# 3One2: One-step Regression Plus One-step Diffusion for One-hot Modulation in Dual-path Video Snapshot Compressive Imaging

**Ge Wang**[1,2], **Xing Liu**[3], **Xin Yuan**[2,3,*]

[1]Zhejiang University, Hangzhou, Zhejiang 310058, China
[2]Westlake University, Hangzhou, Zhejiang 310030, China
[3]Westlake Institute for Optoelectronics, Hangzhou, Zhejiang 311421, China
{wangge,xyuan}@westlake.edu.cn

## Abstract

Video snapshot compressive imaging (SCI) captures dynamic scene sequences through a two-dimensional (2D) snapshot, fundamentally relying on optical modulation for hardware compression and the corresponding software reconstruction. While mainstream video SCI using random binary modulation has demonstrated success, it inevitably results in temporal aliasing during compression. One-hot modulation, activating only one sub-frame per pixel, provides a promising solution for achieving perfect temporal decoupling, thereby alleviating issues associated with aliasing. However, no algorithms currently exist to fully exploit this potential. To bridge this gap, we propose an algorithm specifically designed for one-hot masks. First, leveraging the decoupling properties of one-hot modulation, we transform the reconstruction task into a generative video inpainting problem and introduce a stochastic differential equation (SDE) of the forward process that aligns with the hardware compression process. Next, we identify limitations of the pure diffusion method for video SCI and propose a novel framework that combines one-step regression initialization with one-step diffusion refinement. Furthermore, to mitigate the spatial degradation caused by one-hot modulation, we implement a dual optical path at the hardware level, utilizing complementary information from another path to enhance the inpainted video. To our knowledge, this is the first work integrating diffusion into video SCI reconstruction. Experiments conducted on synthetic datasets and real scenes demonstrate the effectiveness of our method.

## Introduction

Capturing high-speed temporal events is crucial across various scientific and technological fields, including fluid dynamics (Adrian and Westerweel 2011), biomechanics (Lieberman et al. 2010), and materials science (Parab et al. 2019). Conventional high-speed camera imaging techniques face prohibitive hardware and storage transmission costs. Inspired by compressive sensing (Candès, Romberg, and Tao 2006; Donoho 2006), video snapshot compressive imaging (SCI) (Yuan, Brady, and Katsaggelos 2021) employs optical hardware to multiplex a sequence of video frames, each encoded with a distinct modulation pattern (hereafter referred to as a mask), into a single snapshot measurement on a two-dimensional (2D) detector. Additionally, it incorporates hardware-adapted algorithms to reconstruct the original video scene. Within this hardware-encoder plus software-decoder framework, video SCI system achieves orders of magnitude improvements in temporal resolution.

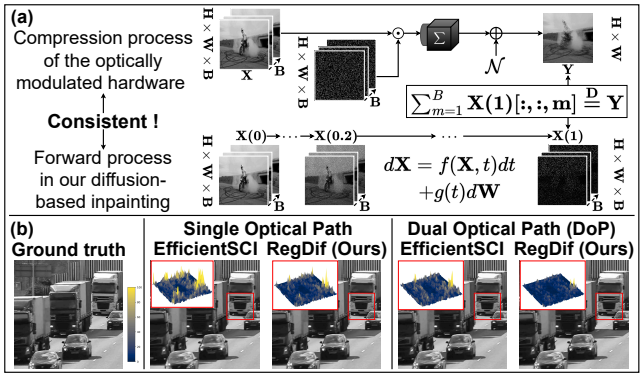


Figure 1: (a) The forward process in our diffusion-base inpainting aligns with the hardware compression process in video SCI. (b) Our method shows superiority in both single and dual-path settings. The 3D heatmap represents the absolute error between the red block and the ground truth.

The performance of video SCI primarily depends on the encoding mask and the corresponding reconstruction algorithm. Mainstream SCI systems (Qiao, Liu, and Yuan 2020; Qiao et al. 2020) typically employ random binary masks, which are physically implemented via digital micromirror devices (DMDs) for hardware encoding. Despite their success, this approach inevitably results in *temporal aliasing during compression*. Specifically, multiplexing of the temporal channel in the encoding process leads to the superposition of information from different frames in the compressed measurement. As illustrated in Fig. 2(a) histogram, different frames in the original scene exhibit distinct colors; however, after modulation by the random binary mask, all colors are blended in the measurement. Under the one-hot mask configuration, the DMD reflects only one sub-frame's signal per spatial location to the sensor during compression. This ensures perfectly decoupled temporal information, where each measurement pixel corresponds exclusively to a single original frame, thus eliminating temporal superposition.

*Corresponding author

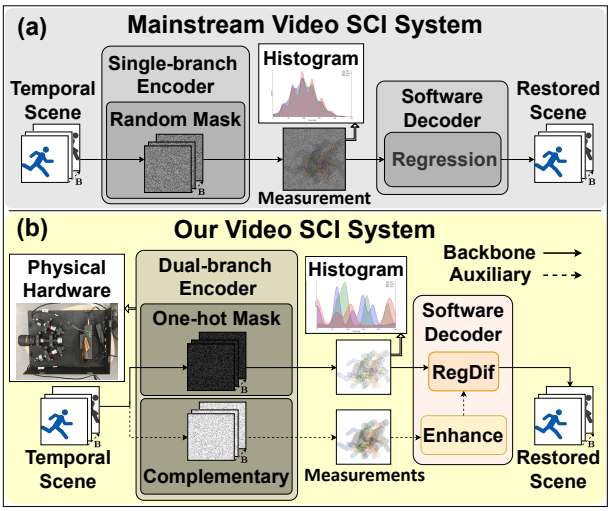


Figure 2: (a) Overall framework of a single-branch system using a random binary mask and its measurement pixel histogram. (b) Overall framework of a dual-branch system using a one-hot mask and its measurement pixel histogram.

As shown in Fig. 2(b) histogram, the measurement retains distinct inter-frame color separation. Furthermore, one-hot mask enables full-dynamic-range measurement capture, preventing bit-depth clipping and proving particularly advantageous in low-light and long-exposure scenarios, as well as in deploying a complete video SCI system on a chip.

Currently, no algorithms are specifically designed for one-hot masks to fully leverage their ability to decouple temporal information (Qiao and Yuan 2022). To fill this gap, we reformulate reconstruction as a generative inpainting task and propose a video SCI-adapted stochastic differential equation (SDE) of the forward process that mirrors the hardware optically modulated encoding, as shown in Fig. 1(a). For our video SCI inpainting input, masked regions dominate each frame's spatial domain. To overcome degradation caused by large masked regions and avoid slow iterative sampling, we develop a hybrid framework dubbed RegDif, where a single regression block provides the coarse estimate, followed by the one-step diffusion for refinement. To our best knowledge, this represents the first application of a diffusion model to video SCI. Additionally, to further mitigate spatial degradation resulting from the one-hot mask, we introduce a dual optical path: the original path captures modulated signals whose corresponding one-hot mask values are 1, while a compensatory path with another 2D sensor records discarded signals with mask values of 0. A fusion module then enhances the inpainted video using the compensatory-path measurement. As shown in Fig. 1(b), selected frames of reconstructed videos show the superiority of our method.

In summary, we propose a novel video SCI system as shown in Fig. 2(b), and the contributions of this work are:

- Leveraging the temporal decoupling property of the one-hot mask, we transform the regression reconstruction problem into a diffusion-based inpainting task and propose a customized SDE of the forward process that aligns with the hardware compression process in video SCI.
- We observe that the pure diffusion method yields unsatisfactory reconstruction performance and speed in video SCI tasks characterized by severe degradation. To address these, we propose a hybrid framework employing one-step regression to obtain a coarse estimate, followed by one-step diffusion refinement. Additionally, we introduce a dual optical path to enhance the inpainted video.
- We conduct comprehensive experiments to demonstrate the exceptional performance of the proposed reconstruction framework tailored for the one-hot mask under both single and dual-optical-path configurations.

## Related Work

### Video Snapshot Compressive Imaging

In the hardware component of video SCI, a random binary mask is commonly implemented by either a digital micromirror device (DMD) (Qiao et al. 2020; Qiao, Liu, and Yuan 2020) or liquid crystal on silicon (LCOS) (Hitomi et al. 2011; Liu et al. 2013). To overcome issues caused by temporal aliasing, learned binary masks via programmable pixel sensors (Martel et al. 2020) and structural masks with reduced refresh rates (Wang, Wang, and Yuan 2023) have been proposed. Additionally, optical designs that capture two objects in a single measurement (Qiao, Liu, and Yuan 2020) and dual paths for complementary measurements of one object (Wang, Liu, and Yuan 2025) have been explored. In this study, we are the first to integrate the one-hot mask with the dual optical path to decouple temporal information and improve light efficiency to mitigate the degradation issue.

In the reconstruction process, operating under default random binary mask settings, traditional model-based methods employ various regularizations, including Total Variation (Yuan 2016), Gaussian Mixture Model (Yang et al. 2014), and Low Rank (Liu et al. 2018). Plug-and-play (PnP) frameworks (Yuan et al. 2020, 2021) integrate pretrained deep image predictors into iterative optimizations for flexibility. For rapid reconstruction with high quality, deep learning-based methods (Cheng et al. 2020; Wang, Cao, and Yuan 2023; Cheng et al. 2021; Ma et al. 2019; Wang et al. 2021, 2022) have been introduced. BIRNAT (Cheng et al. 2020) utilizes a bidirectional recurrent neural network for reconstruction, while RevSCI (Cheng et al. 2021) employs a 3D CNN-based memory-efficient framework. Additionally, STFormer (Wang et al. 2022) and EfficientSCI (Wang, Cao, and Yuan 2023) leverage spatial-temporal transformers and spatial CNNs with temporal transformers, respectively. All previously mentioned deep learning-based methods depend on regression reconstruction. In contrast, we propose the first diffusion-based method tailored for the one-hot mask.

### Video Inpainting

Video inpainting seeks to restore gaps or missing regions with visually consistent content while maintaining spatial and temporal coherence. Critical techniques include flow-based propagation (Gao et al. 2020; Li et al. 2022; Zhang, Fu, and Liu 2022; Zhou et al. 2023) and video transformers (Liu et al. 2021; Li et al. 2022; Zhang, Fu, and Liu

2022), with Propainter (Zhou et al. 2023) integrating both in the architecture. With advancements in diffusion models, recent studies have explored the application of image diffusion models for video editing (Ceylan, Huang, and Mitra 2023; Geyer et al. 2023; Qi et al. 2023; Wu et al. 2023), such as AVID (Zhang et al. 2024) using motion modules and structure guidance. These methods emphasize the temporal continuity of masked objects. In contrast, our video SCI inpainting employs temporally incoherent one-hot masks. Inspired by IR-SDE (Luo et al. 2023), which is used for image restoration, we propose an SDE of the forward process that aligns with the video SCI encoding process and a tailored reconstruction method for severely degraded videos.

# Methodology

## Mathematical Model of Video SCI

All subsequent discussions will focus on grayscale videos, while color processing is illustrated in Appendix D.

As shown in Fig. 2(a), in the single-path hardware encoder with random binary mask, the original $B$ frame input video $\boldsymbol{X} \in \mathbb{R}^{H\times W\times B}$ is modulated by the mask $\tilde{\boldsymbol{M}} \in \mathbb{R}^{H\times W\times B} \sim \text{Bernoulli}(0.5)^{H\times W\times B}$. Subsequently, by compressing data across time, the 2D camera detector captures a compressed measurement $\tilde{\boldsymbol{Y}} \in \mathbb{R}^{H\times W}$. The entire encoding process can be represented as follows:

$$\tilde{\boldsymbol{Y}} = \textstyle\sum_{m=1}^{B} \boldsymbol{X}_m \odot \tilde{\boldsymbol{M}}_m + \tilde{\boldsymbol{Z}},$$

where $\odot$ denotes the Hadamard (element-wise) multiplication. The notation $\boldsymbol{X}_m := \boldsymbol{X}[:,:,m]$ represents the $m$-th frame of the input video, while $\tilde{\boldsymbol{M}}_m := \tilde{\boldsymbol{M}}[:,:,m]$ indicates the corresponding modulation for the $m$-th frame of the video. Additionally, $\tilde{\boldsymbol{Z}} \in \mathbb{R}^{H\times W}$ denotes the measurement noise. Subsequently, a regression-based reconstruction algorithm, denoted as $\mathcal{D}^*_{\text{reg}} : \mathbb{R}^{H\times W} \times \mathbb{R}^{H\times W\times B} \to \mathbb{R}^{H\times W\times B}$, is employed to recover the original video, determined by:

$$\mathcal{D}^*_{\text{reg}} = \arg\min_{\mathcal{D}_{\text{reg}}} ||\boldsymbol{X} - \mathcal{D}_{\text{reg}}(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{M}})||_F.$$

In our encoding, the one-hot mask is generated through:

$$\boldsymbol{M}[h,w,m] = \begin{cases} 1, & \text{if } \lfloor \mathcal{I}[h,w] \rfloor = m, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $\lfloor\cdot\rfloor$ denotes the floor function, which rounds down to the nearest integer, and $\mathcal{I} \sim \text{Uniform}(1, B+1)^{H\times W}$. According to the mask defined in Eq. (1), at any spatial position $(h,w)$, only one temporal channel has a value of 1 while all others are 0, which is why we refer to it as a one-hot mask.

In addition to the one-hot mask $\boldsymbol{M}$, our encoder incorporates a compensatory path alongside the original path to capture signals associated with mask values of 0. Our optical design enables the measurement $\boldsymbol{Y}^C$ from the compensatory path to capture signals that are not detected by the measurement $\boldsymbol{Y}$ from the original path, thereby positioning $\boldsymbol{Y}^C$ as a complement to $\boldsymbol{Y}$. Consequently, the encoding process in our dual optical path system is represented as follows:

$$\boldsymbol{Y} = \textstyle\sum_{m=1}^{B} \boldsymbol{X}_m \odot \boldsymbol{M}_m + \boldsymbol{Z}, \tag{2}$$

$$\boldsymbol{Y}^C = \textstyle\sum_{m=1}^{B} \boldsymbol{X}_m \odot (\boldsymbol{1} - \boldsymbol{M}_m) + \boldsymbol{Z}^C, \tag{3}$$

where $\boldsymbol{Z}/\boldsymbol{Z}^C$ is the corresponding measurement noise in the original/compensatory path, and $\boldsymbol{1} \in \mathbb{R}^{H\times W}$ is an all-one matrix. More hardware encoder details are in Appendix A.

We observe that using one-hot mask $\boldsymbol{M}$ can separate the temporal information of different frames within the measurement $\boldsymbol{Y}$ from Eq. (2). For instance, the non-zero components of $\boldsymbol{Y} \odot \boldsymbol{M}_m$ represent the pixel values at the corresponding positions in the $m$-th frame of the input video with the measurement noise (commonly modeled as a Gaussian distribution). Consequently, it is apt to employ diffusion-based inpainting for reconstruction. Intuitively, we consider the following SDE of the forward process in diffusion:

$$d\boldsymbol{X} = \underbrace{\mu(t)(\boldsymbol{X}_{\text{dst}} - \boldsymbol{X}(t))dt}_{\text{drift}} + \underbrace{\sigma(t)d\boldsymbol{W}}_{\text{dispersion}}, \tag{4}$$

with the boundary condition constraint $\boldsymbol{X}_{\text{src}} := \boldsymbol{X}(0) = \boldsymbol{X}$ and $t \in [0,1]$, where $\mu(t), \sigma(t)$ are time-dependent hyperparameters that characterize speed and stochastic volatility, and $\boldsymbol{X}_{\text{dst}} = \boldsymbol{X} \odot \boldsymbol{M}$, where $\boldsymbol{W}$ represents the Wiener process. It can be observed that the drift term is analogous to the interpolation method that gradually transforms $\boldsymbol{X}_{\text{src}}$ into $\boldsymbol{X}_{\text{dst}}$. Additionally, the role of the dispersion term is to introduce Gaussian noise during this process, serving a similar function as the measurement noise $\boldsymbol{Z}$ in Eq. (2). Fortunately, our investigation reveals that similar SDE methods have been thoroughly examined in prior research (Luo et al. 2023), which focused on image restoration. Consistent with prior research, we discover that if $\mu(t)$ and $\sigma(t)$ are constrained by $\sigma^2(t)/\mu(t) = 2\lambda^2$, where $\lambda^2$ is the stationary variance, Eq. (4) yields a closed-form solution:

$$\boldsymbol{X}(t) = [(1-\bar{\mu}(t))\boldsymbol{X}_{\text{src}} + \bar{\mu}(t)\boldsymbol{X}_{\text{dst}}] + \bar{\sigma}(t)\epsilon, \tag{5}$$

where $\bar{\mu}(t) = 1 - e^{-\theta(t)}$ is the interpolation factor, $\bar{\sigma}(t) = \lambda\sqrt{1-e^{-2\theta(t)}}$, $\theta(t) = \int_0^t \mu(s)ds$ , and $\epsilon \sim \mathcal{N}(0,1)^{H\times W\times B}$. By appropriately configuring $\lambda$ and $\mu(t)$ (where $\sigma(t)$ can be determined by $\lambda\sqrt{2\mu(t)}$), we can achieve the following approximations: as $t$ progresses from 0 to 1, $\bar{\mu}(t)$ transitions from 0 to 1, and $\bar{\sigma}(t)$ transitions from 0 to $\lambda$. This configuration ensures that the diffusion forward process aligns with the video SCI hardware encoding process in Eq. (2). Specifically, it is easy to notice that $\boldsymbol{X}(0)$ is consistent with the input video $\boldsymbol{X}$. Assuming the measurement noise $\boldsymbol{Z} \sim \mathcal{N}(0, \tilde{\lambda}^2)$ (for simplicity, we consider the case where the mean of the noise is zero; in scenarios where a non-zero mean is considered, a modification to $\boldsymbol{X}_{\text{dst}}$ is required), when we set $\lambda = \tilde{\lambda}/\sqrt{B}$, compressed result from $\boldsymbol{X}(1)$ alongside the temporal channel is consistent with $\boldsymbol{Y}$, such that $\sum_{m=1}^{B} \boldsymbol{X}(1)[:,:,m] \overset{\text{distribution}}{=} \boldsymbol{Y}$. Thus, we can characterize the process of compressing the original video $\boldsymbol{X}$ into $\boldsymbol{Y}$ using the video SCI hardware encoder via the diffusion forward process from $\boldsymbol{X}(0)$ to $\boldsymbol{X}(1)$. A more formal statement and the detailed proof are shown in Appendix B.

We can then reverse the process to restore $X(0)$ from $X(1)$ by backing the SDE in time (Song et al. 2020). The corresponding SDE of the reverse process is given by:

$$(d\boldsymbol{X})_{\epsilon_t} = [\mu(t)(\boldsymbol{X}_{\text{dst}} - \boldsymbol{X}(t)) + \tilde{\sigma}(t)\epsilon_t]dt + \sigma(t)d\hat{\boldsymbol{W}}, \tag{6}$$

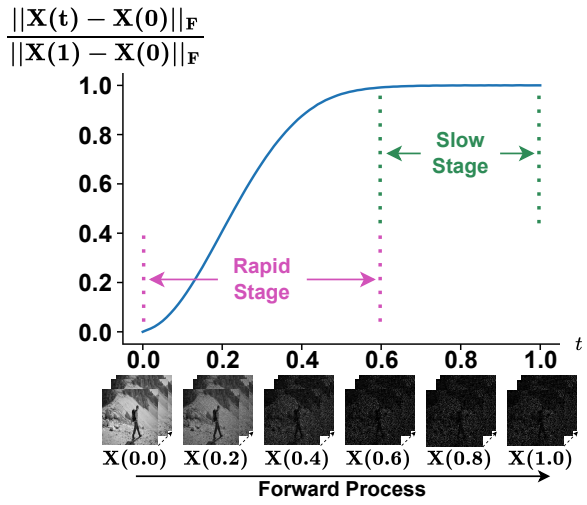


Figure 3: The normalized $\ell_2$-norm distance between $\boldsymbol{X}(t)$ and $\boldsymbol{X}(0)$ as t varies from 0 to 1 in our forward process.

where $\tilde{\sigma}(t) = \sigma(t)^2/\bar{\sigma}(t)$, $\hat{\boldsymbol{W}}$ is a reverse-time Wiener process, and $\epsilon_t$ is the noise we need to approximate by a neural network $\epsilon_{\gamma^*}(\boldsymbol{X}(t), t)$, determined by the trajectory loss:

$$\gamma^* = \arg\min_\gamma \mathbb{E}_{t\sim U(0,1)}[||(d\boldsymbol{X})_t^* - (d\boldsymbol{X})_{\epsilon_\gamma(\boldsymbol{X}(t),t)}||_F], \quad (7)$$

where $(d\boldsymbol{X})_t^*$ denotes the ideal trajectory between $t - dt$ and $t$. In contrast to the commonly used objective function in DDPM (Ho, Jain, and Abbeel 2020), trajectory loss is employed to stabilize training when restoring images affected by complex degradations (Luo et al. 2023). In this study, we retain trajectory loss as degradation in our video inpainting task is more severe than in typical image restoration tasks. After deriving the trained noise predictor $\epsilon_{\gamma^*}(\cdot,\cdot)$, one can use $\mathcal{D}_{\text{reg}}^*$ to retrieve the original video, as shown in Algorithm 1. It is noteworthy that in the practical reverse process, we substitute $\boldsymbol{X}_{\text{dst}}[:,:,m]$, which is unknown, with $\boldsymbol{Y} \odot \boldsymbol{M}_m$ for $m = 1, \ldots, B$, where the error is a higher-order infinitesimal compared with $d\hat{\boldsymbol{W}}$. Further details regarding our diffusion-based inpainting, including the derivation of Eq. (5) from Eq. (4), the acquisition of Eq. (6), and the specifics of Eq. (7), can be found in Appendix B.

## RegDif: One-step Regression + One-step Diffusion

In executing the pure diffusion-based inpainting task described above, we find that the iterative process limits rapid reconstruction and yields suboptimal performance. Our analysis attributes these unsatisfactory reconstruction results to the nonlinear characteristics of the interpolation factor $\bar{\mu}(t)$ in Eq. (5). For instance, in the 100-step diffusion with a simple linear schedule, as illustrated in Fig. 3, we note that after approximately 60 iterations in the forward process, the difference between the current state $\boldsymbol{X}(0.6)$ and the terminal state $\boldsymbol{X}(1)$ becomes negligible. Consequently, during the reverse process, the model learns inconsistent denoising strategies: smooth in the slow stage and fast in the rapid stage. Additionally, the degradation of $\boldsymbol{X}(1)$ caused by the one-hot mask further constrains denoising.

Algorithm 1: Pure Diffusion-based Inpainting $\mathcal{D}_{\text{dif}}^*$

**Input**: Measurement $\boldsymbol{Y} \in \mathbb{R}^{H\times W}$, Mask $\boldsymbol{M} \in \mathbb{R}^{H\times W\times B}$
**Parameter**: Noise Predictor $\epsilon_{\gamma^*}(\cdot,\cdot)$, Iteration Count $N$, Station Variance $\lambda^2$, Discrete Scheduler $\mu = [\mu_1, \ldots, \mu_N]$
**Output**: Reconstructed Video $\boldsymbol{X}$

$\Delta t \leftarrow 1/N$
$H, W, B \leftarrow \text{shape}(\boldsymbol{M})$
$\sigma \leftarrow \lambda\sqrt{2\mu}$
$\theta \leftarrow \text{cumsum}(\mu \cdot \Delta t)$
$\bar{\sigma} \leftarrow \lambda\sqrt{1 - e^{-2\theta}}$
**for** $m = 1, \ldots, B$ **do**
  $\tilde{\boldsymbol{X}}[:,:,m] \leftarrow Y \odot M[:,:,m]$
**end for**
$\boldsymbol{X} \leftarrow \tilde{\boldsymbol{X}}$
**for** $i = N, \ldots, 1$ **do**
  $\hat{\epsilon} \leftarrow \mathcal{N}(0, \Delta t)^{H\times W\times B}$
  $\epsilon \leftarrow \epsilon_{\gamma^*}(\boldsymbol{X}, i \cdot \Delta t)$
  $(\Delta\boldsymbol{X})_\epsilon \leftarrow \left(\mu[i](\tilde{\boldsymbol{X}} - \boldsymbol{X}) + \frac{\sigma[i]^2}{\bar{\sigma}[i]}\epsilon\right)\Delta t + \sigma[i]\hat{\epsilon}$
  $\boldsymbol{X} \leftarrow \boldsymbol{X} - (\Delta\boldsymbol{X})_\epsilon$
**end for**
**return** $\boldsymbol{X}$

To address this, one solution is to design an intricate scheduler $\mu(t)$ with parameter tuning. Instead, we use a simple regressor with one block to replace the slow reverse stage, followed by one-step diffusion denoising. Based on our design, this approach $(i)$ significantly enhances reconstruction speed, $(ii)$ eliminates the need for complex parameter tuning, and $(iii)$ partially mitigates degradation of the masked video via the regressor's coarse prediction.

As shown in Fig. 4, our framework consists of three components: Regressor Initializer, Timestep Predictor, and One-step Denoiser. The Regressor Initializer first smooths the degraded masked video obtained by decoupling temporal information from measurement $\boldsymbol{Y}$ through a *one-hot mask* and updates masked regions using this coarse prediction. Then, the Timestep Predictor derives the timestep of the updated video. After that, the *one-step denoiser* directly denoises the updated video into the final result via Eq. 6 without iteration.

In the Timestep Predictor, we use 3D Convolution with Average Pooling to obtain the current timestep. The Noise Predictor and Regressor Initializer share identical 3D Convolution structures for Feature Extraction and Prediction Head blocks, and both use the same Spatial-temporal ResBlock backbone, differing only in that the Noise Predictor adds Timestep Embedding blocks to predict time-dependent noise. The Regressor Initializer uses only one Spatial-temporal ResBlock, thus referred to as *one-step regression*. Within Spatial-temporal ResBlock, embedded inputs are partitioned along channels, processed through STHB blocks, and fused. Each STHB uses 2D Convolution and 2D State Space Model (SSM) for per-frame local/global information, and Attention for cross-frame temporal information, as shown in Fig. 5. Model details are shown in Appendix C.

In Fig. 4, we present three different loss terms that are

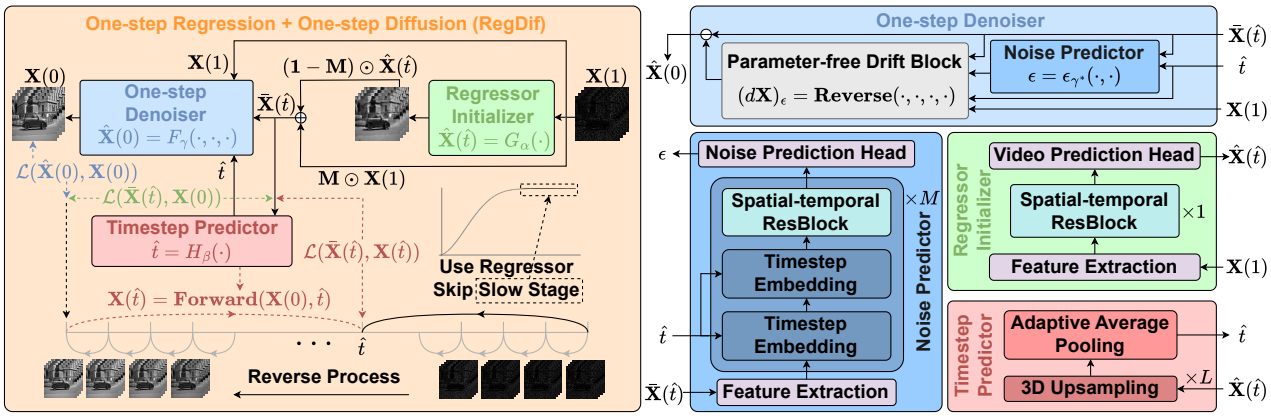


Figure 4: Illustration of the overall RegDif framework (LHS) and key components in RegDif (RHS). In the LHS, black solid line indicates the pipeline of RegDif, gray solid line indicates the reverse process of pure diffusion inpainting. Three colored dash lines indicate pipelines for different loss derivations. The RHS shows architectures of key components mentioned in RegDif.

Algorithm 2: Regress Init + Diffusion Refine $\mathcal{D}^*_{\text{regdif}}$

**Input**: Measurement $\boldsymbol{Y} \in \mathbb{R}^{H\times W}$, Mask $\boldsymbol{M} \in \mathbb{R}^{H\times W\times B}$
**Parameter**: Regressor Initializer $G_{\alpha^*}(\cdot)$, Timestep Predictor $H_{\beta^*}(\cdot)$, Noise Predictor $\epsilon_{\gamma^*}(\cdot,\cdot)$, Station Variance $\lambda^2$, Continuous Scheduler $\mu(\cdot)$
**Output**: Reconstructed Video $\boldsymbol{X}$

$H, W, B \leftarrow \text{shape}(\boldsymbol{M})$
$\sigma(\cdot) \leftarrow \lambda\sqrt{2\mu(\cdot)}$
$\theta(\cdot) \leftarrow \int_0^{\cdot}(\mu(s)ds$
$\bar{\sigma}(\cdot) \leftarrow \lambda\sqrt{1-e^{-2\theta(\cdot)}}$
**for** $m = 1, \ldots, B$ **do**
  $\boldsymbol{X}(1)[:,:,m] \leftarrow Y \odot M[:,:,m]$
**end for**
$\hat{\boldsymbol{X}}(\hat{t}) \leftarrow G_{\alpha^*}(\boldsymbol{X}(1))$
$\bar{\boldsymbol{X}}(\hat{t}) \leftarrow \boldsymbol{M} \odot \boldsymbol{X}(1) + (\boldsymbol{1}-\boldsymbol{M}) \odot \hat{\boldsymbol{X}}(\hat{t})$
$\hat{t} \leftarrow H_{\beta^*}(\bar{\boldsymbol{X}}(\hat{t}))$
$\hat{\epsilon} \leftarrow \mathcal{N}(0,\hat{t})^{H\times W\times B}$
$\epsilon \leftarrow \epsilon_{\gamma^*}(\bar{\boldsymbol{X}}(\hat{t}),\hat{t})$
$(\Delta\boldsymbol{X})_\epsilon \leftarrow \left(\mu(\hat{t})(\boldsymbol{X}(1)-\bar{\boldsymbol{X}}(\hat{t})) + \frac{\sigma(\hat{t})^2}{\bar{\sigma}(\hat{t})}\epsilon\right)\hat{t} + \sigma(\hat{t})\hat{\epsilon}$
$\boldsymbol{X} \leftarrow \bar{\boldsymbol{X}}(\hat{t}) - (\Delta\boldsymbol{X})_\epsilon$
**return** $\boldsymbol{X}$

designed for the joint training of $G_\alpha$, $H_\beta$, and $F_\gamma$:

$$\mathcal{L} = \underbrace{\mathcal{L}(\bar{\boldsymbol{X}}(\hat{t}),\boldsymbol{X}(0))}_{\text{Regression Loss}} + \underbrace{\mathcal{L}(\bar{\boldsymbol{X}}(\hat{t}),\boldsymbol{X}(\hat{t}))}_{\text{Alignment Loss}} + \underbrace{\mathcal{L}(\hat{\boldsymbol{X}}(0),\boldsymbol{X}(0))}_{\text{Diffusion Loss}},$$

where $\mathcal{L}(\cdot,\cdot)$ is the $\ell_2$-norm loss, $\bar{\boldsymbol{X}}(\hat{t})$ is the video updated by the Regressor Initializer, $\boldsymbol{X}(\hat{t})$ is the intermediate state by setting $t = \hat{t}$ in Eq. (5), $\hat{\boldsymbol{X}}(0)$ is the result derived from the One-step Denoiser. Regression Loss ensures the coarse prediction result of the Regressor Initializer being closer to the ground truth, providing improved initialization input for the One-step Denoiser. Alignment Loss aligns the updated video within our reverse denoising process, allowing the application of Eq. (6) for denoising. Finally, Diffusion

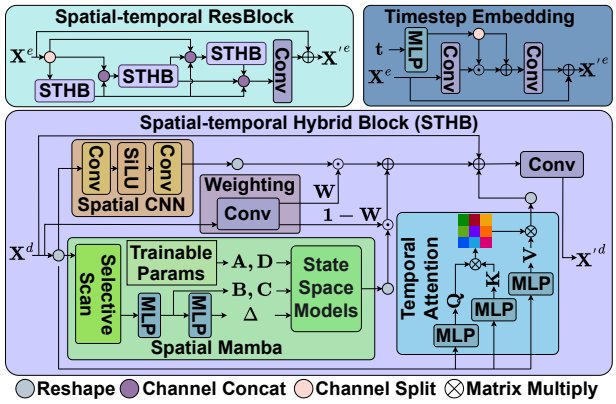


Figure 5: Detailed architectures of the Spatial-temporal ResBlock, the STHB block, and the Timestep Embedding block.

Loss, a special case of trajectory loss in Eq. (7) by setting $(d\boldsymbol{X})^*_{\hat{t}} = \bar{\boldsymbol{X}}(\hat{t}) - \boldsymbol{X}(0)$ and $\hat{\boldsymbol{X}}(0) = \bar{\boldsymbol{X}}(\hat{t}) - (d\boldsymbol{X})_{\epsilon_{\hat{t}}}$, optimizes the parameters of the Noise Predictor in the One-step Denoiser, enabling one-step diffusion from $\bar{\boldsymbol{X}}(\hat{t})$ to $\boldsymbol{X}(0)$. With trained parameters $\alpha^*$, $\beta^*$, and $\gamma^*$, the algorithm $\mathcal{D}^*_{\text{regdif}}$ retrieves the original video, as demonstrated in Algorithm 2.

Furthermore, drawing inspiration from the adapter, we integrate the compensatory measurement from Eq. (3) into our reconstruction framework to enhance reconstruction quality, utilizing the architecture analogous to that of the adapter:

$$\hat{\boldsymbol{X}}^{\text{enhance}}(\hat{t}) = \text{Fusion}_x\left(\hat{\boldsymbol{X}}(\hat{t}), \text{Embedding}_x(\boldsymbol{Y}^C)\right),$$
$$\epsilon^{\text{enhance}}_{\hat{t}} = \text{Fusion}_\epsilon\left(\epsilon_{\hat{t}}, \text{Embedding}_\epsilon(\boldsymbol{Y}^C)\right),$$

where $\hat{\boldsymbol{X}}^{\text{enhance}}(\hat{t})$ and $\epsilon^{\text{enhance}}_{\hat{t}}$ denote the enhanced outputs of the Regressor Initializer and the Noise Predictor, respectively. The modules $\text{Embedding}_x$ and $\text{Embedding}_\epsilon$ serve to obtain the hidden representations of the measurement $\boldsymbol{Y}^C$, which are then fed into the video fusion module $\text{Fusion}_x$ and the noise fusion module $\text{Fusion}_\epsilon$, respectively. Details of the fusion architecture can be found in Appendix C.

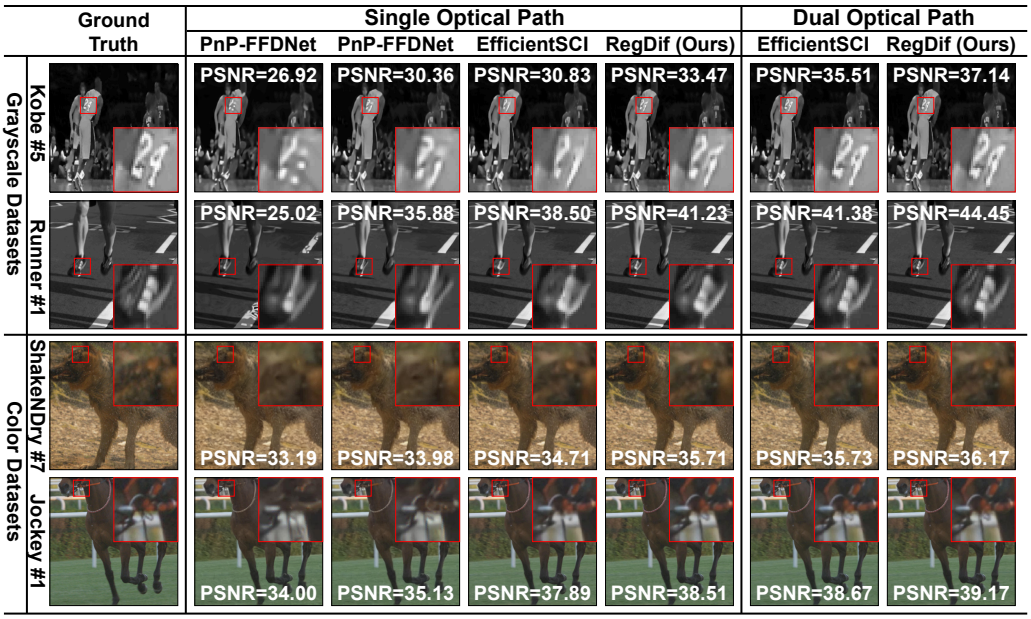


Figure 6: Selected reconstruction frames of simulated grayscale and color datasets. Zoom in for better view.

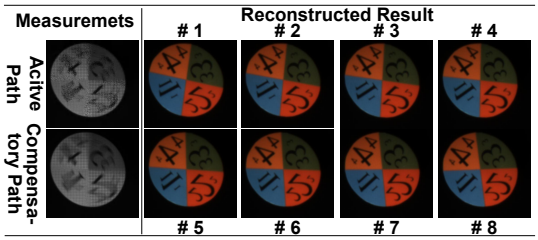


Figure 7: Measurements captured by our dual-path hardware with one-hot mask and reconstructed results of DoP-RegDif.

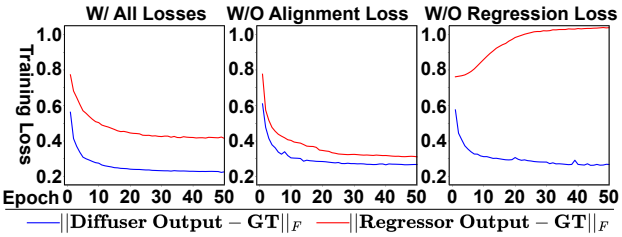


Figure 8: Ablation on the loss functions used in RegDif.

## Experiments

### Datasets and Implementation Details

Following the settings in previous works (Wang, Cao, and Yuan 2023; Cheng et al. 2020), we use DAVIS2017 (Pont-Tuset et al. 2017) with resolution $480 \times 894$ (480P) as the training dataset for the model. In the evaluation stage, we test the RegDif on six simulation grayscale/color datasets with a size of $\{256 \times 256 \times 8, 512 \times 512 \times 3 \times 8\}$. Subsequently, we conduct data acquisition and reconstruction in real-world scenarios using the dual optical path equipped with the one-hot mask. For the reconstruction framework RegDif, as shown in Fig. 4, we set 7 building blocks in the Noiser Predictor and 1 building block in the Regressor Initializer to match 8 building blocks commonly used in previous works (Wang, Cao, and Yuan 2023). We use the PyTorch framework with 4 NVIDIA RTX 3090 GPUs for training with Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.99$) on a $1 \times 10^{-4}$ learning rate over 100 epochs with a decay ratio 0.5 per 25 epochs and then finetune with a $1 \times 10^{-5}$ learning ratio over 20 epochs. The Peak Signal to Noise Ratio (PSNR) and Structural SIMilarity (SSIM) (Wang et al. 2004) are used as the performance indicators of the reconstruction quality. In all experiments, the best and second-best results of the evaluated methods are **highlighted** and underlined.

### Results on Simulation Benchmark Videos

On simulated grayscale datasets, we compare RegDif with model-based methods (GAP-TV (Yuan 2016), PnP-FFDNet (Yuan et al. 2020), PnP-FastDVDnet (Yuan et al. 2021)) and deep learning-based methods (BIRNAT (Cheng et al. 2020), RevSCI (Cheng et al. 2021), STFormer (Wang et al. 2022), EfficientSCI (Wang, Cao, and Yuan 2023)) in the single-path configuration. For dual path, we compare DoP-RegDif (RegDif enhanced by compensatory-path measurement) with DoP-EfficientSCI (EfficientSCI enhanced through the same fusion method as DoP-RegDif). In terms of quantitative comparisons shown in Table 1, our RegDif outperforms previous model-based methods by more than **3.5** dB within the single-path setting. Compared to EfficientSCI, RegDif achieves an improvement of **1.9** dB while maintaining comparable parameters (EfficientSCI: 8 regression blocks; RegDif: 1 regression + 7 diffusion blocks; same 256 hidden dimension with similar architecture). In the dual-path setting, DoP-RegDif yields **2.8** dB over RegDif by in-

| Grayscale Datasets | Kobe | Traffic | Runner | Drop | Crash | Aerial | Average |
|---|---|---|---|---|---|---|---|
| GAP-TV | 22.42, 0.690 | 19.53, 0.627 | 27.44, 0.861 | 31.50, 0.932 | 24.01, 0.836 | 25.18, 0.828 | 25.01, 0.796 |
| PnP-FFDNet | 26.92, 0.912 | 17.01, 0.694 | 25.02, 0.902 | 10.05, 0.341 | 14.91, 0.539 | 21.69, 0.798 | 19.27, 0.698 |
| PnP-FastDVDnet | 30.36, 0.937 | 26.80, 0.920 | 35.88, 0.957 | 42.08, 0.989 | 26.23, 0.906 | 27.25, 0.896 | 31.43, 0.934 |
| BIRNAT | 29.50, 0.900 | 27.92, 0.922 | 36.86, 0.970 | 41.22, 0.989 | 27.80, 0.903 | 28.34, 0.897 | 31.94, 0.930 |
| RevSCI | 28.47, 0.854 | 27.97, 0.921 | 37.04, 0.970 | 41.43, 0.989 | 27.64, 0.896 | 28.69, 0.903 | 31.87, 0.922 |
| STFormer | 30.55, 0.946 | 29.42, 0.952 | 38.76, 0.983 | 42.04, 0.992 | 28.72, 0.950 | 29.79, 0.938 | 33.21, 0.960 |
| EfficientSCI | 30.83, 0.949 | 29.60, 0.955 | 38.50, 0.983 | 42.34, 0.992 | 29.05, 0.955 | 29.53, 0.936 | 33.31, 0.962 |
| RegDif(Ours) | 33.47, 0.971 | 31.32, 0.967 | 41.23, 0.988 | 45.28, 0.995 | 29.58, 0.963 | 30.46, 0.949 | 35.22, 0.972 |
| DoP-EfficientSCI | 35.51, 0.974 | 32.03, 0.973 | 41.38, 0.984 | 45.72, 0.995 | 30.08, 0.978 | 31.76, 0.951 | 36.08, 0.976 |
| DoP-RegDif(Ours) | **37.14**, **0.984** | **33.72**, **0.979** | **44.45**, **0.989** | **47.05**, **0.996** | **31.69**, **0.979** | **34.41**, **0.975** | **38.08**, **0.984** |

Table 1: PSNR (left) and SSIM (right) on six grayscale datasets under single and dual optical path (DoP) with one-hot mask.

| Color Datasets | ShakeNDry | Traffic | Jockey | Beauty | Runner | Bosphorus | Average |
|---|---|---|---|---|---|---|---|
| GAP-TV | 22.58, 0.421 | 19.94, 0.546 | 24.73, 0.521 | 19.75, 0.327 | 28.00, 0.806 | 24.30, 0.501 | 23.22, 0.520 |
| PnP-FFDNet | 33.19, 0.943 | 24.18, 0.856 | 34.00, 0.946 | 35.33, 0.971 | 34.79, 0.951 | 32.93, 0.954 | 32.41, 0.937 |
| PnP-FastDVDnet | 33.98, 0.946 | 26.55, 0.912 | 35.13, 0.952 | 36.05, 0.973 | 37.22, 0.973 | 37.56, 0.978 | 34.41, 0.956 |
| EfficientSCI | 34.71, 0.934 | 28.79, 0.913 | 37.89, 0.946 | 36.84, 0.966 | 41.75, 0.982 | 40.30, 0.967 | 36.71, 0.951 |
| RegDif(Ours) | 35.71, 0.947 | 30.34, 0.937 | 38.51, 0.954 | 37.22, 0.978 | 42.09, 0.985 | 40.48, 0.972 | 37.39, 0.962 |
| DoP-EfficientSCI | 35.73, 0.944 | 30.47, 0.933 | 38.67, 0.954 | 37.43, 0.975 | 42.20, 0.988 | 40.81, 0.971 | 37.55, 0.961 |
| DoP-RegDif(Ours) | **36.17**, **0.954** | **30.86**, **0.949** | **39.17**, **0.969** | **37.73**, **0.986** | **43.74**, **0.989** | **41.74**, **0.979** | **38.24**, **0.971** |

Table 2: PSNR (left) and SSIM (right) on six color datasets under single and dual optical path (DoP) with one-hot mask.

corporating additional measurement and outperforms DoP-EfficientSCI by **2.0** dB. As shown in Fig. 6, our method retrieves more details and textures in grayscale datasets.

On simulated color datasets, reconstruction is more challenging due to spatial masking interacting with the Bayer filter. Beyond changing output channels from 1 to 3, RegDif requires additional masked-region processing during coarse-prediction updates (details are provided in Appendix D). We compare RegDif with GAP-TV, PnP-FFDNet, PnP-FastDvDnet, and EfficientSCI in the single-path configuration and compare DoP-RegDif with DoP-EfficientSCI for dual path. Our RegDif outperforms EfficientSCI by **0.68** dB under the single-path setting, while DoP-RegDif outperforms DoP-EfficientSCI by **0.69** dB under the dual-path setting. As shown in Fig. 6, our method is better at restoring accurate colors and fine structures in color datasets.

## Results on Real Captured Videos

In real-world scenarios, the presence of noise often leads to discrepancies compared to simulations. RegDif, which introduces Gaussian noise in the SDE, provides advantages in reconstructing real scenes. Furthermore, the limited bit depth of the camera causes the restricted dynamic range; in this context, our one-hot mask demonstrates a clear advantage over the random binary mask. We employ a dual optical path with the one-hot mask to capture a dynamic scene featuring a rotating disc displaying four digits, each situated in a region of a different color. After obtaining two complementary measurements, we reconstruct the dynamic scene with DoP-RegDif, and the results are illustrated in Fig. 7.

## Ablation Study

To offer insights into the proposed method, we analyze the impact of loss functions in our framework by recording two metrics during training: the $\ell_2$-norm distance between ground truth and One-step Denoiser output (final output), as well as the distance between ground truth and Regressor Initializer output (coarse output) under three conditions: $(i)$ all three losses are utilized, $(ii)$ the Alignment Loss is omitted, and $(iii)$ the Regression Loss is removed. As shown in Fig. 8, the final output under condition $(i)$ performs the best. When the Alignment Loss is absent (which aligns the coarse output within the diffusion process), the lack of this constraint enhances the coarse output but degrades the final output compared to condition $(i)$. Furthermore, when ablating Regression Loss (which constrains the coarse result), the Alignment Loss may cause deterioration of the coarse output by aligning it within reverse process, leading to a decline in the final output relative to condition $(i)$. Additional details, including ablation studies of model architecture, inference times, and parameter sizes are provided in Appendix E.

# Conclusion

Leveraging the one-hot mask's ability to effectively distinguish information between frames in modulated measurements, we are the first to transform the regression problem in video SCI into a diffusion-based inpainting problem by proposing an SDE of the forward process that aligns with the hardware compression process. To address slow reconstruction speed and poor quality of the pure diffusion in video SCI, we propose a hybrid framework named RegDif, using one-step regression to predict coarse results followed by one-step denoising. Furthermore, we design a hardware-level dual path incorporating an additional measurement to mitigate degradation caused by one-hot masks. Extensive experiments on grayscale, color, and real data demonstrate our method's superiority under one-hot modulation.

## Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2024YFF0505603, National Natural Science Foundation of China under Grant 62271414, Zhejiang Provincial Science Fund for Distinguished Young Scholar Project under Grant LR23F010001, Zhejiang "Pioneer" and "Leading Goose" R&D Program under Grant 2024SDXHDX0006 and 2024C03182, the Key Project of Westlake Institute for Optoelectronics under Grant 2023GD007, and Ningbo Science and Technology Bureau, "Science and Technology Yongjiang 2035" Key Technology Breakthrough Program under Grant 2024Z126.

## References

Adrian, R. J.; and Westerweel, J. 2011. *Particle image velocimetry*. 30. Cambridge university press.

Candès, E. J.; Romberg, J.; and Tao, T. 2006. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2): 489–509.

Ceylan, D.; Huang, C.-H. P.; and Mitra, N. J. 2023. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23206–23217.

Cheng, Z.; Chen, B.; Liu, G.; Zhang, H.; Lu, R.; Wang, Z.; and Yuan, X. 2021. Memory-efficient network for large-scale video compressive sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16246–16255.

Cheng, Z.; Lu, R.; Wang, Z.; Zhang, H.; Chen, B.; Meng, Z.; and Yuan, X. 2020. BIRNAT: Bidirectional recurrent neural networks with adversarial training for video snapshot compressive imaging. In *European Conference on Computer Vision*, 258–275. Springer.

Donoho, D. L. 2006. Compressed sensing. *IEEE Transactions on information theory*, 52(4): 1289–1306.

Gao, C.; Saraf, A.; Huang, J.-B.; and Kopf, J. 2020. Flow-edge guided video completion. In *European Conference on Computer Vision*, 713–729. Springer.

Geyer, M.; Bar-Tal, O.; Bagon, S.; and Dekel, T. 2023. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*.

Hitomi, Y.; Gu, J.; Gupta, M.; Mitsunaga, T.; and Nayar, S. K. 2011. Video from a single coded exposure photograph using a learned over-complete dictionary. In *2011 International Conference on Computer Vision*, 287–294. IEEE.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Li, Z.; Lu, C.-Z.; Qin, J.; Guo, C.-L.; and Cheng, M.-M. 2022. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17562–17571.

Lieberman, D. E.; Venkadesan, M.; Werbel, W. A.; Daoud, A. I.; D'andrea, S.; Davis, I. S.; Mang'Eni, R. O.; and Pitsiladis, Y. 2010. Foot strike patterns and collision forces in habitually barefoot versus shod runners. *Nature*, 463(7280): 531–535.

Liu, D.; Gu, J.; Hitomi, Y.; Gupta, M.; Mitsunaga, T.; and Nayar, S. K. 2013. Efficient space-time sampling with pixel-wise coded exposure for high-speed imaging. *IEEE transactions on pattern analysis and machine intelligence*, 36(2): 248–260.

Liu, R.; Deng, H.; Huang, Y.; Shi, X.; Lu, L.; Sun, W.; Wang, X.; Dai, J.; and Li, H. 2021. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, 14040–14049.

Liu, Y.; Yuan, X.; Suo, J.; Brady, D. J.; and Dai, Q. 2018. Rank minimization for snapshot compressive imaging. *IEEE transactions on pattern analysis and machine intelligence*, 41(12): 2990–3006.

Luo, Z.; Gustafsson, F. K.; Zhao, Z.; Sjölund, J.; and Schön, T. B. 2023. Image restoration with mean-reverting stochastic differential equations. *arXiv preprint arXiv:2301.11699*.

Ma, J.; Liu, X.-Y.; Shou, Z.; and Yuan, X. 2019. Deep tensor admm-net for snapshot compressive imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10223–10232.

Martel, J. N.; Mueller, L. K.; Carey, S. J.; Dudek, P.; and Wetzstein, G. 2020. Neural sensors: Learning pixel exposures for HDR imaging and video compressive sensing with programmable sensors. *IEEE transactions on pattern analysis and machine intelligence*, 42(7): 1642–1653.

Parab, N. D.; Xiong, L.; Guo, Q.; Guo, Z.; Kirk, C.; Nie, Y.; Xiao, X.; Fezzaa, K.; Everheart, W.; Chen, W. W.; et al. 2019. Investigation of dynamic fracture behavior of additively manufactured Al-10Si-Mg using high-speed synchrotron X-ray imaging. *Additive Manufacturing*, 30: 100878.

Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbeláez, P.; Sorkine-Hornung, A.; and Van Gool, L. 2017. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*.

Qi, C.; Cun, X.; Zhang, Y.; Lei, C.; Wang, X.; Shan, Y.; and Chen, Q. 2023. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15932–15942.

Qiao, M.; Liu, X.; and Yuan, X. 2020. Snapshot spatial–temporal compressive imaging. *Optics letters*, 45(7): 1659–1662.

Qiao, M.; Meng, Z.; Ma, J.; and Yuan, X. 2020. Deep learning for video compressive sensing. *Apl Photonics*, 5(3).

Qiao, M.; and Yuan, X. 2022. Coded aperture compressive temporal imaging using complementary codes and untrained neural networks for high-quality reconstruction. *Optics Letters*, 48(1): 109–112.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Wang, G.; Liu, X.; and Yuan, X. 2025. Dual-optical-path coded aperture compressive temporal imaging. *Optics Letters*, 50(6): 1865–1868.

Wang, L.; Cao, M.; and Yuan, X. 2023. Efficientsci: Densely connected network with space-time factorization for large-scale video snapshot compressive imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18477–18486.

Wang, L.; Cao, M.; Zhong, Y.; and Yuan, X. 2022. Spatial-temporal transformer for video snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7): 9072–9089.

Wang, P.; Wang, L.; and Yuan, X. 2023. Deep optics for video snapshot compressive imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10646–10656.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wang, Z.; Zhang, H.; Cheng, Z.; Chen, B.; and Yuan, X. 2021. Metasci: Scalable and adaptive reconstruction for video compressive sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2083–2092.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7623–7633.

Yang, J.; Liao, X.; Yuan, X.; Llull, P.; Brady, D. J.; Sapiro, G.; and Carin, L. 2014. Compressive sensing by learning a Gaussian mixture model from measurements. *IEEE Transactions on Image Processing*, 24(1): 106–119.

Yuan, X. 2016. Generalized alternating projection based total variation minimization for compressive sensing. In *2016 IEEE International conference on image processing (ICIP)*, 2539–2543. IEEE.

Yuan, X.; Brady, D. J.; and Katsaggelos, A. K. 2021. Snapshot compressive imaging: Theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 38(2): 65–88.

Yuan, X.; Liu, Y.; Suo, J.; and Dai, Q. 2020. Plug-and-play algorithms for large-scale snapshot compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1447–1457.

Yuan, X.; Liu, Y.; Suo, J.; Durand, F.; and Dai, Q. 2021. Plug-and-play algorithms for video snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 7093–7111.

Zhang, K.; Fu, J.; and Liu, D. 2022. Flow-guided transformer for video inpainting. In *European conference on computer vision*, 74–90. Springer.

Zhang, Z.; Wu, B.; Wang, X.; Luo, Y.; Zhang, L.; Zhao, Y.; Vajda, P.; Metaxas, D.; and Yu, L. 2024. Avid: Any-length video inpainting with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7162–7172.

Zhou, S.; Li, C.; Chan, K. C.; and Loy, C. C. 2023. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10477–10486.