne2Ne PSNR/SSIM	SDAP PSNR/SSIM	DIP PSNR/SSIM	N2F PSNR/SSIM	ZSN2N PSNR/SSIM	Ours PSNR/SSIM
RENOIR	27.61/0.617	28.68/0.611	<u>30.03</u>/**0.784**	29.43/0.644	28.73/0.608	28.66/0.601	**31.40**/<u>0.747</u>
SIDD Medium	27.60/0.639	30.17/0.676	28.60/**0.865**	<u>30.98</u>/0.709	29.95/0.658	29.92/0.645	**33.75**/<u>0.820</u>
SIDD Val	25.10/0.406	26.33/0.470	<u>28.93</u>/**0.809**	27.02/0.603	25.59/0.435	25.61/0.423	**30.49**/<u>0.715</u>
PolyU	30.18/0.891	36.22/0.920	28.89/0.913	<u>36.40</u>/**0.944**	36.19/0.916	36.08/0.908	**36.76**/<u>0.933</u>

Synthetic noise. We selected Gaussian noise with intensities of $\sigma \in (25,\ 50)$ and Poisson noise with intensities of $\lambda \in (25,\ 50)$ for the denoising test of synthetic noise. Both types of noise are generated in the intensity range of [0, 255]. The datasets used for testing include McMaster18 [8] and Kodak24[1].

Real-world sRGB noise. We selected four real-world sRGB noise datasets to test our method and the compared methods, which include RENOIR [3], SIDD Medium [1], SIDD Validation [1] and PolyU [26] Datasets. The images in the RENOIR dataset are captured under low light and high ISO settings, containing significant real-world noise. This dataset can validate the denoising performance of methods on real noisy images under low light conditions. SIDD Datasets contain images captured by five different smartphone cameras under various lighting conditions, with ten distinct scenes and diverse noise levels. PolyU is obtained by capturing static scenes with multiple commercial cameras, with noise originating from camera sensors and photon noise.

Microscope noise. We also evaluate the denoising methods on the Fluorescence Microscopy Denoising (FMD) [32] dataset to evaluate their ability to suppress fluorescence microscopy noise. Denoising experiments were conducted on the fish cell dataset within FMD.

4.2 Synthetic Experiments

Experiment implementation. In the synthetic noise experiments, we added Gaussian noise with intensities of $\sigma \in (25, 50)$ and Poisson noise with intensities of $\lambda \in (10, 25, 50)$ to the McMaster18 and Kodak24 datasets for experimentation. All images were center-cropped to 224×224. For self-supervised methods, these methods were trained on 500 noisy images from the SenseNoise-500 [31] dataset, and then experiments were conducted using the trained models. Ne2Ne, N2V, and DIP require input noise intensity information, so we provided the synthetic noise intensities (σ, λ) to these models.

Experimental results. The PSNR/SSIM of the denoised images are shown in Table 1, and the visualized results are presented in Figures 3 (a-b). The self-supervised methods show poor denoising performance for both Gaussian and Poisson noise. N2V and Ne2Ne leave significant residual noise in the denoised images, while SDAP suffer from excessive denoising, causing the images to become overly smooth and losing details. The poor performance of the self-supervised methods may be attributed to their training on the real-world noisy dataset, SenseNoise-500. As a result, their models are not well-suited to the statistical characteristics of synthetic noise, which impedes effective denoising. DIP's denoising performance is unstable due to its heavy reliance on the unstable early stopping mechanism. The visualized results show that the denoised images from DIP still contain noise. Both N2F and ZSN2N are based on the N2N theory, which aligns well with the statistical characteristics of synthetic noise. As a result, they achieve relatively good denoising performance. Our

[1]http://r0k.us/graphics/kodak/

CoT-SD demonstrates excellent performance in the experiments, successfully reconstructing denoised images. The visual results show that our denoising method removes most of the noise while preserving the structure and fine-detail of the original image.

4.3 Real-World Experiments

Experiment implementation. We conducted experiments on our CoT-SD and comparison methods using four real-world noisy datasets: RENOIR, SIDD Medium, SIDD Val, and PolyU. The scenes in these four datasets include low-light conditions, images captured by different smartphone brands, images captured by various camera brands, as well as indoor and outdoor scenarios.

Experimental results. Table 2 presents the PSNR/SSIM results for our CoT-SD and comparison methods in real-world denoising. The visualized results are presented in Figures 3 (c-d). The proposed CoT-SD demonstrates excellent PSNR performance across various denoising scenarios in all four datasets. This demonstrates the applicability of the proposed CoT-SD for denoising in various scenarios, as well as its generalizability and robustness to complex real-world noise. In contrast to synthetic noise, self-supervised methods have demonstrated promising performance in denoising real-world noise. SDAP achieved the best SSIM scores on the RENOIR, SIDD Medium, and SIDD val datasets. This is due to the images denoised by SDAP are highly smoothed, retaining only the primary structural features of the image, which results in a visually perceptible similarity between the denoised and original image structures. Since SSIM primarily evaluates structural information, SDAP achieved higher SSIM scores. For zero-shot methods, N2F and ZSN2N, which rely on noise statistical properties for denoising, exhibited poor performance when faced with complex real-world noise. In contrast, DIP, which does not rely on noise statistical properties but instead leverages the prior information from the image itself and the network's low impedance denoising, demonstrated good performance in complex real-world noise denoising tasks.

Table 3: Quantitative comparison of CoT-SD and other compared methods on the Fluorescence Microscopy Denoising dataset.

Metrics	Self-Supervised Methods			Zero-Shot Methods			
	N2V	Ne2Ne	SDAP	DIP	N2F	ZSN2N	Ours
PSNR	24.98	27.83	29.43	25.60	<u>31.74</u>	30.43	**31.89**
SSIM	0.498	0.710	0.837	0.615	<u>0.876</u>	0.829	**0.881**

4.4 Microscope Noise Experiments

Experiment implementation. We conducted experiments on our proposed CoT-SD and comparison methods on the FMD dataset. Unlike the previously mentioned real-world datasets, the FMD dataset was specifically designed to tackle the Poisson-Gaussian noise challenges inherent in real-world fluorescence microscopy images. It comprises 12,000 clear images captured with commercial confocal, two-photon, and wide-field microscopes, alongside 60,000 noisy images at varying noise levels generated through image averaging. Note that these denoising experiments were conducted on the fish cell dataset within FMD, because it provides a controlled and representative sample with well-characterized noise patterns, allowing for a fair and consistent evaluation of denoising performance.

Experimental results. The PSNR/SSIM results are shown in Table 3, and visual results are shown in Figure 3 (e). CoT-SD demonstrates superior performance on the FMD dataset, achieving a PSNR/SSIM of 31.89/0.881, significantly outperforming all compared methods. In contrast, the zero-shot methods ZSN2N and N2F attain 30.43/0.829 and 31.74/0.876, respectively. The inferior performance of N2F in real-world denoising, contrasted with its commendable results on the FMD dataset, can be attributed to the specific statistical characteristics of microscopy noise. The Poisson-Gaussian noise present in the FMD dataset aligns closely with the assumptions underlying N2F. By leveraging a masking strategy to exploit the spatial independence of noise, N2F effectively learns the mixed distribution of Poisson and Gaussian noise, thereby achieving a notable denoising performance.

5 Conclusion

In this paper, we proposed the Chain-of-Thought Semantic Denoising (CoT-SD) framework, which leverages stepwise reasoning and multimodal feature alignment for high-quality zero-shot denoising. Experimental results demonstrate that CoT-SD outperforms existing state-of-the-art methods, showing improved denoising performance on real-world noisy images. The approach effectively preserves image details and enhances generalization, making it a promising solution for challenging denoising tasks.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *CVPR*, pages 1692–1700, 2018.

[2] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPR*, July 2017.

[3] Josue Anaya and Adrian Barbu. Renoir-a dataset for real low-light noise image reduction. *arXiv preprint arXiv*, 1409:6, 2014.

[4] Jun Cheng, Tao Liu, and Shan Tan. Score priors guided deep variational inference for unsupervised real-world single image denoising. In *ICCV*, pages 12937–12948, October 2023.

[5] Wim De Neys. Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, 46:e111, 2023.

[6] Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in Neural Information Processing Systems*, 36:70757–70798, 2023.

[7] Tao Huang, Songjiang Li, Xu Jia, Huchuan Lu, and Jianzhuang Liu. Neighbor2neighbor: Self-supervised denoising from single noisy images. In *CVPR*, pages 14781–14790, 2021.

[8] Sandip M Kasar and Sachin D Ruikar. Image demosaicking by nonlocal adaptive thresholding. In *2013 International Conference on Signal Processing, Image Processing & Pattern Recognition*, pages 34–38. IEEE, 2013.

[9] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *CVPR*, pages 2129–2137, 2019.

[10] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2Noise: Learning image restoration without clean data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2965–2974. PMLR, 10–15 Jul 2018.

[11] Jason Lequyer, Reuben Philip, Amit Sharma, Wen-Hsin Hsu, and Laurence Pelletier. A fast blind zero-shot denoiser. *Nature Machine Intelligence*, 4(11):953–963, 2022.

[12] Xinyang Li, Yixin Li, Yiliang Zhou, Jiamin Wu, Zhifeng Zhao, Jiaqi Fan, Fei Deng, Zhaofa Wu, Guihua Xiao, Jing He, et al. Real-time denoising enables high-sensitivity fluorescence time-lapse imaging beyond the shot-noise limit. *Nature Biotechnology*, 41(2):282–292, 2023.

[13] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *CVPR*, pages 18278–18289, 2023.

[14] Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of thought empowers transformers to solve inherently serial problems. *arXiv preprint arXiv:2402.12875*, 1, 2024.

[15] Xin Lin, Chao Ren, Xiao Liu, Jie Huang, and Yinjie Lei. Unsupervised image denoising in real-world scenarios via self-collaboration parallel generative adversarial branches. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12642–12652, October 2023.

[16] Yilin Liu, Jiang Li, Yunkui Pang, Dong Nie, and Pew-Thian Yap. The devil is in the upsampling: Architectural decisions made simpler for denoising with deep image prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12408–12417, October 2023.

[17] Youssef Mansour and Reinhard Heckel. Zero-shot noise2noise: Efficient image denoising without any data. In *CVPR*, pages 14018–14027, 2023.

[18] Yizhong Pan, Xiao Liu, Xiangyu Liao, Yuanzhouhan Cao, and Chao Ren. Random subsamples generation for self-supervised real image denoising. In *ICCV*, pages 12150–12159, 2023.

[19] Chang Qiao, Yunmin Zeng, Quan Meng, Xingye Chen, Haoyu Chen, Tao Jiang, Rongfei Wei, Jiabao Guo, Wenfeng Fu, Huaide Lu, et al. Zero-shot learning enables instant denoising and super-resolution in optical fluorescence microscopy. *Nature Communications*, 15(1):4180, 2024.

[20] Yuhui Quan, Mingqin Chen, Tongyao Pang, and Hui Ji. Self2self with dropout: Learning self-supervised denoising from single image. In *CVPR*, pages 1887–1895, 2020.

[21] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *CVPR*, pages 9446–9454, 2018.

[22] Jiachuan Wang, Shimin Di, Lei Chen, and Charles Wang Wai Ng. Noise2info: Noisy image to information of noise for self-supervised image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16034–16043, October 2023.

[23] Zejin Wang, Jiazheng Liu, Guoqing Li, and Hua Han. Blind2unblind: Self-supervised image denoising with visible blind spots. In *CVPR*, pages 2027–2036, 2022.

[24] Zichun Wang, Ying Fu, Ji Liu, and Yulun Zhang. Lg-bpn: Local and global blind-patch network for self-supervised real-world denoising. In *CVPR*, pages 18156–18165, 2023.

[25] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[26] J Xu, H Li, Z Liang, D Zhang, and L Zhang. Real-world noisy image denoising: A new benchmark. arxiv 2018. *arXiv preprint arXiv:1804.02603*.

[27] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *ICCV*, pages 12302–12311, 2023.

[28] 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 *CVPR*, pages 5728–5739, 2022.

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

[30] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE TIP*, 27(9):4608–4622, 2018.

[31] Yi Zhang, Dasong Li, Ka Lung Law, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. Idr: Self-supervised image denoising via iterative data refinement. In *CVPR*, pages 2098–2107, 2022.

[32] Yide Zhang, Yinhao Zhu, Evan Nichols, Qingfei Wang, Siyuan Zhang, Cody Smith, and Scott Howard. A poisson-gaussian denoising dataset with real fluorescence microscopy images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11702–11710, 2019.

[33] Yunhao Zou, Chenggang Yan, and Ying Fu. Iterative denoiser and noise estimator for self-supervised image denoising. In *ICCV*, pages 13265–13274, October 2023.