CoT-SD: Chain-of-Thought Semantic Denoising

BMVC 2025

Yanlin Jiang[1,2], Yuchen Liu[1,2], Mingren Liu[3]

[1] School of Information Science and Technology, Beijing University of Technology, China;

[2] Beijing Artificial Intelligence Institute; [3] Alibaba Cloud, China

Motivation

- Existing zero-shot and self-supervised denoisers tend to over-smooth structures or mismatch complex noise statistics, leading to systematic loss of fine-grained details in real images and a persistent performance gap.
- Denoising simultaneously requires large-scale color regularization and fine texture restoration. A single-step mapping struggles to balance these competing objectives.
- Pure staging without explicit information transfer risks structural drift at stage boundaries and inconsistent attention across stages.

Contribution

- 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.

Method

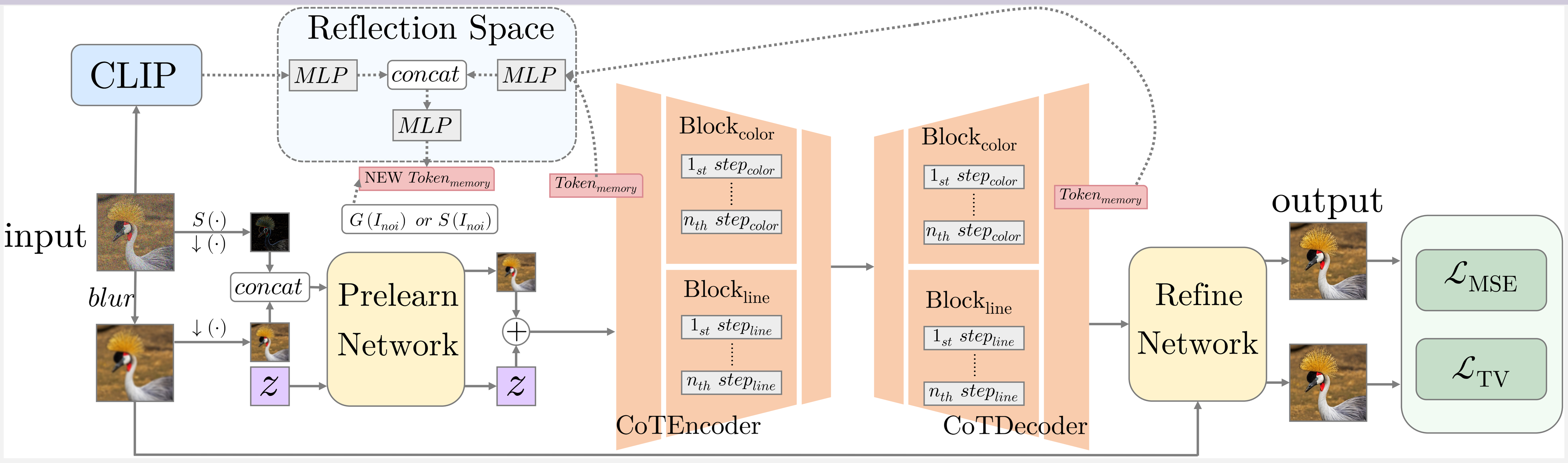

Figure 1: The overall architecture of CoT-SD.

- **Observation Prior.** Inspired by human slow-thinking, we first obtain a preliminary observation o by perceiving the image's contours and structure. We further introduce a learnable guessing latent z to express an initial hypothesis about textures and lines, and feed ($[o;z]$) into the multi-step CoTEncoder, which increases the degrees of freedom and stability for subsequent reasoning.
- **Stepwise Reconstruction.** Denoising is decomposed into $Stage_{color}$ (coarse color) and $Stage_{line}$ (fine details). Blocks are frozen/activated per stage and run multi-step iterations, progressing from foundation to details.
- **Memory Token.** $Token_{memory}$ is concatenated with patch features to carry prior-stage understanding (global color, local structures). It is refreshed via reflection-space alignment each step and persisted across stages to transfer $Stage_{color}$ priors into $Stage_{line}$; attention supervision focuses on low-frequency color vs high-frequency texture.
- **Reflection Space.** We deploy the MLPs to map and fuse network image features with CLIP semantics in a shared reflection space, then map back to update. $Token_{memory}$ Alignment is trained with DIV2K image/CLIP features using cosine-similarity loss, etc.
- **Result Inspection.** The output of the stepwise reconstruction img_{denoi} is passed through a Refine Network $R(\cdot)$ for inspection and detail enhancement, which can provide high flexibility for the inspection task.
- **Loss Function.** The total loss function is as follows.

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

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

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

Experimental Results

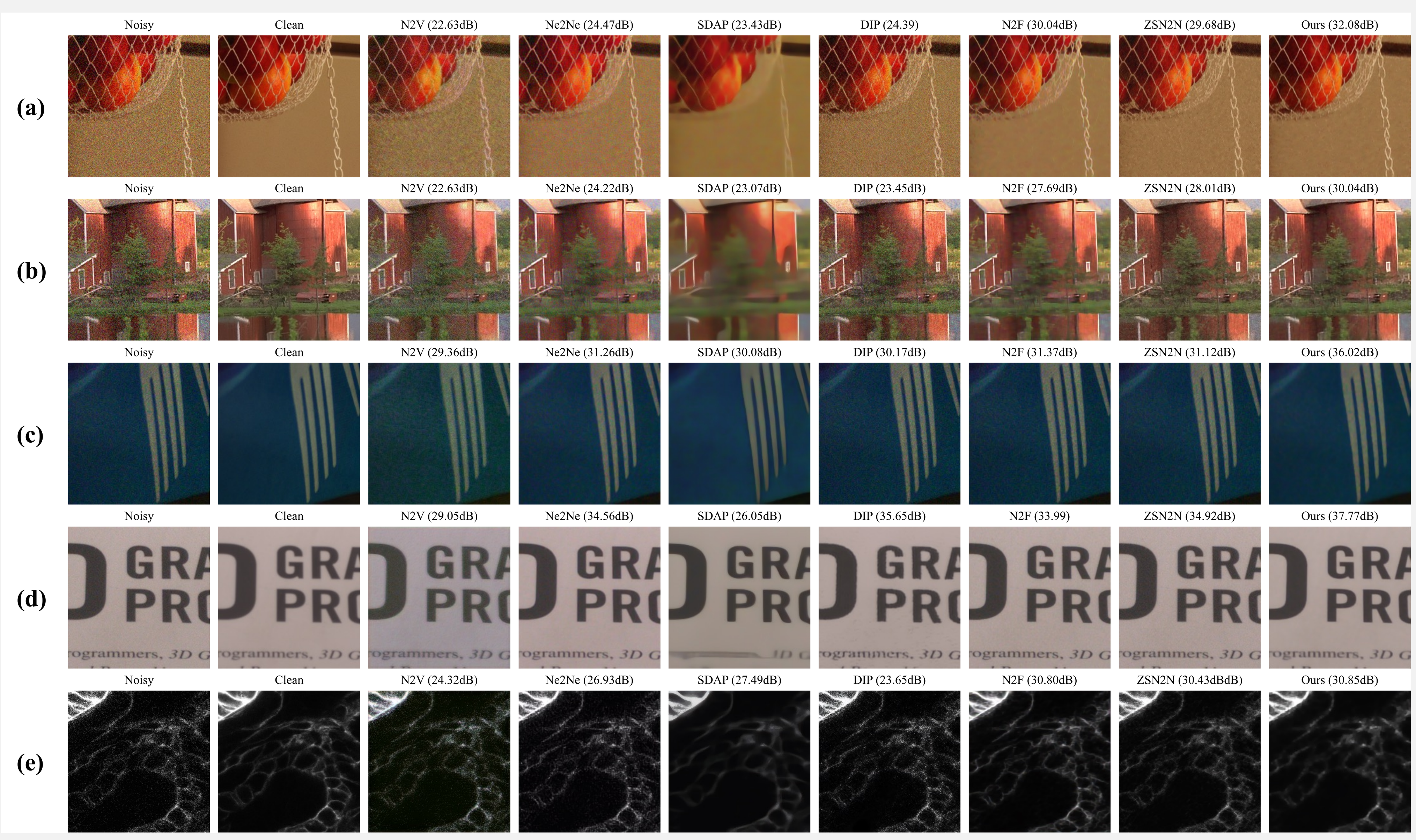

Figure 2: Visual quality comparison on (a) McMaster18, (b) Kodak24, (c) RENOIR, (d) SIDD Medium, and (e) Fluorescence Microscopy Denoising datasets.

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 <u>underlined</u> indicates the second best.

Datasets	Noise	Self-Supervised Methods N2V PSNR/SSIM	Ne2Ne PSNR/SSIM	SDAP PSNR/SSIM	Zero-Shot Methods 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	<u>29.03/0.799</u>	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	<u>24.35/0.646</u>	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	<u>28.55/0.835</u>	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	<u>30.39/0.875</u>	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	<u>28.67/0.785</u>	**29.96/0.815**
	σ=50	18.71/0.319	19.27/0.311	22.50/0.529	18.73/0.292	<u>24.69</u>/**0.633**	24.58/0.608	**24.95**/<u>0.631</u>
	λ=25	21.39/0.437	22.21/0.451	22.98/0.600	20.26/0.405	26.64/<u>0.736</u>	<u>26.95</u>/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	<u>28.95/0.806</u>	**30.23/0.848**

Table 2: Comparison of self-supervised and zero-shot methods on various datasets.

Datasets\Methods	Self-Supervised Methods N2V PSNR/SSIM	Ne2Ne PSNR/SSIM	SDAP PSNR/SSIM	Zero-Shot Methods 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>

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

Metrics	Self-Supervised Methods N2V	Ne2Ne	SDAP	Zero-Shot Methods 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**

- We compare the denoising performance of our method with existing self-supervised (N2V, Ne2Ne, SDAP) and zero-shot (DIP, N2F, ZSN2N) denoisers, reporting PSNR/SSIM with representative visualizations. Figure 2 shows the zero-shot comparison, and Table 1,2,3 demonstrate the quantitative results.
- Synthetic Experiments: On McMaster18 and Kodak24, we add Gaussian noise ($\sigma \in \{25{,}50\}$) and Poisson noise ($\lambda \in \{10{,}25{,}50\}$). Our 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.
- Real-World Experiments: We evaluate on RENOIR, SIDD Medium, SIDD Val, and PolyU, covering low-light scenes, multiple smartphone/camera brands, and both indoor/outdoor scenarios. The CoT-SD demonstrates excellent PSNR performance across various denoising scenarios in all four datasets. This demonstrates the applicability of the CoT-SD for denoising in various scenarios, as well as its generalizability and robustness to complex real-world noise.
- Microscope Noise Experiments: On the Fluorescence Microscopy Denoising (FMD) dataset, CoT-SD demonstrates superior performance on the FMD dataset, achieving a PSNR/SSIM of 31.89/0.881, significantly outperforming all compared methods.