

A Limitations

Compared to traditional diffusion-based inverse problem solvers (DIS), CoSIGN reduces number of sampling steps to 1-2 NFEs. However, the training of a ControlNet for each inverse task may limit the generalizability of the proposed method. Although we demonstrated its robustness against number of angles and noise scales in sparse-view CT reconstruction, a performance gap still exists when adapting the trained ControlNet to a different task. Future works may explore ways to utilize few-shot adaptation method for the training of ControlNet, or improve zero-shot inference ability of the proposed method.

B Implementation Details

We implement our proposed algorithm (CoSIGN) based on the consistency model (CM) codebase⁵ so that we can make use of the CM checkpoint pretrained on the LSUN bedroom dataset [52]. The UNet structure of CM contains 6 resolution levels for the input size of 256×256 . There are two residual blocks for each resolution level in both the encoder and the decoder. In the architecture of the additional encoder for guiding the CM backbone with the conditional input, we replaced each decoder layer with a zero-initialized convolution layer. We also add a zero-convolution layer before the additional encoder. We maintain the middle block in CM at the end of the additional encoder. The output of the middle block will pass through a zero-initialized convolution layer before entering the CM. We inject these conditions into CM by directly adding them with the skip connections between the encoder and the decoder. For medical images, we change the input channel of the first layer in both CM and the additional encoder into one.

In experiments on natural images, we train the additional encoder for 50k steps with a batch size of 144. In experiments on medical images, we start from training the diffusion model since no pretrained checkpoint is available. Specifically, we first train an EDM [16] model on LDCT training set [29] for 9k steps with a batch size of 144. Then we distill this diffusion model into a consistency model by training for another 12k steps. Finally, we train the additional encoder for 9k steps with the CM backbone frozen. We do not train these models for further steps since it might induce over-fitting on such a small dataset.

We adopt the forward operator of different inverse problems from DPS codebase⁶ and add hard measurement constraints like DDNM [48] into it.

For evaluation, we adopt codes from DPS [3] to calculate PSNR and SSIM whereas codes from CM [44] to calculate FID. Following [44], the intermediate noise level is determined by ternary search in multistep sampling.

⁵ https://github.com/openai/consistency_models

⁶ <https://github.com/DPS2022/diffusion-posterior-sampling>

C Implementation Details of Baselines

SwinIR

For super-resolution, we follow the default setting in 4x superresolution used in the original codebase provided by [21], and trained for 500k iterations on the LSUN-bedroom training set. For nonlinear deblurring and box-inpainting, we train swinir by mapping the degraded images to the ground truth images for 500k iterations.

DPS and MCG. For DPS and MCG, we use the original codebase provided by [3,6] and pre-trained DDPM models [13] trained on LSUN-bedroom training sets. We follow the default setting with NFE being 1000.

DDRM. For DDRM, we follow the original code provided by [17] with DDPM models trained on LSUN-bedroom training sets. We use the default parameters as displayed by [17] with NFE being 20.

CM As [44] did not report quantitative results on inverse problem solving tasks, we used the iterative inpainting and the iterative super-resolution functions in their codebase to reproduce their results. We try to keep measurement consistency and improve image quality by maximizing the number of iterations.

I²SB, CDDB The original I²SB model was trained on ImageNet. To compare it with our method, we fine-tuned it on LSUN-bedroom for 6k steps with a batch size of 256. We initialized the model for nonlinear deblur task using the checkpoint for Gaussian deblur task since no checkpoint is available on this task. Experiments on CDDB is also re-conducted on these fine-tuned models.

FBP-UNet For FBP-UNet, we use the model structure as described in [15] and then train the model with input images being FBP reconstructions and output being ground truth images of the 9000 2D CT slices from 40 patients.

D Additional Results

D.1 Distortion Metrics

In Appendix D.1, we present distortion metrics of three natural image restoration tasks. It should be noticed that unlike CT reconstruction with 23 angles, these inverse tasks are considerably aggressive. To make up for information loss while maintaining image quality, some degree of hallucination is necessary [3,35]. However, PSNR/SSIM strictly penalize hallucination as they rely on pixel-level differences.

We would like to clarify that DDB methods [4,22] outperform ours in low NFE region not because they produce higher-fidelity images, but because they "trade accuracy with quality". When working within 1-2 NFES, DDB methods generate samples closer to $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$ rather than any clear \mathbf{x}_0 from the real distribution $p(\mathbf{x})$. As shown in Fig. 3 and Fig. 4, methods with superior distortion metrics mostly generate blurry samples, indicating that they seek the mean of all possible reconstructions rather than a single clear result.

Having acknowledged this, we agree that distortion may be detrimental in certain inverse tasks, especially those with medical applications. Therefore, we provide distortion metrics of medical image reconstruction tasks in Tab. 2a, and those of natural image reconstruction tasks in Appendix D.1 for reference.

	Block Inpainting		SR×4		Nonlinear Deblur	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
SwinIR 1	20.21	0.796	27.04	0.805	23.32	0.685
DDRM 20	18.90	0.624	24.95	0.691	-	-
CM 39	18.16	0.660	24.91	0.742	-	-
DPS 1000	18.93	0.630	25.07	0.723	24.68	0.702
I ² SB 2	<u>23.21</u>	0.715	<u>27.23</u>	<u>0.816</u>	<u>28.30</u>	<u>0.843</u>
I ² SB 999	20.68	0.685	24.63	0.721	26.78	0.792
CDDB 2	23.74	0.859	27.31	0.819	28.51	0.847
CDDB 999	22.97	0.847	25.27	0.740	27.80	0.836
Ours 1	22.28	0.828	25.38	0.764	25.75	0.791
Ours 2	22.61	0.841	26.13	0.769	26.86	0.816

Table 6: Distortion metrics of solving natural image inverse problems on LSUN bedroom validation set. Baselines using around 1000 NFEs are shadowed in grey and excluded for ranking. **Bold:** best; underline: second best.

D.2 Results on Natural Image Restoration

We provide additional visual results on natural image restoration of both CoSIGN and the baselines in Fig. 8, Fig. 9 and Fig. 10. All images are randomly selected from the dataset without cherry picking. As depicted in these images, the visual quality of our results surpasses all existing methods in comparable NFE region, and is also comparable with those obtained with hundreds of NFEs.

D.3 Results on Medical Image Restoration

In Fig. 11, we provide additional visual results on medical image restoration of both CoSIGN and the baselines. The selected images encompass CT scans of abdomen, head and chest. It can be seen from the images that compared with baselines, images reconstructed with CoSIGN are both high-fidelity and noiseless.

D.4 Derivation of Hard Consistency Formula in the Linear and Noiseless Case

If the forward operator \mathbf{A} is linear, full-rank and the measurements are noiseless (i.e., $\mathbf{y} = \mathbf{A}(\mathbf{x})$), suppose we want to find the closest point to \mathbf{x}_0 that is consistent to the measurement \mathbf{y} , then we can pose the optimization problem as

$$\hat{\mathbf{x}}_0 = \arg \min_{\mathbf{z} \in \mathbb{R}^n} \{\|\mathbf{z} - \mathbf{x}_0\|_2^2\} \quad \text{s.t. } \mathbf{A}(\mathbf{z}) = \mathbf{y} \quad (14)$$

Then, the solution to this optimization problem is given by

$$\hat{\mathbf{x}}_0 = \mathbf{x}_0 - (\mathbf{A}^+ \mathbf{A} \mathbf{x}_0 - \mathbf{A}^+ \mathbf{y}), \quad (15)$$

Proof: Consider $\mathbf{t} = \mathbf{z} - \mathbf{x}_0$, then the previous optimization objective can be written to

$$\hat{\mathbf{x}}_0 = \arg \min_{\mathbf{z} \in \mathbb{R}^n} \{\|\mathbf{t}\|_2^2\} \quad \text{s.t. } \mathbf{A}(\mathbf{t}) = \mathbf{y} - \mathbf{A} \mathbf{x}_0 \quad (16)$$

Then we can decompose \mathbf{t} into a null space component and a perpendicular range space component, such that $\mathbf{t} = \mathbf{t}_{N(\mathbf{A})} + \mathbf{t}_{R(\mathbf{A}^T)}$, where $N(\mathbf{A}) \perp R(\mathbf{A}^T)$. We also have $\mathbf{A} \mathbf{t} = \mathbf{A} \mathbf{t}_{R(\mathbf{A}^T)} = \mathbf{A} \mathbf{A}^T k = \mathbf{y} - \mathbf{A} \mathbf{x}_0$, by $\mathbf{t}_{R(\mathbf{A}^T)} = \mathbf{A}^T k$. Then $k = (\mathbf{A} \mathbf{A}^T)^{-1} (\mathbf{y} - \mathbf{A} \mathbf{x}_0)$, and then $\mathbf{t}_{R(\mathbf{A}^T)} = \mathbf{A}^T (\mathbf{A} \mathbf{A}^T)^{-1} (\mathbf{y} - \mathbf{A} \mathbf{x}_0) = \mathbf{A}^\dagger (\mathbf{y} - \mathbf{A} \mathbf{x}_0)$

We also have $\|\mathbf{t}\|_2^2 = \|\mathbf{t}_{N(\mathbf{A})}\|_2^2 + \|\mathbf{t}_{R(\mathbf{A}^T)}\|_2^2$, observe when $\mathbf{t}_{N(\mathbf{A})} = 0$, $\|\mathbf{t}\|_2^2$ is minimized. Hence, $\mathbf{z} = \mathbf{x}_0 + \mathbf{A}^\dagger (\mathbf{y} - \mathbf{A} \mathbf{x}_0)$.

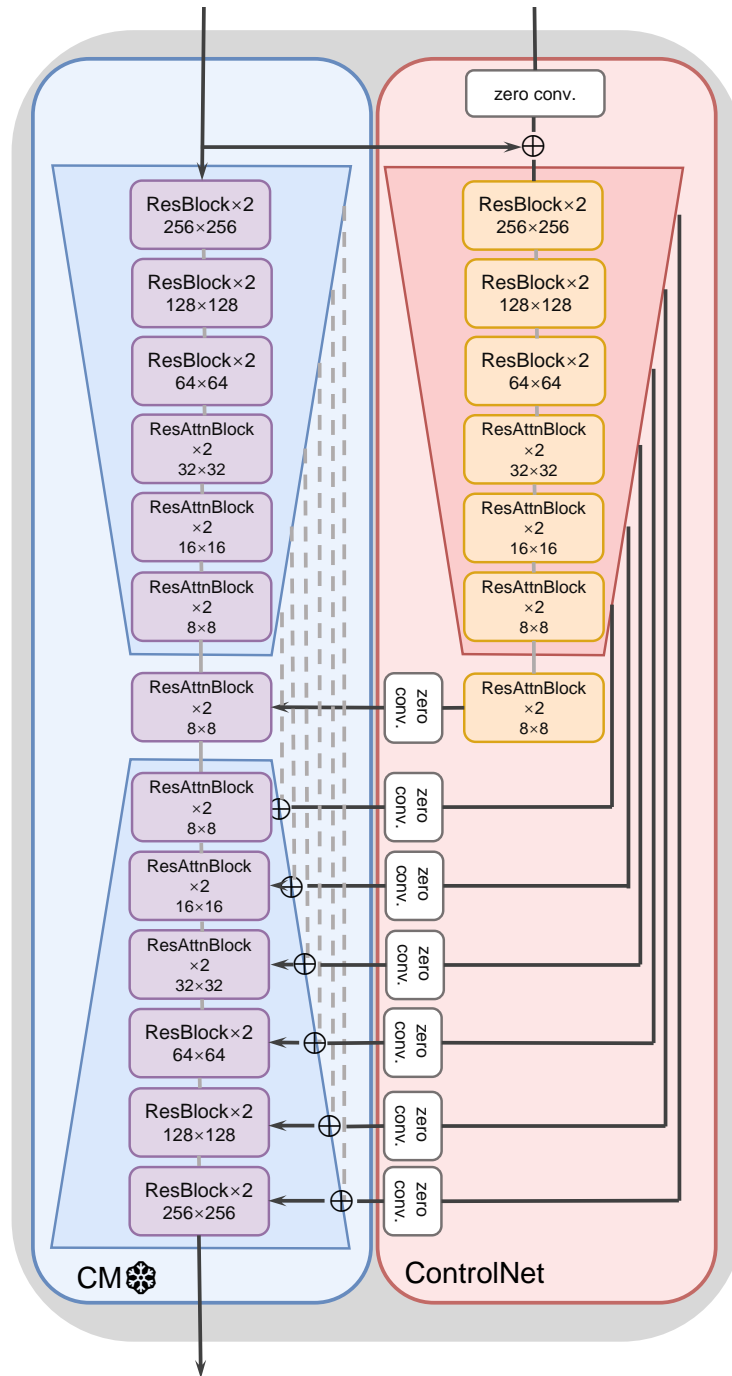


Fig. 7: Illustration of our proposed CoSIGN model structure. “ResAttnBlock×2” denotes a “ResBlock-Attention Block-ResBlock” structure.

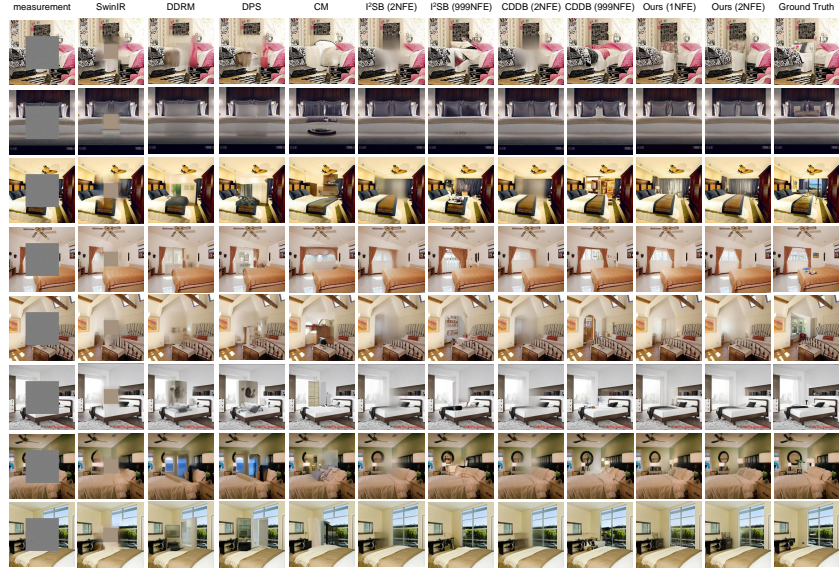


Fig. 8: Additional results on central block inpainting on LSUN bedroom validation set.

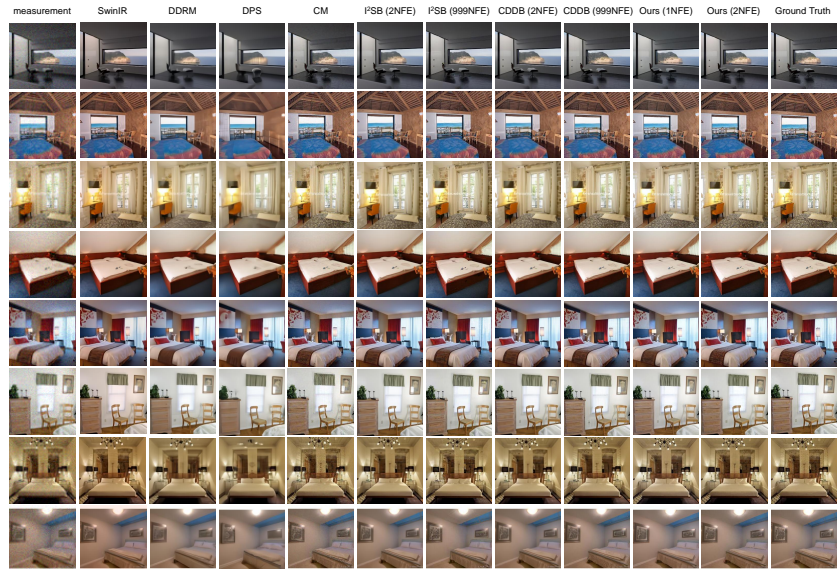


Fig. 9: Additional results on super-resolution on LSUN bedroom validation set.



Fig. 10: Additional results on nonlinear-deblur on LSUN bedroom validation set.

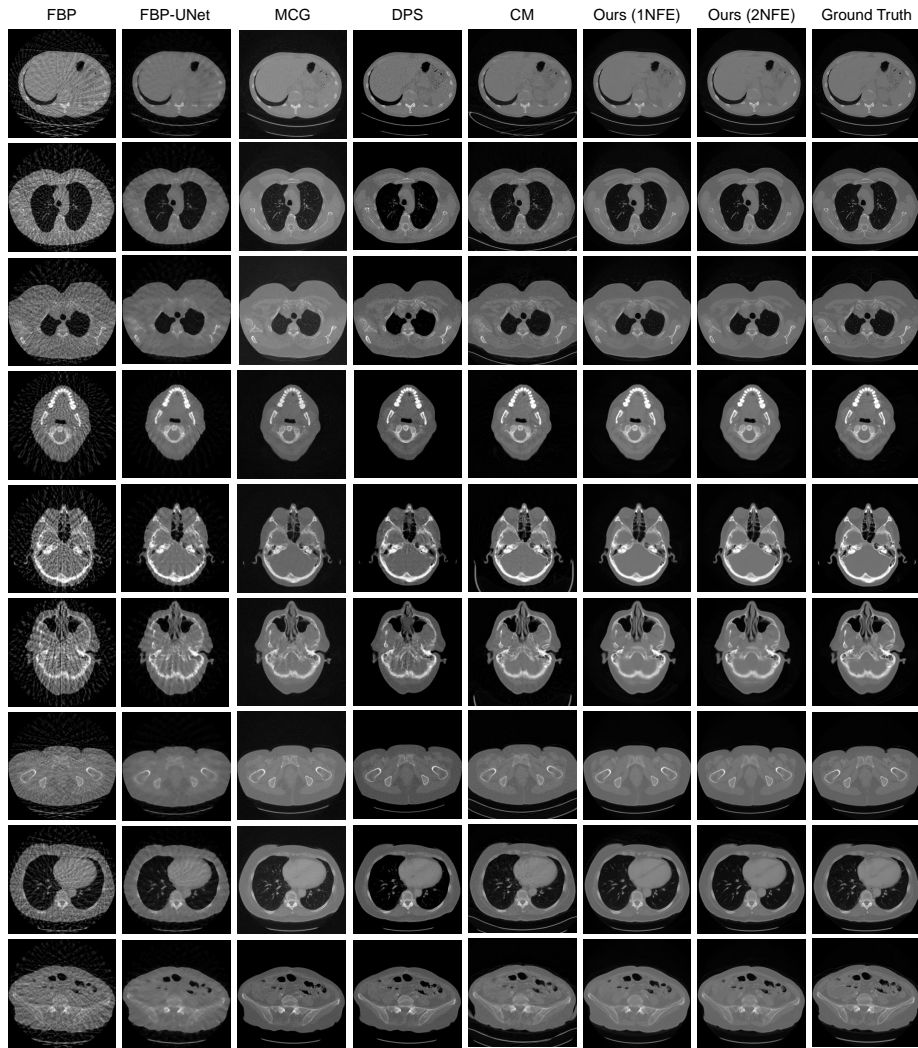


Fig. 11: Additional results on sparse-view CT reconstruction with 23 angles on LDCT validation set.