Solving Inverse Problems in Compressive Imaging with Score-Based Generative Models

Zhen Yuen Chong[1,2,†] zhenyuen@connect.hku.hk
Yaping Zhao[1,3,†,*] zhaoyp@connect.hku.hk
Zhongrui Wang[1,3] zrwang@eee.hku.hk
Edmund Y. Lam[1,3,*] elam@eee.hku.hk

[1]*The University of Hong Kong, Pokfulam, Hong Kong SAR*
[2]*University of Cambridge, St. Edmund's College, Cambridge CB3 0BN, UK*
[3]*ACCESS — AI Chip Center for Emerging Smart Systems, Hong Kong SAR*

https://github.com/zhenyuen/score_sci

***Abstract*—Snapshot Compressive Imaging (SCI) is a technique for capturing high-dimensional data through snapshot measurements using a two-dimensional (2D) detector. This approach is accomplished via coded aperture compressive temporal imaging (CACTI), which involves applying a temporally variant mask to spatially encode each sequential signal before aggregating the encoded information into a single compressed measurement. The objective of our work is to develop algorithms capable of reconstructing each video frame as a 3D data cube from its 2D measurement. To achieve this goal, we introduce multiple approaches that utilize unconditional and pre-trained 2D score models for video frame reconstruction. Our method involves modeling both the forward perturbation process of the data distribution and its reverse process as stochastic differential equations (SDEs). We also employ score-based deep learning models to estimate the scores of the data distribution across different time steps. Differing from many applications, our sampling process relies on the observed measurement, which directly corresponds to pixel values, rather than class labels. We demonstrate that employing traditional score-based generative methods with 2D score models in SCI, or integrating them into the plug-and-play framework as a deep generative prior, presents challenges. Furthermore, we propose ideas to address these limitations for future research.**

***Index Terms*—Snapshot Compressive Imaging, Computational Imaging, Generative Models**

I. Introduction

Snapshot compressive imaging (SCI) [1] is a technique that captures high-dimensional data as snapshot measurements utilizing a two-dimensional (2D) detector. This is accomplished through coded aperture compressive temporal imaging (CACTI) [2], which employs a temporal-variant mask to spatially encode each sequential signal before aggregating the encoded signal into a single compressed measurement. This research focuses on video frame reconstruction within CACTI systems using score-based generative methods. CACTI systems find applications in various fields such as medical imaging, astronomy, and surveillance, given their ability to capture high-dimensional data as snapshot measurements using a two-dimensional detector. However, reconstructing video frames from 2D measurements poses challenges due to the high-dimensional data and the compressed nature of the measurements.

Score-based generative methods have shown promise in addressing this challenge, but their application in CACTI systems is new and unexplored. Examples of traditional non-deep learning approaches include the alternating direction method of multipliers (ADMM) [3] [4] or generalized alternating projection method with a denoiser such as ADMM-TV [5] or GAP-TV [6]. Recent deep learning-based methods include the likes of deep fully-connected networks [7], GAP-net [8], Deep Tensor ADMM-net [9] and BIRNAT [10]. With the recent advancements in generative models, we seek to explore their feasibility in this context which has not been widely explored. In subsequent sections, we show that their integration into the popular plug-and-play framework as a deep generative prior is not straightforward.

We propose novel approaches involving unconditional and pre-trained 2D score models for video frame reconstruction from 2D measurements in CACTI systems. These algorithms characterize both the forward perturbation and its reverse using stochastic differential equations (SDEs), employing score-based deep learning models to estimate data distribution scores over different time intervals. Furthermore, the research highlights the significance of training score models with images closely resembling the scene compositions of the target video.

This research contributes by enhancing the efficiency and effectiveness of video frame reconstruction algorithms in CACTI systems, applicable across various fields. The limitations of traditional score-based generative methods in CACTI systems are highlighted, providing a framework for future research in this area. Overall, this study addresses the critical challenges within CACTI systems and lays the groundwork for more advanced, accurate video frame reconstruction techniques. The primary contributions include:

- Introducing algorithms that leverage unconditional and pre-trained 2D score models for video frame reconstruction in CACTI systems. These algorithms model forward and reverse perturbation processes as stochastic differential equations, employing score-based deep learning models to estimate data distribution scores across different time steps.
- Demonstrating the significance of training score models

with images that closely resemble the scene compositions of the target video, unlike general-purpose denoisers such as FFDNet [11], which do not require retraining in current CACTI algorithms.
- Identifying the constraints of traditional score-based generative methods in CACTI systems and proposing strategies to overcome them in future research. These include the necessity for appropriate training datasets, the relatively slower speed of score-based generative methods compared to end-to-end approaches, and the challenges in integrating score-based generative methods into the plug-and-play framework as a deep generative prior.

II. BACKGROUND

A. Snapshot Compressive Imaging

The mathematical model [12], [13] of Coded Aperture Compressive Temporal Imaging (CACTI) is represented by:

$$Y = \sum_{k=1}^{N_t} M_k \odot X_k + Z \tag{1}$$

such that $X \in \mathbb{R}^{N_x \times N_y \times N_t}$ is the 3D data cube, $M \in \mathbb{R}^{N_x \times N_y \times N_t}$ is the mask, $Y \in \mathbb{R}^{N_x \times N_y}$ is the 2D measurement and $Z \in \mathbb{R}^{N_x \times N_y}$ is the measurement noise. We denote the k-th video frame as X_k and its corresponding mask as M_k. The Hadamard (element-wise) product is represented by the symbol $\odot$. We define:

$$x = \begin{bmatrix} x_1^T & \dots & x_{N_t}^T \end{bmatrix}^T \tag{2}$$

$$\Phi = \begin{bmatrix} D_1 & \dots & D_N \end{bmatrix}_t \tag{3}$$

such that $x_k = vec(X_k)$ represents the vectorized form of X_k by stacking its columns, and $D_k = Diag(vec(M_k))$ is an ill-conditioned diagonal matrix with elements corresponding to the vector representation of M_k. This gives the forward expression

$$y = \Phi \mathrm{x} + z \tag{4}$$

such that $y = vec(Y)$ and $z = vec(Z)$ with a sampling rate of $1/N_t$. Note that the forward model of Compressed Sensing (CS) differs from that of SCI. CS utilizes a dense binary matrix as its sub-sampling mask, while SCI uses a sparse binary mask for modulation. Compression-based projection gradient descent is commonly used in conventional reconstruction algorithms and can be applied iteratively to approximate the solution [14]. However, the quality of the denoiser used in this approach bounds the accuracy of the solution [1].

B. Score-based Generative Models

Song et al. [15] introduced an encompassing framework for score-based generative models, including score-based maximum likelihood density (SMLD) models [16] and denoising diffusion probabilistic models (DDPM) [17]. Their method entails discretizing various stochastic differential equations (SDEs) within a unified structure. The class of SDEs examined by Song et al. is expressible using the Itô interpretation:

$$dx_t = f(t)x_t dt + g(t)d\bar{w}_t \tag{5}$$

Equation (5) defines a system containing a vector-valued function $f(t)$ (referred to as the drift coefficient), a scalar-valued function $g(t)$ (known as the diffusion coefficient), and a standard Wiener process $\bar{w}$ for $0 \leq t \leq T$. The coefficients display global Lipschitz properties in both state and time, ensuring a unique and robust solution, as elucidated in [16]. Anderson [18] presents the reverse-time SDE as follows:

$$dx_t = f'(x_t, t)dt + g(t)d\bar{w}_t \tag{6}$$

$$f'(x_t, t) = f(t)x_t - g^2(t)\nabla_x \log p_t(x_t) \tag{7}$$

Here, t flows backwards from T to 0, with dt as an infinitesimally small negative time step [15]. The noise-conditioned score network (NCSN), also known as a score model, approximates the score of the data distribution at each time step. The score at time t is $\nabla_x \log p_t(x_t)$, where $s\theta^*(x_t, t)$ represents the Noise Conditional Score Network (NCSN) [15].

The forward SDE signifies the perturbation process from the underlying data distribution p_0, introducing noise across incremental time steps until it converges into the known noisy prior distribution $p_T = \pi(x)$ [15]. Typically, this prior holds no data distribution information and is often modeled as a Gaussian distribution. On the other hand, the reverse-time SDE represents the inverse process of transforming the prior distribution back to the desired data distribution, depending on the data distribution's score. The specific PC sampling technique utilized defines the sampling process. For instance, the unconditional sampling predictor step through the Euler-Maruyama algorithm [19] is shown as:

$$x_{t_{i-1}} = x_{t_i} - f'(x_{t_i}, t) + \frac{g(t_i)}{\sqrt{N}} z_i \tag{8}$$

$$f'(x_{t_i}, t) = \frac{f(t_i)x_{t_i}}{N} - \frac{g^2(t_i)s_{\theta^*}(x_{t_i}, t_i)}{N} \tag{9}$$

The perturbation kernel, also called the transitional distribution, is modeled as a conditional Gaussian distribution: $p_{0,t}(x_t|x_0) = \mathcal{N}(x_t|\alpha(t)x_0, \beta^2(t)I)$. Here, $\alpha(t), \beta(t) : [0, 1] \to \mathbb{R}$ are tractable and derivable from $f(t)$ and $g(t)$ for the considered classes of SDEs. This approach was introduced by Särkkä and Solin (2019) [20].

Particularly, the Variance Exploding Stochastic Differential Equation (VESDE) discretization, also known as Score-based Maximum Likelihood Density (SMLD), demonstrated superior performance compared to other methods like Denoising Diffusion Probabilistic Models (DDPM) [21]. Therefore, SMLD is adopted in our experiments. Within SMLD, the transitional distribution originates from the Markov Chain $x_i i = 1^N$, which outlines the forward perturbation process for N noise scales.

$$x_i = x_{i-1} + z_{i-1}\sqrt{\sigma_i^2 - \sigma_{i-1}^2}. \tag{10}$$

The noise scales $\sigma_i i = 1^N$ in SMLD follow a positive geometric sequence, with $zi - 1 \sim \mathcal{N}(0, I)$, where $i = 1, \dots, N$. The

choice of σ_i critically influences method success, balancing the data impact minimization and the challenges posed by score matching.

Specifically, $\sigma_1 = \sigma_{\min}$ is set to a small value to minimize its impact on the data distribution, while $\sigma_N = \sigma_{\max}$ is chosen significantly large to address challenges such as low-density data regions and low-dimensional data distributions in high-dimensional ambient spaces [16]. Hence, the stochastic process formulation of SMLD reads:

$$dx = \sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^t \sqrt{2\ln\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)} d\bar{w}_t \tag{11}$$

where (11) corresponds to the VESDE. The noise scale is denoted as $\sigma(t) = \sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^t$, and the transitional PDF is presented as $p_{0,t}(x(t)|x(0)) = \mathcal{N}(x(t); x(0), [\sigma^2(t) - \sigma^2(0)]I)$.

C. NCSNs and PC Sampling

Song et al. introduced the NCSN++ model, capable of generating high-fidelity face samples through training on the FFHQ dataset [15]. The NCSN++ architecture, adapted from a 4-cascaded RefineNet architecture designed for image segmentation [22], undergoes modifications to suit the current task. The score model is trained on a dataset denoted by $x^{(1)}, \ldots, x^{(n)} \sim p_0(x)$, utilizing denoising score matching and a loss function given by:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_t \mathbb{E} x_t^{(i)}[h(x_t^{(i)}, t)] \tag{12}$$

$$h(x_t^{(i)}, t) = \|s_\theta(x_t^{(i)}, t) - \nabla_{x_t^{(i)}} p_{0t}(x_t^{(i)}|x_0^{(i)})\|_2^2 \tag{13}$$

Samples of $x_t^{(i)}$ can be obtained by simulating the reverse-time SDE [15]. The prior distribution is defined as $\pi(x) = \mathcal{N}(0, \sigma_{\max} I)$. The original paper uses 2000 noise scales by default, with a batch size of 64. The sampling process employs Predictor-Corrector (PC) methods [23], where score-based Markov Chain Monte-Carlo (MCMC) methods refine the initial approximation from a numerical SDE solver, like the Euler-Maruyama method in (9). This maintains consistency between the predicted sample's marginal distribution and the posterior p_t [15].

III. Methods

A. Conditional Reverse-Time SDE

Here, we propose the use of pre-existing sampling methods conditioned on the observed measurement [15] for reconstructing video frames.

$$dx_t = f'(x, y, t)dt + g(t)d\bar{w}_t \tag{14}$$

$$f'(x, y, t) = f(t)x_t - g^2(t)\nabla_x \log p_t(x_t|y) \tag{15}$$

To reconstruct video frames using sampling methods conditioned on the observed measurement, it is necessary to train score models to approximate $\nabla_{x_t} \log p_t(x_t|y)$ using paired data, denoted as (x_t, y). However, this approach has certain disadvantages, including the computational expense and limitations associated with supervised learning techniques [15]. Therefore, an alternative approach is to apply Bayes' Theorem to obtain the following expression:

$$\log p_t(x_t|y) = \log\left(\frac{p(y|x_t)p(x_t)}{p(y)}\right) \tag{16}$$

This allows for the estimation of $\nabla_{x_t} \log p(x_t)$ without the need for conditioning on any input. It is important to note that the variable y typically represents class labels in most literature on conditional sampling and is stochastic in nature. If $\nabla_{x_t} \log p(y|x_t)$ is intractable, it can be approximated with another score model. In SCI however, the observed measurement given by $y = \Phi x$ represents a linear constraint, and $p(y|x_t) = \delta(y - \Phi x_t)$ corresponds to the unit impulse (Dirac Delta) function, in which its score is undefined. To address this, we introduce a tractable stochastic process, denoted as $\{y_t\}_{t=0}^1$, where $y_t = \Phi x_t$. Considering the unconditional stochastic process $\{x_t\}_{t=0}^1$, we can write:

$$x_t = a(t)x_0 + \beta(t)z, \quad z \sim \mathcal{N}(0, I) \tag{17}$$

The stochastic process y_t can be expressed as

$$y_t = \alpha(t)y_0 + \beta(t)\Phi z \tag{18}$$

where y_t is Gaussian distributed. The conditional reverse-time SDE can then be rewritten as:

$$dx_t = f'(x_t, y_t, t)dt + g(t)d\bar{w}_t \tag{19}$$

$$f'(x_t, y_t, t) = f(t)x_t - g^2(t)\nabla_x \log p_t(x_t|y_t) \tag{20}$$

$$\nabla_{x_t} \log p_t(x_t|y_t) = \mathcal{L}(y_t, x_t) + \nabla_{x_t} \log p_t(x_t) \tag{21}$$

$$\mathcal{L}(y_t, x_t) = \nabla_{x_t} \log p_t(y_t|x_t) \tag{22}$$

$$= -\Phi^T \Sigma^{-1}(y_t - \Phi x_t) \tag{23}$$

such that $\Sigma = (\beta(t)\Phi)(\beta(t)\Phi)^T$.

B. Proximal Optimization

In the preceding section, we recognized the necessity of introducing additional constraints to our sampling approach, contingent on the observed measurement y. In response, we reverted to the unconditional sampling process outlined in (7). However, to ensure the consistency of sampled video frames with y, we inserted an intermediate proximal optimization step within the iterative procedure. This introduced an intermediate parameter $x'_{t_{i-1}}$, defined as follows:

$$x'_{t_{i-1}} = \underset{z}{\operatorname{argmin}}\{k_1(z, x_{t_i}) + k_2(z, u_{t_i})\} \tag{24}$$

$$k_1(z, x_{t_i}) = (1 - \lambda)\|z - x_{t_i}\|_2^2 \tag{25}$$

$$k_2(z, u_{t_i}) = \min_{y_{t_i} = \Phi u_{t_i}} \lambda\|z - u_{t_i}\|_2^2 \tag{26}$$

Here, y_{t_i} remains the stochastic variable defined in (18), with λ acting as a balancing hyper-parameter. For λ, setting it to 0 renders an unconditional sampling process, while setting it to 1 ensures full alignment with the measurement at the

expense of sample quality. Under the assumption of Φ being full rank, a feasible solution might exist for certain λ, as expressed below:

$$x'_{t_{i-1}} = z^* = B^{-1}((1-\lambda)x'_{t_i} - \Phi(\Phi\Phi^T)^{-1}y_{t_i} \quad (27)$$

$$B = (1-\lambda)I - \lambda\Phi^T(\Phi\Phi^T)^{-1}\Phi \quad (28)$$

Alternatively, we can modify (25) to:

$$x'_{t_{i-1}} = \underset{z}{\operatorname{argmin}}\{k_1(z, x_{t_i}) + k_3(z, x_{t_i}, u_{t_i})\} \quad (29)$$

$$k_3(z, x_{t_i}, u_{t_i}) = \lambda\|\tilde{z}\|_2^2 \quad (30)$$

$$\tilde{z} = z - \arg\min_{u_{t_i}} \|x_{t_i} - u_{t_i}\|_2^2 \quad (31)$$

which gives a tractable expression for $x'_{t_{i-1}}$

$$x'_{t_{i-1}} = x_{t_i} + \lambda\Phi^T(\Phi\Phi^T)^{-1}(y_{t_i} - \Phi x_{t_i}) \quad (32)$$

he rationale behind (30) can be explained as follows: the second term signifies the Euclidean projection of x_{t_i} onto a manifold defined by the set $u_{t_i} \in \mathbb{R}^n \mid y_{t_i} = \Phi u_{t_i}$. This term is a typical inclusion in Plug-and-Play (PnP) algorithms. Through embracing the PnP framework, we can deduce the ensuing update rule for our sampling process.

$$x'_{t_i} = x_{t_i} + \lambda\Phi^T(\Phi\Phi^T)^{-1}(y_{t_i} - \Phi x_{t_i}) \quad (33)$$

$$x_{t_{i-1}} = q(x'_{t_i}, z_i, s_{\theta^*}(x'_{t_i}, t_i)) \quad (34)$$

Here, the unconditional sampling step, denoted as $q(x'_{t_i}, z_i, s_{\theta^*}(x'_{t_i}, t_i))$ and defined in (9), is employed to update the intermediate parameter x'_{t_i}. This mirrors the denoising step in the PnP framework, where x_t converges towards the desired signal domain. In our methodology, however, this step corresponds to a Langevin sampling process involving the introduction of noise instead of its elimination. We observed that introducing noise without suitable noise scales led to suboptimal outcomes. Attempts to enhance accuracy by omitting noise addition resulted in worse peak signal-to-noise ratio (PSNR) values, as x_t struggled to escape local minima.

We carried out experiments using both $y_t \sim p(y_t|y)$ and y directly in our approach. When the latter was employed, the corresponding modification to the update rule in (30) took the form:

$$x'_{t_{i-1}} = x_{t_{i-1}} + \lambda\Phi^T(\Phi\Phi^T)^{-1}(y - \Phi x_{t_{i-1}}) \quad (35)$$

C. Projecting Expectations

Equation (11) defines the VESDE process used in our approach. By considering the forward-time SDE, we can express the mean of x_t as follows:

$$\bar{x}_t = \mathbb{E}[x_t] = \int_0^t \sigma_{\min}(\frac{\sigma_{\max}}{\sigma_{\min}})^s \sqrt{2\ln\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)} d\bar{w}_s + x_0 \quad (36)$$

$$= \int_0^t \sigma_{\min}(\frac{\sigma_{\max}}{\sigma_{\min}})^s \sqrt{2\ln\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)} ds + x_0 \quad (37)$$

$$\bar{x}_t = 2\sigma_{\min}\frac{(\sigma_{\max}/\sigma_{\min})^t - 1}{\sqrt{\ln(\sigma_{\max}/\sigma_{\min})}} I + x_0 \quad (38)$$

Thus, the mean of y_t is given by

$$\bar{y}_t = \mathbb{E}[y_t] = \Phi\mathbb{E}[x_t] \quad (39)$$

$$\bar{y}_t = \left[\frac{2\sigma_{\min}((\sigma_{\max}/\sigma_{\min})^t - 1)}{\sqrt{\ln(\sigma_{\max}/\sigma_{\min})}}\right]\Phi I + y_0 \quad (40)$$

When simulating the unconditional reverse-time SDE for sampling, the mean of $x_{t_{i-1}}$ is calculated without the addition of Gaussian noise. To illustrate this approach, we can consider the Euler Maruyama predictor as an example:

$$\bar{x}_{t_{i-1}} = x_{t_i} - \frac{f(t_i)x_{t_i}}{N} + \frac{g(t_i)^2}{N} s_{\theta^*}(x_{t_i}, t_i) \quad (41)$$

Our objective through the proximal optimization step is to project x_{t_i} onto the manifold defined by $\{x_{t_i} \in \mathbb{R}^n \mid y_{t_i} = \Phi x_{t_i}\}$. Since both x_{t_i} and y_{t_i} are stochastic variables, we consider the projection of their expectations (mean) instead. This leads to the following update rule:

$$\bar{x}_{t_{i-1}} = x'ti - \frac{f(t_i)x'_{t_i}}{N} + \frac{g(t_i)^2}{N} s_{\theta^*}(x'_{t_i}, t_i) \quad (42)$$

$$\bar{x}'_{t_{i-1}} = \bar{x}_{t_{i-1}} + \Phi^T(\Phi\Phi^T)^{-1}(\bar{y}_{t_{i-1}} - \Phi\bar{x}_{t_{i-1}}) \quad (43)$$

$$x'_{t_{i-1}} = \bar{x}'_{t_{i-1}} + \frac{g(t_i)z_i}{\sqrt{N}} \quad (44)$$

such that the intermediate parameter $x'_{t_{i-1}}$ represents the projection of $x_{t_{i-1}}$ onto the manifold defined by y_{t_i}. While this method leads to only minor differences in peak signal-to-noise ratio (PSNR) compared to previous techniques, it provides valuable insight into the quality of the noisy reconstructions generated using our approach.

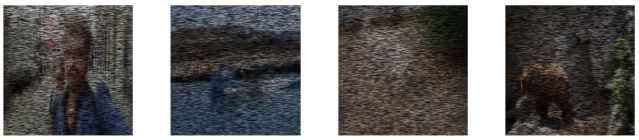
Since the proximal optimization step involves projecting x_{t_i} onto the manifold defined by y_{t_i}, subsequent updates are performed towards the projection $x'_{t_{i-1}}$ rather than x_{t_i} itself. This results in an entirely new Markov process that is no longer described by the VESDE. Unfortunately, due to our limited knowledge and time constraints, we have yet to identify tractable forward and reverse-time expressions for this process. However, the one-to-many mapping from the noisy prior distribution to the data distribution suggests the potential use of non-Gaussian perturbation kernels to further constrain the mapping to approximately one-to-one. This may require a modification of the pre-existing training objective of the score model.

D. Integration with Pre-existing Methods

Consider the case where a single video frame is captured in a singular measurement. Here, y_0 signifies a sparse measurement of the underlying video frame x_0. In such a setup, our algorithm manages to achieve a notably higher peak signal-to-noise ratio (PSNR) of approximately 33 dB. As depicted in Figure 1, disparities between the original image and its reconstruction are minimal, though some minor details like the bear's fur (highlighted in red) exhibit slight loss upon closer examination.

From this observation, we deduce that enhanced reconstruction might be attainable by initially approximating the

Fig. 1: Reconstruction (right) of an image (left) from its snapshot measurement.



(a) (b) (c) (d)

Fig. 2: The observed measurements corresponding to video data (a) `Human`, (b) `Boat`, (c) `Dog`, and (d) `Bear`, respectively.

unmasked pixel values, denoted as $\widehat{x}_0$, for each video frame in snapshot compressive imaging. This preliminary approximation can be achieved through existing algorithms like GAP-TV [6]. Subsequently, the masked pixel values are reconstructed using our score model. This results in a modification to the sampling process as follows:

$$x'_{t_i} = x_{t_i} + \lambda\varphi(\hat{x}_0 - x_{t_i}) \tag{45}$$

$$\bar{x}_{ti-1} = q(xt_i{}', z_i, s_{\theta^*}(x'_{t_i}, t_i)) \tag{46}$$

In this context, φ takes the form of $\text{Diag}(\text{vec}(M_1, M_2, \ldots, M_t))$, where $\varphi \in \mathbb{R}^{N_x \times N_y \times N_t}$ represents a diagonal matrix formed by extracting elements from the vectorized representation of binary modulation masks, denoted as M. This refined approach is labeled as Score-GAP-TV and is presented in Table I.

IV. Experiments

A. Experimental Setting

1) Greyscale reconstruction: Initially, our project concentrated on the reconstruction of grayscale video frames. To circumvent the need for extensive training time to create a score model from the ground up, we employed pre-trained model checkpoints available in [15]. Nonetheless, these checkpoints were trained using RGB images sourced from the CIFAR 10 and FFHQ datasets, whereas the grayscale testing images provided for CACTI had dimensions of 256×256. To accommodate the RGB score model's input dimensions, we replicated the images across their color channel axis, then converted the output back to grayscale by averaging pixel values over color channels. Unfortunately, the computed scores for each color channel diverged, causing distortion in grayscale images. Consequently, we shifted to utilizing RGB inputs directly.

In our experimental comparisons, we opted for GAP-TV due to its ease of implementation, reasonable accuracy, and speed. However, the implementation accessible to us solely supported grayscale inputs, necessitating its re-implementation to accommodate RGB inputs.

Algorithm	PSNR (dB)	No. iter	Time (min)
Conditional reverse-time SDE	16.61	2000	14.16
Proximal optimization	17.89	2000	13.76
Modified proximal optimization	22.89	2000	13.25
Score-TV-GAP	23.09	2000	13.82
GAP-TV (*)	23.74	40	0.08

TABLE I: Quantitative evaluation of proposed algorithms.

Data	No. iter	PSNR	Time (min)
Boat	2200	34.12	15.11
Human	2200	33.68	15.05
Bear	2200	28.52	15.04
Dog	2200	34.14	15.13

TABLE II: PSNR values corresponding to the reconstruction of different images with our algorithm in the CS problem setting.

2) Custom Dataset and Model Training: For optimal performance of the score model, it was empirically observed that training images should exhibit scene compositions similar to the testing images. Unfortunately, none of the provided pre-trained checkpoints had images with dimensions and compositions matching those of the CACTI testing video frames. To tackle this, we generated custom testing datasets from DAVIS 2017 by subjecting its training video frames to a simulated SCI system with modulation masks identical to the original CACTI dataset. However, initial attempts at training the score model using the custom dataset proved unsuccessful due to unstable training, premature convergence, and extended training times. Consequently, we had to reduce the batch size from 64 to 8 owing to hardware memory constraints. The training process took around 10 hours for 50,000 iterations—noticeably less than the 1,300,000 iterations for convergence outlined in [15] with optimal hyperparameters. Moreover, due to time constraints, hyperparameter tuning through trial and error was unfeasible.

B. Evaluation and Analysis of Reconstruction Results

1) Conditional Reverse-Time SDE: The performance of the conditional reverse-time SDE method is detailed in Table 1. The average PSNR for other methods, excluding ours and GAP-TV (*), is sourced from [24]. GAP-TV (*) indicates our attempt to replicate results in [24], which was hampered by scarce hyper-parameter information and potential implementation errors. Each algorithm ran on distinct hardware, thus the time taken isn't indicative of computational complexity. Furthermore, average PSNR for alternative algorithms, except ours and GAP-TV (*), were appraised on different datasets.

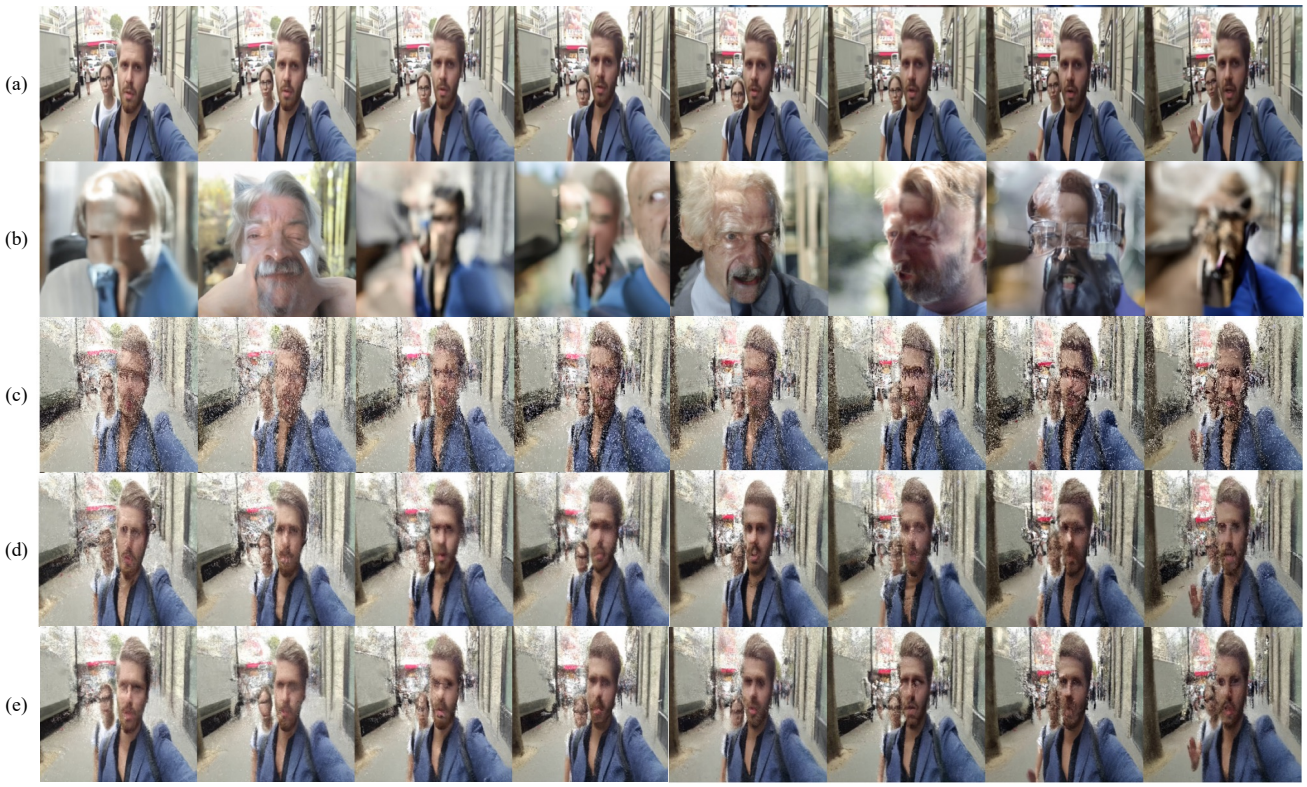


Fig. 3: (a) The original video frames in the video `Human`, and reconstruction results of the measurement using (b) conditional reverse-time SDE, (c) proximal optimization, (d) modified proximal optimization, (e) integration with pre-existing methods, respectively.

Fig. 4: (a) The original video frames in the video `Boat`, and reconstruction results of the measurement using (b) proximal optimization, (c) modified proximal optimization, (d) integration with pre-existing methods, respectively.

Our sampling process utilized an NVIDIA GTX 3090 without computational optimization, yielding 8 video frame samples. The employed score model checkpoint was trained on the Flickr Faces HQ (FFHQ) dataset. Original video frames (rows 1 and 3) and corresponding reconstructions (rows 2 and 4) are presented in Figure 3.

The major factors causing significantly distinct scene compositions in the reconstructed video frames compared to the original stem from introducing Gaussian noise to the measurement y, forming the tractable stochastic process $\{y_t\}_{t=0}^1$. The reconstructions, $\{x_t\}_{t=0}^1$, are no longer constrained solely by y but by $\{y_t\}_{t=0}^1$. Although noise diminishes over time, $\{y_t\}_{t=0}^1$ acts as a weaker constraint on $\{x_t\}_{t=0}^1$, especially early in the sampling process. In score-based generative methods, larger features often emerge in initial sampling stages, setting scene composition and influencing later-stage details.

Score models aren't directly trained to approximate the data, x_0, but its score, $\nabla_{x_t} \log p(x_t)$. This enables broad mode coverage via Langevin sampling [25]. Consequently, generated scene compositions differ from the desired. While the conditioning term, $\nabla \log p(y_t|x_t)$ ensures linear combinations of sparse measurements $x_{t,k}$ align with y_t, it lacks per-frame-based model penalization. This effect is evident in Figure 3. The score model, trained on human faces, aims to generate human faces that align with our observed measurement rather than identifying the right scene composition. Additionally, 2D generative score models fail to capture inter-frame temporal features, treating each frame as an independent image. Despite the overall structural differences leading to a low PSNR, the output is highly detailed and smooth.

2) Proximal Optimization: Table 1 displays the performance of the proximal optimization method using $\lambda = 0.9$, which greatly boosts peak signal-to-noise ratio (PSNR) while trimming computation compared to alternatives. Figure 3 showcases a sample reconstruction using our adapted proximal optimization approach, aligning more closely with the observed measurement y in contrast to our previous conditional reverse-time SDE method, thanks to its strict linear constraint. Despite the higher PSNR achieved, proximal optimization-generated samples exhibit noise and artifacts relative to the prior method. This seems to result from the divergence of x_t through the proximal optimization step, breaking away from the state trajectories dictated by the VESDE. Consequently, the score model struggles to correctly approximate the posterior score at time t_i using the intermediate parameter x'_{t_i}, which deviates from its previously-known state x_{t_i}.

Understanding the poor reconstruction quality involves revisiting the single video frame scenario discussed in Section III-D, achieving markedly higher PSNR. This suggests that the substantial PSNR drop in our methods stems from the inability to perfectly recover pixel values at unmasked locations in multi-frame settings, unlike in single-frame settings. Further analysis reveals that the single-frame setup is akin to (i) compressed sensing without applying any basis transformation before sparse measurements (ii) image in-painting with uniformly distributed missing pixels. In such cases, generative score models have demonstrated state-of-the-art reconstruction [15], [21].

The difference between these problems and SCI lies in the presence of reference points: available unmasked pixel values (for in-painting) or scene composition preservation amid noise (for compressed sensing with basis transformation, e.g., DCT). These reference points constrain the state trajectories in the sampling process, ensuring consistency with measurements and enabling pre-trained models to generalize to novel image types. In compressed sensing for medical imaging, reference points offer a tractable solution to Equation (8) in [21]. This simultaneously minimizes distances between x'_{t_i} and x_{t_i} alongside $\hat{x}'_{t_i}$ and the linear manifold $\{u_{t_i} \in \mathbb{R}^n \mid y_{t_i} = \Phi u_{t_i}\}$ without retraining.

However, in SCI, measurements correspond to linear combinations of unmasked pixel values across multiple frames. In all our algorithms, reference points are computed by either (i) scaled difference ($\beta(t)$-scaled noise difference between y_t and Φx_t) or (ii) scaled difference (λ-scaled difference between y_0 and Φx_t). These differences are then averaged over all unmasked pixels across frames for each pixel location i, represented as $count_i = \sum_k \mathbb{I}[M_{k,i}]$, with $\mathbb{I}$ as the indicator function and M_k as the sub-sampling mask for frame k. This corresponds to L2 loss minimization for each unmasked pixel. In scenes with significant inter-frame changes, the estimates deviate from actual values. These approximated pixel locations function as reference points guiding the sampling process. Errors in these reference points result in incorrect predictions of surrounding pixels by the score model, further compounding through the sampling process.

3) Integration with Pre-existing Methods: Table 1 presents the performance of integrating our score-based generative approach with GAP-TV, using $\lambda = 0.8$. The disparity in peak signal-to-noise ratio (PSNR) between this and prior methods is negligible, yet the reconstructions exhibit enhanced smoothness, evident in Figure 4. Originally, the algorithm was anticipated to yield heightened PSNR across video frames approximated solely by GAP-TV. Surprisingly, we noted that the reconstruction process led to a lower PSNR over approximated video frames. This observation aligns with our earlier argument that errors in reference pixel approximation translate to inaccurate predictions of surrounding masked pixels. As a consequence, the PSNR of reconstructed video frames becomes constrained by the PSNR of approximated counterparts. From an information theory standpoint, this adheres to the data processing inequality, asserting that predicted pixels' accuracy is limited by approximated reference pixels.

C. Further Observations and Discussions

One key drawback of the Langevin sampling method is the necessity for sampling in small time intervals, facilitating the approximation of reverse-time perturbation kernels as tractable Gaussians. This requirement mandates an equivalent number of steps for both forward perturbation and reverse-time sampling. However, when both processes are understood through the score-based generative framework proposed in

Fig. 5: (a) The original video frames in the video `Dog`, and reconstruction results of the measurement using (b) proximal optimization, (c) modified proximal optimization, (d) integration with pre-existing methods, respectively.

Fig. 6: (a) The original video frames in the video `Bear`, and reconstruction results of the measurement using (b) proximal optimization, (c) modified proximal optimization, (d) integration with pre-existing methods, respectively.

[15], the shift from discrete-time Markov chains to continuous Markov processes inherently satisfies this condition via infinitesimal time steps. Only under these conditions can a workable objective function, as outlined in (13), be formulated to train the score model. Notably, reducing iterations in the reverse process adversely impacts the reconstruction quality in our conditional reverse-time stochastic differential equation (SDE) method. Nonetheless, for the proximal optimization method, the reconstruction quality variation between 2000 and 200 iterations is minimal. This is due to the deviation of x_t from VESDE, breaking the Gaussian perturbation kernel assumption and leading to suboptimal reconstructions. We substituted VESDE with Variance-Preserving Stochastic Differential Equation (VPSDE), the continuous-time counterpart to Diffusion Denoising Probabilistic models (DDIM) [26]. VPSDE employs 1000 noise scales and a distinct sampling schedule starting at $t \in (0.6, 1]$. With the proximal optimization step, the reconstruction quality at 1000 and 400 iterations is marginal.

From our experiments, we conclude that the proximal optimization step ensures consistency with y at the expense of subpar and noisy reconstruction quality. Conversely, the conditional reverse-time SDE generates higher-quality samples that structurally diverge from the intended data x. We also established that integrating pre-trained 2D generative score models into existing frameworks, such as plug-and-play for snapshot compressive imaging inverse problems, is intricate. Crafting such an algorithm would likely involve formulating a specialized class of stochastic differential equations that follow state trajectories linearly constrained by the snapshot measurement and consequently, a narrower mode coverage. This would likely necessitate distinct objective functions and training procedures from those applied here. We conjecture that adopting 3D diffusion models might yield significantly improved outcomes, as they can capture temporal information among video frames, capitalizing on the structural similarity of video frames to identify accurate scene compositions. Unlike conventional denoisers that approximate the target scene from measurement y, limiting recoverable detail, generative score methods theoretically have the potential to construct scene compositions akin to the target from the ground up (Gaussian white noise) and filling in necessary details to achieve higher quality reconstructions.

V. Future Work

A. Processing Speed

Recent advancements in score-based generative modeling have proposed strategies to enhance sampling efficiency, primarily focusing on two avenues: utilizing non-Markovian processes or leveraging non-Gaussian perturbation kernels. Nichol et al. [27] introduced an innovative sampling scheme that updates at intervals of T/S, introducing a new schedule $\tau_1, \ldots, \tau_s$ with $\tau_1 < \ldots < \tau_s$ and $S < T$. This is accomplished through a deterministic, non-Markovian process. The notion of non-Gaussian perturbation kernels, as presented in [25] by Xiao et al. in the context of Denoising Diffusion GANs, is also significant. This concept relaxes the need for infinitesimally small time steps by employing GANs to approximate non-Gaussian, multimodal transitional distributions. However, this approach yields a many-to-one mapping from clean images to their noisy prior distributions, leading to heightened sample diversity, which is not ideal for SCI. Furthermore, diminishing time steps while retaining Gaussian perturbation kernels compromises reconstruction quality. Therefore, for future endeavors, we recommend the adoption of non-Gaussian perturbation kernels.

B. Reconstruction Quality

The trilemma in image or video frame reconstruction involves balancing the generation of high-quality samples, increasing sampling speed, and improving mode coverage/sample diversity [25]. Two prevalent methods for generating high-quality samples are Generative Adversarial Networks (GANs) and score-based methods. Unlike GANs, score-based methods prioritize mode coverage over sampling speed, which is a drawback. However, for SCI, extensive mode coverage might lead to unfavorable outcomes due to excessive sample diversity. It is expected that single-shot models like GANs could provide better Peak Signal-to-Noise Ratio (PSNR) with significantly faster processing.

The stochastic nature of traditional score-based methods, induced by added noise, is undesirable. Implementing post-processing denoising techniques [28] is an option for enhancing reconstruction quality.

Another strategy is to tailor the approach for specific SCI applications. Training a score model on video frames with varying scene compositions has shown poor accuracy in previous experiments. By focusing on specific applications, where the overall scene compositions are known, datasets can be curated for training and testing. However, this approach might limit model flexibility.

Finally, we suggest the utilization of 3D U-Nets as score models capable of directly handling video inputs instead of images. These score models employ 3D convolutions to capture both temporal and spatial features. Video Diffusion models [29] use this concept to generate coherent video frames from noise. Additionally, customized 3D U-Nets with gated attention mechanisms, such as FLAVR [30], have effectively captured temporal features for video frame interpolation, though they demand substantial computational resources for training.

VI. Conclusion

Our current best approach achieves a Peak Signal-to-Noise Ratio (PSNR) of approximately 23 dB, requiring around 14 minutes for reconstructing video frames from a single snapshot measurement. These results are subpar compared to other state-of-the-art algorithms like DEQ [31], which achieves around 30 dB PSNR at notably faster speeds. While we anticipate a 2 to 3 dB PSNR improvement with finely tuned hyperparameters, it remains insufficient compared to other methods. Hence, greater emphasis should be placed on

devising new forward perturbation and reverse-time stochastic processes to attain practical speeds and heightened accuracies. The exploration of non-Markovian processes and non-Gaussian perturbation kernels is warranted. Although 3D Diffusion models are viable, they entail high computational costs. Given the nascent nature of score models in deep learning, novel strategies are still emerging to fully leverage their image-generation capabilities. We hope that our project has presented new opportunities for future research in this area.

REFERENCES

[1] Xin Yuan, David J Brady, and Aggelos K Katsaggelos. Snapshot compressive imaging: Theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 38(2):65–88, 2021.

[2] Patrick Llull, Xuejun Liao, Xin Yuan, Jianbo Yang, David Kittle, Lawrence Carin, Guillermo Sapiro, and David J Brady. Coded aperture compressive temporal imaging. *Optics express*, 21(9):10526–10545, 2013.

[3] Stephen P. Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[4] Stanley Chan. Performance analysis of plug-and-play admm: A graph signal processing perspective. *IEEE Transactions on Computational Imaging*, PP:1–1, 01 2019.

[5] Zhiwei (Tony) Qin, Donald Goldfarb, and Shiqian Ma. An alternating direction method for total variation denoising. *Optimization Methods and Software*, 30(3):594–615, 2015.

[6] Xin Yuan. Generalized alternating projection based total variation minimization for compressive sensing, 2015.

[7] Michael Iliadis, Leonidas Spinoulas, and Aggelos K. Katsaggelos. Deep fully-connected networks for video compressive sensing, 2017.

[8] Ziyi Meng, Shirin Jalali, and Xin Yuan. Gap-net for snapshot compressive imaging, 2020.

[9] Jiawei Ma, Xiao-Yang Liu, Zheng Shou, and Xin Yuan. Deep tensor admm-net for snapshot compressive imaging. pages 10222–10231, 10 2019.

[10] Ziheng Cheng, Ruiying Lu, Zhengjue Wang, Hao Zhang, Bo Chen, Ziyi Meng, and Xin Yuan. *BIRNAT: Bidirectional Recurrent Neural Networks with Adversarial Training for Video Snapshot Compressive Imaging*, pages 258–275. 11 2020.

[11] 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, sep 2018.

[12] Qing Yang and Yaping Zhao. Revisit dictionary learning for video compressive sensing under the plug-and-play framework. In *Seventh Asia Pacific Conference on Optics Manufacture and 2021 International Forum of Young Scientists on Advanced Optical Manufacturing (APCOM and YSAOM 2021)*, volume 12166, pages 2018–2025. SPIE, 2022.

[13] Yaping Zhao. Mathematical cookbook for snapshot compressive imaging. *arXiv preprint arXiv:2202.07437*, 2022.

[14] Shirin Jalali and Xin Yuan. Snapshot compressed sensing: Performance bounds and algorithms. *IEEE Transactions on Information Theory*, 65(12):8005–8024, 2019.

[15] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[16] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[18] Brian. D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12:313–326, 1982.

[19] N. Ikeda and S. Watanabe. *Stochastic Differential Equations and Diffusion Processes*. ISSN. Elsevier Science, 2014.

[20] Simo Särkkä and Arno Solin. *Applied Stochastic Differential Equations*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2019.

[21] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021.

[22] Guosheng Lin, Fayao Liu, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for dense prediction. *IEEE transactions on pattern analysis and machine intelligence*, 42(5):1228–1242, 2019.

[23] 4 predictor-corrector methods. In Leon Lapidus and John H. Seinfeld, editors, *Numerical Solution of Ordinary Differential Equations*, volume 74 of *Mathematics in Science and Engineering*, pages 152–241. Elsevier, 1971.

[24] Xin Yuan, Yang Liu, Jinli Suo, and Qionghai Dai. Plug-and-play algorithms for large-scale snapshot compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1447–1457, 2020.

[25] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*, 2021.

[26] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[27] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *CoRR*, abs/2102.09672, 2021.

[28] Yaping Zhao, Haitian Zheng, Zhongrui Wang, Jiebo Luo, and Edmund Y Lam. Manet: improving video denoising with a multi-alignment network. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2036–2040. IEEE, 2022.

[29] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

[30] Tarun Kalluri, Deepak Pathak, Manmohan Chandraker, and Du Tran. Flavr: Flow-agnostic video representations for fast frame interpolation, 2022.

[31] Yaping Zhao, Siming Zheng, and Xin Yuan. Deep equilibrium models for video snapshot compressive imaging. *arXiv preprint arXiv:2201.06931*, 2022.