Snapshot Compressive Imaging Using Domain-Factorized Deep Video Prior

Yu-Chun Miao, Xi-Le Zhao*, Jian-Li Wang, Xiao Fu, Yao Wang

***Abstract*—Snapshot compressive imaging (SCI) aims at efficiently capturing high-dimensional data (e.g., multi-spectral images and videos) using a two-dimensional detector, which is a hardware-friendly data acquisition paradigm. However, because of the complex structure of videos (such as the dynamic background and moving foreground), it is challenging to reconstruct a video from the captured measurement. Existing model-based methods for video SCI reconstruction are inadequate to reconstruct the complex structure of videos, and existing supervised deep learning-based methods are with poor adaptability to videos in real scenarios. Inspired by the physically interpreted video decomposition, we suggest an unsupervised video SCI reconstruction method with tailored deep video prior and affine transformation, namely, `FactorDVP-T`. Our `FactorDVP-T` infers the parameters of the neural networks and the underlying structure of the original video from the captured measurement using a non-reference loss function in an unsupervised manner. Under `FactorDVP-T`, a video is first factorized into the moving foreground and static background. The background is further factorized into temporal bases and spatial coefficients, where each factor can be modeled individually using the designated unsupervised networks in `FactorDVP-T`. Moreover, to tackle the dynamic background in real scenarios, we integrate the affine transformation into `FactorDVP-T`. Benefiting from the expressive power of unsupervised networks embedded in the physically interpreted video decomposition framework, our methods can reconstruct the videos more effectively and better adapt to various videos in real scenarios, as compared with the model-based methods and supervised deep learning-based methods respectively. Extensive experiments on various videos show that our `FactorDVP-T` can better adapt to different videos, compared with the state-of-the-art model-based and supervised deep learning-based SCI reconstruction methods.**

***Index Terms*—Affine transformation, deep video prior (DVP), physically interpreted video decomposition, video compressive sensing.**

I. Introduction

Snapshot compressive imaging (SCI) [1], [2] offers an important means for high-quality data (e.g., multi-spectral images (MSIs) and videos) acquisition and transmission/broadcasting under resource (e.g., storage and streaming bandwidth) limitations. In a nutshell, SCI systems first compress multiple video frames/MSI bands into one frame/band with the assistance of a sensing mask and then recover the original data when needed. Such compression may substantially reduce the memory complexity and communication overhead when storing or transmitting high-quality data. In [3], [1], it was shown that accurate recovery for such compression is possible.

Reconstructing the original data from a single measurement is an ill-posed inverse problem. In essence, this problem is a special compressive sensing (CS) problem [4] that uses a particular hardware-friendly compressing strategy. Like other CS-type problems, e.g., sparse signal reconstruction [5], [6], matrix sensing/completion [7], [8], and tensor recovery [9], [10], one of the key factors leading to success is how to deeply explore the prior knowledge of the original data. Early SCI reconstruction methods often employ hand-crafted priors (e.g., sparsity, low-rankness, and smoothness), such as those used in `GMM-TP` [11], `MMLE-GMM` [12], `GAP-TV` [13] and `DeSCI` [14]. These methods are effective in a certain extent, but the hand-crafted priors oftentimes are inadequate to handle the real-world data.

Recently, triggered by the expressing power of deep neural networks that is far beyond the capability of hand-crafted priors [15], [16], [17], [18], [19], [20], [21], [22], a plethora of supervised deep learning-based methods were developed. These methods often use training data to learn a neural network that can directly recover the original data from a single measurement. A branch of these methods mainly concentrates on exploring different neural network architectures to learn a mapping from a single measurement to the reconstructed video, such as fully connection network (FCN) [23], convolution neural network (CNN) [24], [25], [26], recurrent neural network (RNN) [27], and graph neural network (GNN) [28]. Another branch of supervised deep learning-based methods engages in constructing neural networks by unfolding the iterative optimization algorithms, such as those in [29], [30], [31], [32], [33], [34]. It is worth noting that for such supervised deep learning-based methods when a network is trained for a specific SCI system, it cannot be used in other SCI systems that provide different modulation patterns or different compression rates. As a result, a number of methods are proposed by incorporating a pre-trained deep denoising network such as FFDNet [35] into the plug-and-play (PnP) framework, such as those in [36], [37]. Although these supervised deep learning-based and PnP-based methods obtain high-quality recovered

*Corresponding authors. Tel.: +86 28 61831016, Fax: +86 28 61831280.

the Key Project of Applied Basic Research in Sichuan Province (No. 2020YJ0216), the Applied Basic Research Project of Sichuan Province (No. 2021YJ0107) and National Key Research and Development Program of China (No. 2020YFA0714001). The work of X. Fu is supported by NSF ECCS-2024058 and NSF ECCS-1808159.

Y.-C. Miao, X.-L. Zhao, and J.-L. Wang are with the Research Center for Image and Vision Computing, School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu 611731, P.R.China (e-mails: szmyc1@163.com; xlzhao122003@163.com; wangjianli_123@163.com).

X. Fu is with the School of Electrical Engineering and Computer Science, Oregon State University (OSU), Corvallis, OR 97331, United States (e-mail: xiao.fu@oregonstate.edu).

Y. Wang is with the Institute for Information and System Sciences, School of Mathematics and Statistics, Xian Jiaotong University, Xian 710049, China (e-mail: yao.s.wang@gmail.com).

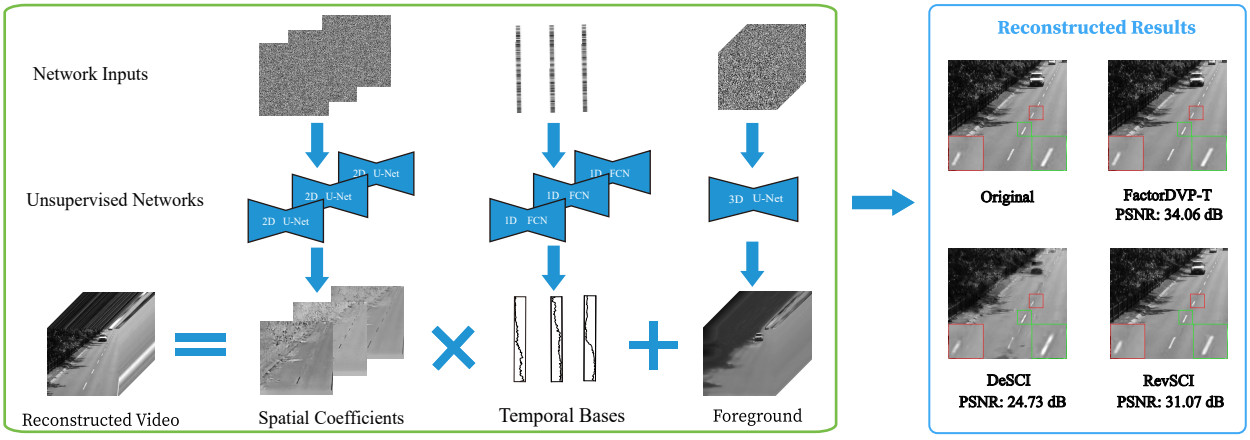

Fig. 1. (a) Illustration of the proposed domain-factorized deep video prior. (b): The original frame, the reconstructed frame by the model-based method (`DeSCI`), the supervised deep learning-based method (`RevSCI`), and our unsupervised method `FactorDVP-T`.

data utilizing the data-driven priors, their limitation is widely recognized:

Poor adaptability to videos in real scenarios. In real scenarios, high-dimensional data is expensive to acquire and always with complex structure, especially for videos with moving foreground and dynamic background. For the data-driven priors utilized in the supervised deep learning-based and PnP-based methods, it is hard to collect training data containing all the distributions in real scenarios, and the reconstruction performance would significantly degrade when the test data is out of the training data's distribution. Thus, poor adaptability to videos in real scenarios is an important limitation for the data-driven priors in supervised deep learning-based and PnP-based methods.

Bearing these concerns in mind, in this work, we put forth an unsupervised video SCI reconstruction method with tailored deep video prior (DVP), which can cleverly address the fore-mentioned limitation by virtue of the expressive power of unsupervised networks embedded in the physically interpreted video decomposition framework. For the videos with static background, we take wisdom from classical video modeling [38] to factorize a video into a static background and a moving foreground. The background is low-rank, and could be further factorized into several spatial coefficients and temporal bases, which can be modeled using 2D U-Net [39] and FCN, respectively. The foreground can be modeled by 3D convolution-based U-Net (3D U-Net) [39]. Moreover, to tackle the dynamic background in real scenarios, we introduce the affine transformation for dynamic background modeling. Our method is termed as `FactorDVP-T`. Our contributions are mainly two folds:

- Inspired by physically interpreted video decomposition, we designated unsupervised networks capturing the latent factors and organically reconciled to deliver promising performance; see Fig. 1 for illustration. The suggested methods which allow us to unsupervisedly reconstruct the video only from the captured measurement itself without any training data, can flexibly adapt to various videos in real scenarios.
- Experiments on a wide range of videos (including 14 grayscale videos with static background, 3 color videos with dynamic background, and 2 real videos) verify that the proposed `FactorDVP-T` can more flexibly adapt to various videos in real scenarios, as compared with the state-of-the-art model-based and supervised deep learning-based methods, see Section V.

The rest of this paper is organized as follows. Section II briefly introduces some pertinent background information. Section III introduces some related works. Section IV elaborates more details about the proposed method. Section V provides the experiment results. Section VI reports the ablation study. Section VII concludes this paper.

II. Notation and Preliminary

In this section, we first take a brief introduction of the basic notations applied in this paper. Then, we introduce the mathematical model of video SCI system.

A. Notation

A scalar, vector, matrix, and tensor are denoted as x, $\boldsymbol{x}$, $\boldsymbol{X}$, and $\underline{\boldsymbol{X}}$, respectively. $[\boldsymbol{x}]_i$, $[\boldsymbol{X}]_{i,j}$, and $[\underline{\boldsymbol{X}}]_{i,j,k}$ denote the i-th, (i,j)-th, and (i,j,k)-th element of $\boldsymbol{x} \in \mathbb{R}^{n_1}$, $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$, and $\underline{\boldsymbol{X}} \in \mathbb{R}^{n_1 \times n_2 \times b}$, respectively. $\underline{\boldsymbol{X}}^{(i)} \in \mathbb{R}^{n_1 \times n_2}$ denotes the i-th frontal slices. The Frobenius norms of $\underline{\boldsymbol{X}}$ is denoted as $\|\underline{\boldsymbol{X}}\|_F = \sqrt{\sum_{i,j,k}[\underline{\boldsymbol{X}}]^2_{i,j,k}}$. The Hadamard product of $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$ and $\boldsymbol{Y} \in \mathbb{R}^{n_1 \times n_2}$ are denoted as $[\boldsymbol{X} \odot \boldsymbol{Y}]_{i,j} = [\boldsymbol{X}]_{i,j}[\boldsymbol{Y}]_{i,j}$. And $\boldsymbol{X} \circ \boldsymbol{y} \in \mathbb{R}^{n_1 \times n_2 \times b}$ is the outer product of $\boldsymbol{y} \in \mathbb{R}^b$ and $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$, which is defined as $[\boldsymbol{X} \circ \boldsymbol{y}]_{i,j,k} = [\boldsymbol{X}]_{i,j}[\boldsymbol{y}]_k$. The $\mathrm{vec}(\boldsymbol{X})$ operator represents $\mathrm{vec}(\boldsymbol{X}) = [[\boldsymbol{X}]^T_{:,1}, \ldots, [\boldsymbol{X}]^T_{:,n_2}]^T$.

B. Degradation Model of Video SCI System

As an important branch with broad application prospects of CS [40], video SCI aims to capture high speed videos with low hardware requirement. Unlike conventional high-speed cameras [41] that put high requirement on the quality of imaging

devices and inevitably lead to high hardware costs, video SCI systems consistently exploit the coded aperture compressive temporal imaging (CACTI) [1]. The image compression procedure of CACTI system can be logically abstracted into two stages: sampling stage and integration stage. In the sampling stage, it samples a set of consecutive frames of the input video stream with the assistance of a random binary mask In the integration stage, the sampled frames are integrated into a single measurement along the temporal dimension, and two-dimensional sensors are utilized to capture this measurement. With this compression procedure, the memory complexity, bandwidth, and communication overhead can be substantially reduced, therefore video SCI systems place low requirement on the hardware quality.

Mathematically, given a video with B frames, each of which contains $n\ (=n_1\times n_2)$ pixels, the imaging model of SCI can be formulated as follows:

$$\boldsymbol{y}=\boldsymbol{\Phi}\boldsymbol{x}+\boldsymbol{n}, \tag{1}$$

where $\boldsymbol{\Phi}\in\mathbb{R}^{n\times nB}$ represents the sensing mask, $\boldsymbol{x}\in\mathbb{R}^{nB}$ is the desired video, $\boldsymbol{y}\in\mathbb{R}^{n}$ is the captured measurement, and $\boldsymbol{n}$ denotes the noise.

Eqn. (1) is consistent with the formulation of CS [40]. Different from traditional CS, the sensing mask considered here has very specific structure:

$$\boldsymbol{\Phi}=\left[\underline{\boldsymbol{D}}^{(1)},\ldots,\underline{\boldsymbol{D}}^{(B)}\right], \tag{2}$$

where $\{\underline{\boldsymbol{D}}^{(k)}\}_{k=1}^{B}$ are diagonal matrices. Specifically, taking video $\underline{\boldsymbol{X}}\in\mathbb{R}^{n_1\times n_2\times b}$ as an example, it is modulated and compressed by shifted random binary sensing mask $\underline{\boldsymbol{M}}\in\mathbb{R}^{n_1\times n_2\times b}$. The measurement $\boldsymbol{Y}\in\mathbb{R}^{n_1\times n_2}$ is given by

$$\boldsymbol{Y}=\sum_{k=1}^{b}\underline{\boldsymbol{X}}^{(k)}\odot\underline{\boldsymbol{M}}^{(k)}+\boldsymbol{N}, \tag{3}$$

where $\odot$ denotes the Hadamard product, and $\underline{\boldsymbol{N}}$ represents the noise.

Notably, by defining $\underline{\boldsymbol{D}}^{(k)}=\mathrm{diag}(\mathrm{vec}(\underline{\boldsymbol{M}}^{(k)}))\in\mathbb{R}^{n_1n_2\times n_1n_2}$, for $k=1,...,B$, $\boldsymbol{y}=\mathrm{vec}(\boldsymbol{Y})\in\mathbb{R}^{n_1n_2}$, $\boldsymbol{n}=\mathrm{vec}(\boldsymbol{N})\in\mathbb{R}^{n_1n_2}$, and $\boldsymbol{x}=[\mathrm{vec}(\underline{\boldsymbol{X}}^{(1)})^T,...,\mathrm{vec}(\underline{\boldsymbol{X}}^{(B)})^T]^T\in\mathbb{R}^{n_1n_2B}$, we have the vector formulation of Eqn. (3), i.e., Eqn. (1).

III. Related Work

In order to highlight the relationship between the proposed methods and previous methods, in this section, we introduce two families of methods separately, i.e., deep image prior (DIP)-based methods and unsupervised SCI reconstruction methods.

A. DIP-Based Methods

DIP was first proposed in [42], which is an unsupervised image restoration framework that aims at alleviating the data adaptability problems of supervised deep learning-based methods. The main idea of this work is that a neural network with an appropriate structure can encode much critical prior information of nature images in an unsupervised manner. Under DIP framework, the minimization problem is formulated as:

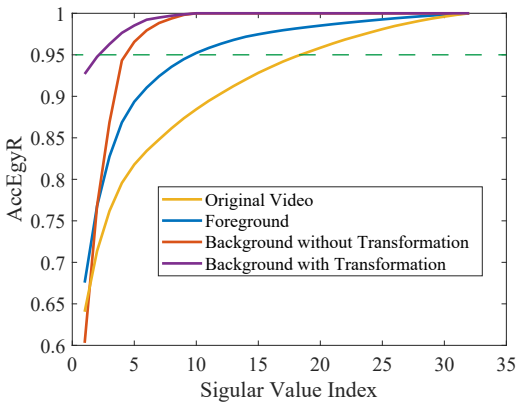

Fig. 2. Comparison of low-rankness between the foreground, background without transformation, background with transformation, and original video consisting of foreground and background by computing the accumulation energy ratios (AccEgyR). Here, the AccEgyR is defined as: AccEgyR$=\sum_i\sigma_i$, where σ_i is the i-th singular value.

$$\min_{\boldsymbol{\theta}} E\left(\mathcal{G}_{\boldsymbol{\theta}}(\boldsymbol{z});\ \boldsymbol{Y}\right), \tag{4}$$

where $\mathcal{G}_{\boldsymbol{\theta}}(\cdot):\mathbb{R}^{n_1\times n_2\times n_3}\rightarrow\mathbb{R}^{n_1\times n_2\times n_3}$ is the network capturing the unsupervised deep prior of the desired data, $\boldsymbol{\theta}$ denotes the network parameters, and $\boldsymbol{z}$ is the random but fixed input of $\mathcal{G}_{\boldsymbol{\theta}}(\cdot)$. The reconstructed data can be estimated as

$$\widehat{\boldsymbol{X}}=\mathcal{G}_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{z}), \tag{5}$$

where $\widehat{\boldsymbol{X}}$ is the reconstructed data, and $\widehat{\boldsymbol{\theta}}$ is the estimated network parameters.

Due to its unsupervised nature, DIP has received much attention in a wide range of application domains, such as image super-resolution [43], semantic photo manipulation [44], video motion transfer [45], and hyperspectral imaging [39]. Recently, Gandelsma et al. [46] showed that when utilizing multiple DIPs to jointly reconstruct a single observed image, these DIPs tend to split the image into different layers with simple patch distribution. Motivated by these methods, Miao et al [47] design two types of DIPs, i.e, deep spatial prior and deep spectral prior, to model the abundance maps and end-members contained in the HSIs, based on the classical spatio-spectral decomposition. Although such DIP-based methods have achieved promising results on RGB images and HSIs, it is hard for them to reconstruct the videos accurately. In this paper, we will explore how to extend DIP to video for video SCI.

B. Unsupervised SCI Reconstruction Methods

To the best of our knowledge, there is only one unsupervised SCI reconstruction method, i.e., `PnP-DIP` [48]. In `PnP-DIP`, DIP is integrated into the PnP regime, leading to exciting results on MSIs. However, the poor reconstructed quality of this method on complex videos prevents it from wide applications. The reason is that, as compared with MSIs,

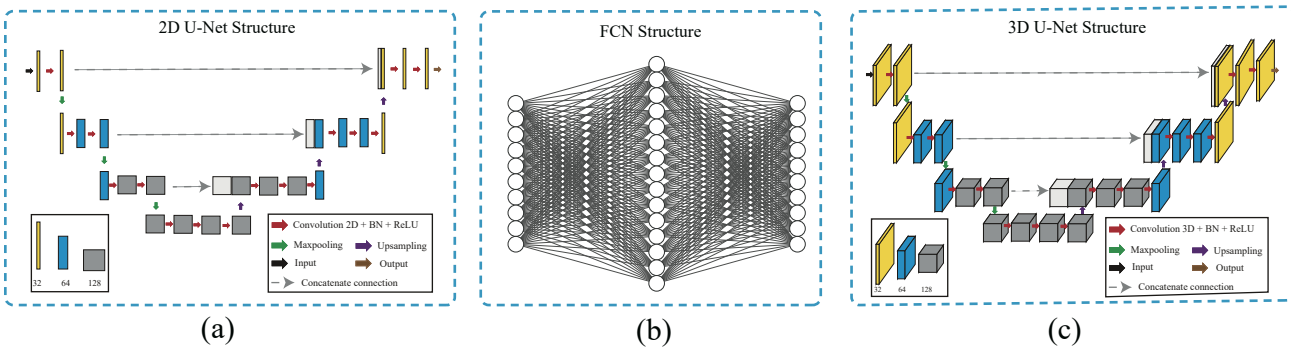

Fig. 3. (a) The network structure of 2D U-Net. (b) The network structure of FCN. (c) The network structure of 3D U-Net.

videos are always with more complex structures, e.g., dynamic background and moving foreground, which preclude the direct exploration of the underlying structure of videos. In this paper, we will explore how to leverage the underlying structure of a video.

IV. Proposed Method

A video could be represented as a third-order tensor $\underline{\boldsymbol{X}} \in \mathbb{R}^{n_1 \times n_2 \times b}$, which could be factorized into a static background $\underline{\boldsymbol{B}} \in \mathbb{R}^{n_1 \times n_2 \times b}$ and a moving foreground $\underline{\boldsymbol{F}} \in \mathbb{R}^{n_1 \times n_2 \times b}$, i.e.,

$$\underline{\boldsymbol{X}} = \underline{\boldsymbol{B}} + \underline{\boldsymbol{F}}. \tag{6}$$

Note that the background $\underline{\boldsymbol{B}}$ lies in a subspace of low dimension along the temporal dimension; see Fig. 2. The background $\underline{\boldsymbol{B}}$ can be further factorized into temporal bases and spatial coefficients [49], [50], [51]. Thus, video $\underline{\boldsymbol{X}}$ can be represented as

$$\underline{\boldsymbol{X}} = \sum_{i=1}^{r} \boldsymbol{S}_i \circ \boldsymbol{c}_i + \underline{\boldsymbol{F}}, \tag{7}$$

where $\circ$ denotes the outer product, $\boldsymbol{S}_i \in \mathbb{R}^{n_1 \times n_2}$ denotes the i-th spatial coefficient, $\boldsymbol{c}_i \in \mathbb{R}^{b}$ denotes the i-th temporal base, and r denotes the number of temporal bases.

A. Integrating Domain-Factorized DVP to SCI

In this subsection, we design a domain-factorized DVP by designating unsupervised networks to capture the corresponding latent factors; see Fig. 1 for illustration. The motivations of our designs are as follows:

For the background, inspired by [47], the physically interpretation of the latent factors (i.e., $\{\boldsymbol{S}_i\}_{i=1}^r$ and $\{\boldsymbol{c}_i\}_{i=1}^r$) makes it possible to utilize some neural network structures to model these factors. According to [52], the spatial coefficients reveal similar qualities of the natural images. Hence, it is reasonable to utilize 2D convolutional neural networks (2D CNN) designed for nature images to capture the unsupervised deep prior of the spatial coefficient $\boldsymbol{S}_i$. Moreover, the i-th temporal base $\boldsymbol{c}_i$ can be considered as a relatively simple 1D continuous smooth signal. Therefore, $\boldsymbol{c}_i$ can be approximated by FCN accurately.

For the foreground, in most cases, it contains much complex information—which is difficult to be modeled in real scenarios. Since 3D convolutional neural networks (3D CNN) can capture motion information encoded in multiple adjacent frames in videos and show promising results on video action recognition [53], it is reasonable to employ 3D CNN to model the unsupervised deep prior of the foreground.

Following the above perspectives, we model the video as follows:

$$\underline{\boldsymbol{X}} = \sum_{i=1}^{r} \mathcal{S}_{\boldsymbol{\theta}_i}(\boldsymbol{z}_i) \circ \mathcal{C}_{\boldsymbol{\zeta}_i}(\boldsymbol{w}_i) + \mathcal{F}_{\boldsymbol{\varsigma}}(\boldsymbol{u}), \tag{8}$$

where $\mathcal{S}_{\boldsymbol{\theta}_i}(\cdot) : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$, $\mathcal{C}_{\boldsymbol{\zeta}_i}(\cdot) : \mathbb{R}^{b} \to \mathbb{R}^{b}$, and $\mathcal{F}_{\boldsymbol{\varsigma}}(\cdot) : \mathbb{R}^{n_1 \times n_2 \times b} \to \mathbb{R}^{n_1 \times n_2 \times b}$ are the neural networks capturing the unsupervised deep prior of the i-th spatial coefficient, the i-th temporal base, and the foreground, respectively; $\boldsymbol{\theta}_i$, $\boldsymbol{\zeta}_i$, and $\boldsymbol{\varsigma}$ collect the corresponding network parameters, respectively; $\boldsymbol{z}_i$, $\boldsymbol{w}_i$, and $\boldsymbol{u}$ are random but fixed input of the neural networks responsible for generating the i-th spatial coefficient, the i-th temporal base, and the foreground, respectively.

Our detailed designs for $\mathcal{S}_{\boldsymbol{\theta}_i}$, $\mathcal{C}_{\boldsymbol{\zeta}_i}$, and $\mathcal{F}_{\boldsymbol{\varsigma}}$ are as follows:

Unsupervised Deep Background Prior. The background is factorized into spatial coefficients and temporal bases. As mentioned, the spatial coefficients reveal similar qualities to natural images and focus on conveying spatial information. In this work, we utilize the 2D U-Net architecture in [39] for modeling the spatial coefficients. Moreover, the temporal bases are relatively easy to be modeled, as they can be regarded as 1D smooth signals. From this perspective, we utilize FCN architecture to capture the unsupervised deep prior of the temporal bases. The corresponding detailed structures are shown in Fig. 3 (a) and (b).

Unsupervised Deep Foreground Prior. As mentioned, the 3D CNN is with a powerful capacity to capture motion information encoded in multiple adjacent frames in videos. Because of this, we employ the 3D U-Net architecture in [39] for modeling the foreground $\underline{\boldsymbol{F}}$. The corresponding detailed structure is shown in Fig. 3 (c).

Based on the above-mentioned unsupervised deep prior designs, the proposed video SCI reconstruction model could be formulated as:

$$\min_{\{\boldsymbol{\theta}_i, \boldsymbol{\zeta}_i\}_{i=1}^r, \boldsymbol{\varsigma}} \left\| \boldsymbol{Y} - \sum_{k=1}^{b} \underline{\boldsymbol{X}}^{(k)} \odot \underline{\boldsymbol{M}}^{(k)} \right\|_F^2, \tag{9}$$

where

$$\underline{\boldsymbol{X}} = \sum_{i=1}^{r} \mathcal{S}_{\boldsymbol{\theta}_i}(\boldsymbol{z}_i) \circ \mathcal{C}_{\boldsymbol{\zeta}_i}(\boldsymbol{w}_i) + \mathcal{F}_{\boldsymbol{\varsigma}}(\boldsymbol{u}).$$

B. Domain-Factorized DVP with Affine Transformation

Due to the moving cameras and the changing circumstance, the background sometimes is dynamic, which fails to obey the low-rank assumption. Thus, how to exploit the underlying structure is a nontrivial problem. Addressing this problem, hereby, we integrate affine transformation to the aforementioned domain-factorized DVP, namely, `FactorDVP-T`.

From Fig. 2, we can observe that the dynamic background $\underline{\boldsymbol{B}} \in \mathbb{R}^{n_1 \times n_2 \times b}$ is implicitly low-rank. In other words, there exists an explicitly low-rank tensor $\underline{\boldsymbol{B}}_L \in \mathbb{R}^{n_1 \times n_2 \times b}$ and an affine transformation $\mathcal{T}_{\boldsymbol{\gamma}}$ parameterized by $\boldsymbol{\gamma}$, satisfying $\underline{\boldsymbol{B}} = \mathcal{T}_{\boldsymbol{\gamma}}(\underline{\boldsymbol{B}}_L)$. Here, $\mathcal{T}_{\boldsymbol{\gamma}} : \mathbb{R}^{n_1 \times n_2 \times b} \to \mathbb{R}^{n_1 \times n_2 \times b}$ is defined as: $[\underline{\boldsymbol{B}}]_{ijk} = [\mathcal{T}_{\boldsymbol{\gamma}}(\underline{\boldsymbol{B}}_L)]_{ijk} = I(\underline{\boldsymbol{B}}_L^{(k)}, (\hat{i}, \hat{j}))$, where $\underline{\boldsymbol{\gamma}} \in \mathbb{R}^{2\times 3\times b}$ is the learnable parameters,

$$\begin{bmatrix} \hat{i} \\ \hat{j} \end{bmatrix} = \begin{bmatrix} [\underline{\boldsymbol{\gamma}}]_{11k} & [\underline{\boldsymbol{\gamma}}]_{12k} & [\underline{\boldsymbol{\gamma}}]_{13k} \\ [\underline{\boldsymbol{\gamma}}]_{21k} & [\underline{\boldsymbol{\gamma}}]_{22k} & [\underline{\boldsymbol{\gamma}}]_{23k} \end{bmatrix} \begin{bmatrix} i \\ j \\ 1 \end{bmatrix},$$

$I(\cdot)$ is the bilinear interpolation function, and $I(\underline{\boldsymbol{B}}_L^{(k)}, (\hat{i}, \hat{j}))$ returns the interpolation result of the matrix $\underline{\boldsymbol{B}}_L^{(k)}$ at the coordinate $(\hat{i}, \hat{j})$. Notably, we perform the affine transformation on each frame of the generated background separately. That is to say, the parameters of affine transformation on each frame are independent of each other.

Integrating affine transformation into the aforementioned domain-factorized DVP, the video SCI reconstruction model could be reformulated as:

$$\min_{\{\boldsymbol{\theta}_i, \boldsymbol{\zeta}_i\}_{i=1}^r, \boldsymbol{\varsigma}, \boldsymbol{\gamma}} \left\| \boldsymbol{Y} - \sum_{k=1}^{b} \underline{\boldsymbol{X}}^{(k)} \odot \underline{\boldsymbol{M}}^{(k)} \right\|_F^2, \tag{10}$$

where

$$\underline{\boldsymbol{X}} = \mathcal{T}_{\boldsymbol{\gamma}}(\sum_{i=1}^{r} \mathcal{S}_{\boldsymbol{\theta}_i}(\boldsymbol{z}_i) \circ \mathcal{C}_{\boldsymbol{\zeta}_i}(\boldsymbol{w}_i)) + \mathcal{F}_{\boldsymbol{\varsigma}}(\boldsymbol{u}).$$

To simplify the representation, we denote the objective function in Eqn. (10) is denoted as:

$$\min_{\{\boldsymbol{\theta}_i, \boldsymbol{\zeta}_i\}_{i=1}^r, \boldsymbol{\varsigma}, \boldsymbol{\gamma}} \ \mathsf{Loss}\left(\{\boldsymbol{\theta}_i, \boldsymbol{\zeta}_i\}_{i=1}^r, \boldsymbol{\varsigma}, \boldsymbol{\gamma}\right). \tag{11}$$

It is worth noting that Eqn. (11) is differentiable w.r.t. γ, as affine transformation and bilinear interpolation are differentiable w.r.t. $\boldsymbol{\gamma}$, respectively [54]. Any off-the-shelf neural network optimizer can be considered to solve the problem w.r.t. $\{\boldsymbol{\theta}_i, \boldsymbol{\zeta}_i\}_{i=1}^r, \boldsymbol{\varsigma}, \boldsymbol{\gamma}$ in Eqn. (11), since this problem is essentially a regression problem with neural models. Here, we consider the adaptive moment estimation (Adam) algorithm that is empirically validated for such complex network learning problems [55]. Letting t denote the iteration number, the solution of the optimization problem in Eqn. (11) is as follows:

$$\boldsymbol{\theta}_i^{t+1} \leftarrow \boldsymbol{\theta}_i^t - \alpha^t \nabla_{\boldsymbol{\theta}_i} \ \mathsf{Loss}\left(\{\boldsymbol{\theta}_i, \boldsymbol{\zeta}_i\}_{i=1}^r, \boldsymbol{\varsigma}, \boldsymbol{\gamma}\right) \tag{12a}$$

$$\boldsymbol{\zeta}_i^{t+1} \leftarrow \boldsymbol{\zeta}_i^t - \alpha^t \nabla_{\boldsymbol{\zeta}_i} \ \mathsf{Loss}\left(\{\boldsymbol{\theta}_i, \boldsymbol{\zeta}_i\}_{i=1}^r, \boldsymbol{\varsigma}, \boldsymbol{\gamma}\right) \tag{12b}$$

$$\boldsymbol{\varsigma}^{t+1} \leftarrow \boldsymbol{\varsigma}^t - \alpha^t \nabla_{\boldsymbol{\varsigma}} \ \mathsf{Loss}\left(\{\boldsymbol{\theta}_i, \boldsymbol{\zeta}_i\}_{i=1}^r, \boldsymbol{\varsigma}, \boldsymbol{\gamma}\right) \tag{12c}$$

$$\boldsymbol{\gamma}^{t+1} \leftarrow \boldsymbol{\gamma}^t - \alpha^t \nabla_{\boldsymbol{\gamma}} \ \mathsf{Loss}\left(\{\boldsymbol{\theta}_i, \boldsymbol{\zeta}_i\}_{i=1}^r, \boldsymbol{\varsigma}, \boldsymbol{\gamma}\right) \tag{12d}$$

for $i = 1, \ldots, r$ and α^t is the step size of iteration t. Notably, the activation functions in networks $\mathcal{S}_{\boldsymbol{\theta}_i}$, $\mathcal{C}_{\boldsymbol{\zeta}_i}$, and $\mathcal{F}_{\boldsymbol{\varsigma}}$ are not differentiable at one point, thus sub-gradient is utilized in this algorithm denoted by ∇. Moreover, the gradient w.r.t. $\boldsymbol{\theta}_i$, $\boldsymbol{\zeta}_i$, $\boldsymbol{\varsigma}$, and $\boldsymbol{\gamma}$ can be computed by the standard back-propagation algorithm [56]. The algorithm is summarized in Algorithm 1, which is dubbed *domain-factorized deep video prior with affine transformation* (`FactorDVP-T`).

Algorithm 1 `FactorDVP-T` for video SCI reconstruction.

Input: the measurement $\underline{\boldsymbol{Y}} \in \mathbb{R}^{n_1 \times n_2}$, the number of temporal bases r, and the max iteration T.

1: sample random $\boldsymbol{z}_i$, $\boldsymbol{w}_i$, and $\boldsymbol{u}$ from uniform distribution.
2: **for** $t = 1$ to T **do**
3: $\quad \widehat{\boldsymbol{S}}_i = \mathcal{S}_{\boldsymbol{\theta}_i^{t-1}}(\boldsymbol{z}_i)$, $\widehat{\boldsymbol{c}}_i = \mathcal{C}_{\boldsymbol{\zeta}_i^{t-1}}(\boldsymbol{w}_i)$, $\widehat{\underline{\boldsymbol{F}}} = \mathcal{F}_{\boldsymbol{\varsigma}^{t-1}}(\boldsymbol{u})$;
4: $\quad \widehat{\mathcal{T}}_{\boldsymbol{\gamma}} = \mathcal{T}_{\boldsymbol{\gamma}^{t-1}}$;
5: $\quad$ Update $\{\boldsymbol{\theta}_i\}_{i=1}^r$, $\{\boldsymbol{\zeta}_i\}_{i=1}^r$, $\boldsymbol{\varsigma}$, and $\boldsymbol{\gamma}$ using Adam [55];
6: **end for**
7: $\widehat{\underline{\boldsymbol{X}}} = \widehat{\mathcal{T}}_{\boldsymbol{\gamma}}(\sum_{i=1}^r \widehat{\boldsymbol{S}}_i \circ \widehat{\boldsymbol{c}}_i) + \widehat{\boldsymbol{F}}$;

Output: the reconstructed video $\widehat{\underline{\boldsymbol{X}}}$.

V. Experiments

In this section, we validate the proposed `FactorDVP-T` on diverse datasets, including fourteen grayscale videos with static background and three color videos with dynamic background. Notably, rather than only using the commonly used benchmarks in [14], [36], [24], other videos are also utilized to verify the data adaptability of our methods. We select a wide range of compared methods to showcase the effectiveness of our method, including two model-based methods, two PnP-based methods, an unsupervised deep learning-based method, and a supervised deep learning-based method. First, we provide the experiment settings in Subsection V-A. Then, we evaluate the proposed methods on simulation and real datasets in Subsection V-B and Subsection V-C respectively.

A. Settings

In the simulation experiment, we select fourteen different grayscale videos with static background (i.e., *Aerial*, *Vehical*, *Kobe*, *Traffic*, *Drop*, *Runner*, *Truck*, *River*, *Trees*, *Water*, *Fountain*, *Highway*, *Lobby*, and *Bridge*)[1,2] and three color videos (i.e., *Waterfall*, *Flower*, and *Park*)[3] with dynamic background to test the performance of the proposed `FactorDVP-T`. Among them, video *Aerial*, *Vehical*, *Kobe*, *Traffic*, *Drop*, and

[1] http://trace.eas.asu.edu/yuv/
[2] http://perception.i2r.a-star.edu.sg/bkmodel/bkindex.html
[3] http://trace.eas.asu.edu/yuv/

TABLE I
QUANTITATIVE COMPARISON OF THE RECONSTRUCTED RESULTS BY DIFFERENT METHODS. THE **BEST** AND SECOND BEST VALUES ARE HIGHLIGHTED IN BOLD AND UNDERLINED, RESPECTIVELY.

Data	*Fountain*			*Highway*			*Trees*			*Water*			*Aerial*			*Truck*			*River*			*Drop*			*Runner*		
Index	PSNR	SSIM	Time	PSNR	SSIM	Time	PSNR	SSIM	Time	PSNR	SSIM	Time	PSNR	SSIM	Time	PSNR	SSIM	Time	PSNR	SSIM	Time	PSNR	SSIM	Time	PSNR	SSIM	Time
GAP-TV	21.59	0.71	0.34	23.01	0.79	0.02	17.09	0.52	0.34	22.34	0.66	0.34	25.05	0.82	0.54	24.47	0.85	0.54	24.44	0.77	0.51	34.74	0.97	0.53	28.81	0.90	0.53
PnP	19.58	0.65	0.48	22.35	0.75	0.19	15.75	0.40	0.44	22.93	0.68	0.45	24.02	0.81	0.16	23.62	0.82	0.15	21.13	0.62	0.15	40.70	0.98	0.15	32.15	0.93	0.15
PnP-TV	23.98	0.82	2.62	27.75	0.88	1.10	19.75	0.68	3.07	24.81	0.65	2.67	24.31	0.83	1.11	27.22	0.90	1.12	26.28	0.84	1.10	35.50	0.97	1.10	29.95	0.92	1.12
DeSCI	25.84	0.81	40.85	24.73	0.82	595.61	22.41	0.63	40.95	24.09	0.83	40.12	25.33	0.85	643.62	29.28	**0.95**	633.56	26.26	0.84	627.22	**43.22**	**0.99**	642.65	**38.76**	**0.97**	635.18
DVP	25.44	0.66	19.31	20.88	0.62	16.29	21.14	0.49	18.29	24.86	0.80	18.29	21.19	0.66	16.01	19.61	0.69	15.98	24.81	0.69	16.27	18.78	0.51	16.13	19.42	0.64	16.29
RevSCI	30.09	**0.89**	0.03	31.07	0.93	0.02	**26.62**	**0.85**	0.02	**32.94**	0.88	0.02	**29.10**	**0.92**	0.03	**30.47**	0.92	0.02	**33.76**	**0.96**	0.02	39.84	**0.99**	0.02	36.14	**0.97**	0.02
FactorDVP-T	**31.81**	**0.89**	14.78	**34.06**	**0.97**	15.14	24.59	0.72	15.14	32.51	**0.96**	14.86	26.84	0.86	15.42	29.57	0.94	16.44	30.81	0.91	15.35	36.69	0.97	16.24	30.76	0.89	15.99

Data	*Lobby*			*Bridge*			*Vehicle*			*Kobe*			*Traffic*			*Waterfall*			*Flower*			*Park*			*Average*		
Index	PSNR	SSIM	Time	PSNR	SSIM	Time	PSNR	SSIM	Time	PSNR	SSIM	Time	PSNR	SSIM	Time	PSNR	SSIM	Time	PSNR	SSIM	Time	PSNR	SSIM	Time	PSNR	SSIM	Time
GAP-TV	22.30	0.82	0.01	24.83	0.75	0.01	24.82	0.838	0.52	26.45	0.85	0.53	20.89	0.71	0.55	25.81	0.73	1.63	19.82	0.76	1.68	19.45	0.59	1.61	23.87	0.76	0.60
PnP	22.02	0.79	0.11	21.84	0.63	0.12	25.42	0.85	0.15	30.50	0.92	0.14	24.18	0.83	0.14	21.82	0.53	0.49	16.99	0.51	0.53	17.42	0.55	0.45	23.67	0.72	0.26
PnP-TV	26.62	0.90	0.68	26.29	0.79	0.67	25.30	0.80	1.10	26.42	0.84	1.11	21.60	0.74	1.11	27.03	0.71	3.36	20.53	0.77	3.42	20.45	0.66	3.34	25.51	0.80	1.75
DeSCI	24.09	0.83	40.12	25.84	0.81	40.85	27.04	0.91	638.76	**33.25**	**0.95**	635.26	**28.72**	**0.92**	641.37	24.65	0.70	1930.85	19.86	0.75	1929.66	17.58	0.51	1928.36	27.11	0.83	687.35
DVP	24.86	0.80	18.29	25.44	0.66	19.31	18.34	0.54	15.99	18.77	0.51	16.15	18.61	0.57	16.22	23.76	0.49	97.01	17.71	0.63	88.83	12.23	0.23	102.21	20.93	0.59	30.99
RevSCI	31.46	0.95	0.02	30.96	0.88	0.03	**27.94**	**0.92**	0.02	31.80	0.92	0.02	27.93	**0.92**	0.02	31.52	0.93	0.11	23.36	0.88	0.12	22.99	0.76	0.10	**30.47**	**0.91**	0.04
FactorDVP-T	**32.51**	**0.96**	14.86	**31.81**	**0.89**	14.78	26.05	0.85	15.41	25.54	0.74	16.08	23.38	0.76	15.98	**34.52**	**0.95**	46.64	**25.26**	**0.91**	46.68	**23.88**	**0.77**	46.32	29.44	0.88	20.91

Runner are selected from the commonly used benchmarks for video SCI, and others are used to validate the data adaptability of the proposed methods. Additionally, two real captured videos *UCF* and *Handlen* [57] is selected in our real experiment. Before the experiment, the pixel values of all datasets are normalized to [0, 1] band-by-band.

To thoroughly evaluate the performance of this method, we compare it with two model-based methods (e.g., `GAP-TV` [13] and `DeSCI` [14]), two PnP-based methods (e.g., `PnP` [36] and `PnP-TV` [37]), a supervised deep learning-based method (e.g., `RevSCI` [24]), and an unsupervised learning-based method (e.g., `DVP` [58]). Peak signal-to-noise ratio (PSNR) and structure similarity (SSIM) are selected to evaluate the performance of these methods [47]. For `GAP-TV`, `PnP`, `PnP-TV`, and `DeSCI`, their parameters are manually adjusted according to the authors' suggestions to uplift their performance. For `RevSCI`, we retrain the networks with our sensing mask and training datasets generated from DAVIS2017 [59] following the authors' suggestions. For `DVP`, we set the max iteration as 10000, and report the highest PSNR and SSIM during the iteration. In the proposed `FactorDVP-T`, only one parameter needs to be manually tuned, e.g., the number of temporal bases r—which is selected from $\{1, 3\}$ for all selected videos. Moreover, the learning rate is selected from $\{0.01,\ 0.001,\ 0.0001\}$, and the max iteration is set as 4000 and 10000 for grayscale videos and color videos, respectively.

The experiments of `PnP`, `PnP-TV`, `DVP`, `RevSCI`, and `FactorDVP-T` are executed using Python on a computer with an AMD Ryzen 7 5800X 8-Core processor @ 3.79 GHz, 32.0 GB of RAM, and an NVIDIA GeForce RTX 3090 GPU. The experiments of `GAP-TV`, and `DeSCI` are implemented in MATLAB (2021a) on the same computer.

B. Simulated Data Experiment Results

Table I lists PSNR, SSIM, and execution time (in minutes) of the compared methods on grayscale and color videos respectively. The best and the second-best results for each quality index are highlighted in boldface and underlined, respectively. We can observe that our method generally outperforms `DeSCI` with an average improvement of 2.33 dB in PSNR, especially on some datasets with fine details, e.g., *Fountain*, *Highway*, *Trees*, and *Water*. For the testing dataset in `RevSCI`, the supervised `RevSCI` performs better than our unsupervised method. The underlying reason may be the distribution of these testing datasets is similar to its training dataset. For the wild datasets not in the testing dataset, our method performs better than `RevSCI`, e.g., *Fountain*, *Highway*, *Waterfall*, *Flower*, and *Park*. This observation reveals the adaptability of our method as compared with the model-based and supervised deep learning-based methods.

The reconstructed results on gray-scale and color videos by the compared methods are shown in Fig. 4. For better visualization, two regions are chosen and enlarged. We can observe that, `PnP`, `PnP-TV`, and `DeSCI` are with relatively dissatisfactory performance, especially for parts with so many fine details. Compared with these three methods, `RevSCI` performs better by training on a large number of external datasets, but some blurring details still exist. In contrast, our `FactorDVP-T` could preserve the most detailed information and demonstrates the best performance among the compared methods, which is consistent with its good performance on PSNR and SSIM. We conjecture that such promising results can be attributed to the strong adaptability to various videos of the unsupervised networks embedded in the physicallyly interpreted video decomposition framework, which is beneficial to preserve complex scenes with plenty of details.

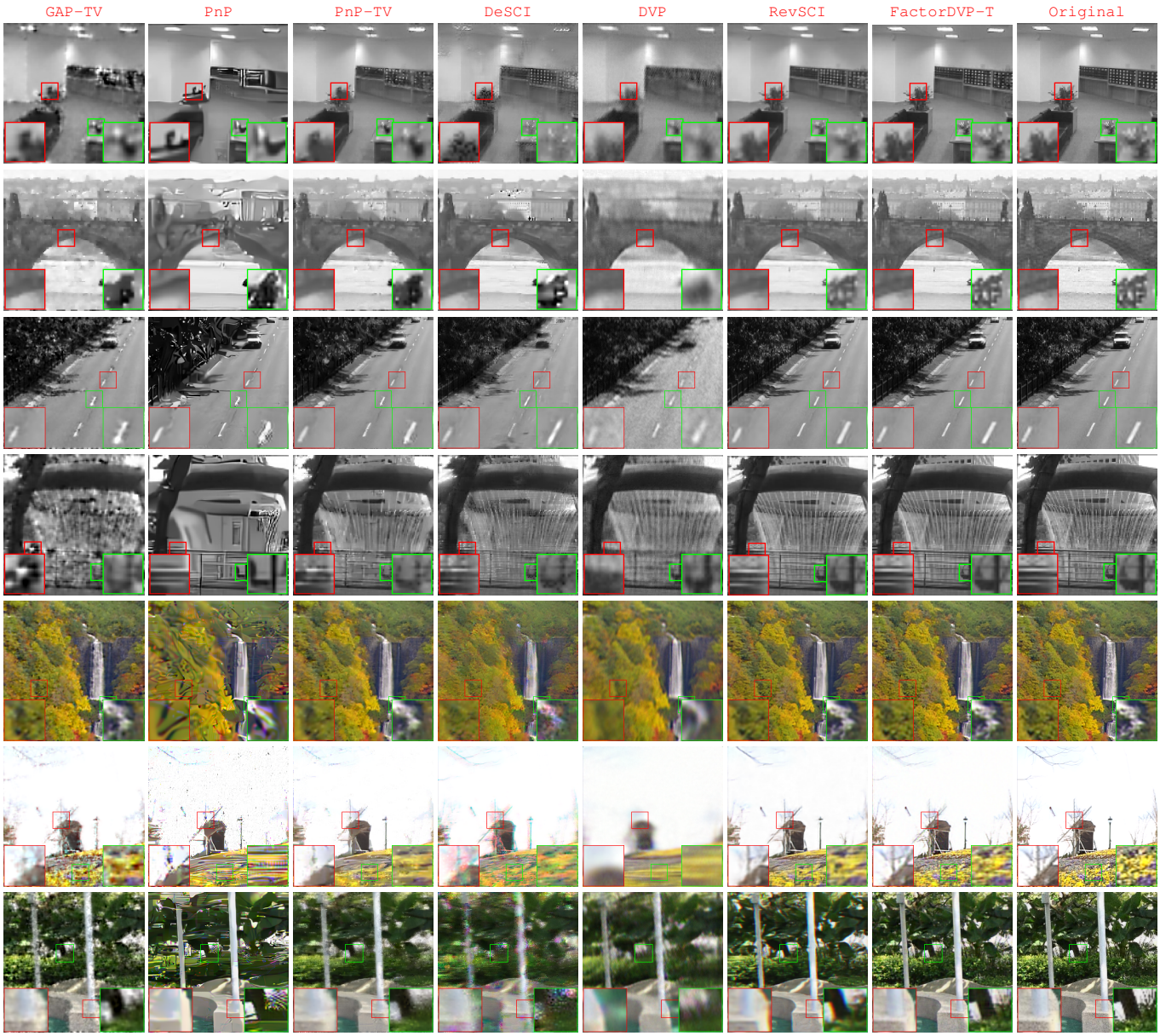

Fig. 4. The reconstructed results produced by the compared methods. From top to bottom: the 1-st frame of *Lobby*, *Bridge*, *Highway*, *Fountain*, *Waterfall*, *Flower*, and *Park*, respectively. From left to right: the reconstructed results produced by GAP-TV, PnP, PnP-TV, DeSCI, DVP, RevSCI, the proposed FactorDVP-T, and the original videos, respectively.

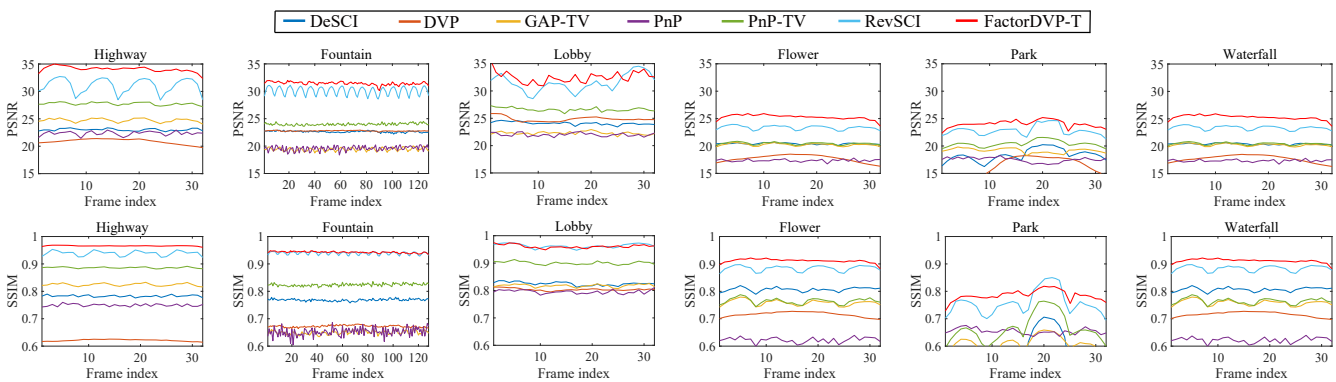

Fig. 5. PSNR and SSIM values of all frames obtained by different methods on video *Highway*, *Fountain*, *Lobby*, *Flower*, *Park*, and *Waterfall*.

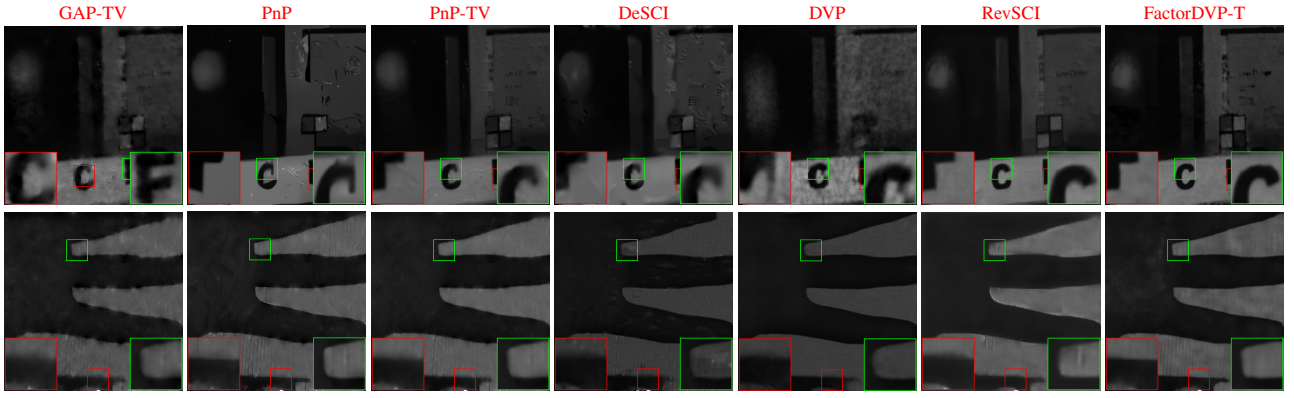

Fig. 6. Visual comparison on real video *UCF* and *Handlen*. From top to bottom: the selected frame of *UCF* and *Handlen*, respectively. From left to right: the reconstructed results produced by `GAP-TV`, `PnP`, `PnP-TV`, `DeSCI`, `DVP`, `RevSCI`, and the proposed `FactorDVP-T`, respectively.

To test our methods performance on every frame, each frame's PSNR and SSIM values on six selected data are shown in Fig. 5. As observed, in most frames, the PSNR values of the proposed `FactorDVP-T` are higher than those of the other compared methods.

Moreover, in order to verify our motivation, the learned background and foreground by our method are shown in Fig. 7. Videos *Highway*, *Water*, *Truck*, and *Drop* are selected as examples. We can observe that the deep background prior and deep foreground prior could be indeed captured in `FactorDVP-T` respectively.

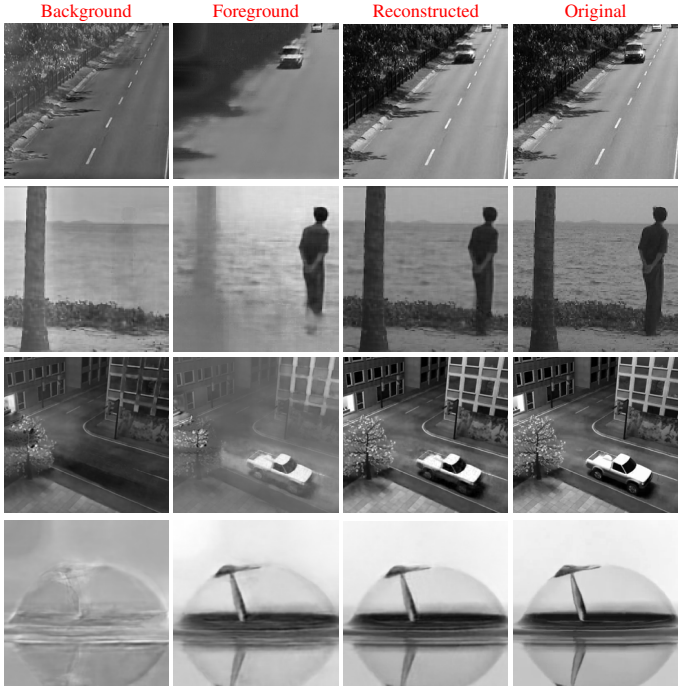

Fig. 7. The learned backgrounds and foregrounds by our method. From top to bottom: the selected frame of video *Highway*, *Water*, *Truck*, and *Drop*, respectively.

C. Real Data Experiment Results

In this part, we demonstrate the efficacy of the proposed `FactorDVP-T` with real data captured from SCI systems. We consider two real grayscale videos *UCF* and *Handlen* which are captured by the CACTI system [1]. For video *UCF*, the snapshot measurement of size 256 $\times$ 256 pixels encodes 10 frames of the same size. For video *Handlen*, the snapshot measurement of size 256 $\times$ 256 pixels encodes 14 frames of the same size. The sensing masks are the same as that used in [14]. The corresponding results are shown in Fig. 6. It can be seen clearly that all of the model-based methods, including `GAP-TV`, `PnP`, `PnP-TV`, and `DeSCI`, lose much detailed information and lead to blurring details. In contrast, the supervised method `RevSCI` and our unsupervised method `FactorDVP-T` both offer visually more pleasing results. Notably, as compared with supervised `RevSCI`, our unsupervised method could preserve more sharp edges and fine details; see the zoom-in region for edges and details preservation. This further verifies the strong adaptability to videos of our methods.

VI. DISCUSSIONS

In this section, we present some necessary discussions about the proposed `FactorDVP-T`.

A. Effectiveness of The Deep Foreground Prior

To verify the effectiveness of deep foreground prior, we compare it with some commonly used hand-crafted priors (e.g., sparse prior and total variation prior) in the proposed `FactorDVP-T`. The reconstructed videos and the corresponding foregrounds are shown in Fig. 8. We can observe that our designed deep foreground prior could be captured more faithfully relative to the sparse prior and total variation prior. And the performance of deep background prior with deep foreground prior and affine transformation (i.e., `FactorDVP-T`) is much more visually pleasing compared with other methods. These observations support our idea for imposing the deep foreground prior on the foreground.

B. Effectiveness of U-Net Structure

In this part, we evaluate the effectiveness of the 2D and 3D U-Net structure in `FactorDVP-T` by replacing them with the corresponding state-of-the-art network (e.g., ResNet). To

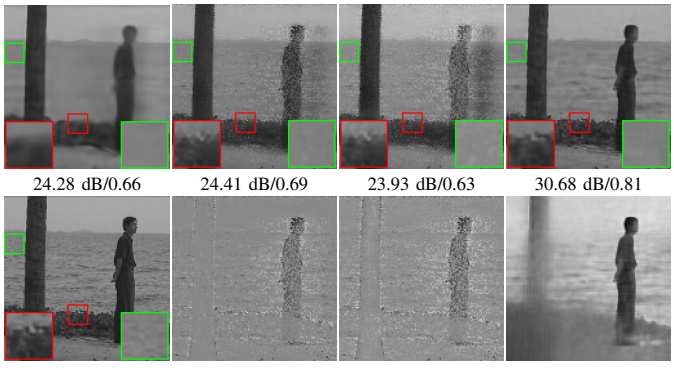

Fig. 8. Effectiveness of the deep foreground prior, sparse prior, and total variation prior for foreground modeling. 1st row: the reconstructed video by deep background prior, deep background prior with sparse prior, deep background prior with total variation prior, and deep background prior with deep foreground prior, with affine transformation, respectively. 2nd row: the original video, the foreground captured by sparse prior, total variation prior, and our deep foreground prior, respectively.

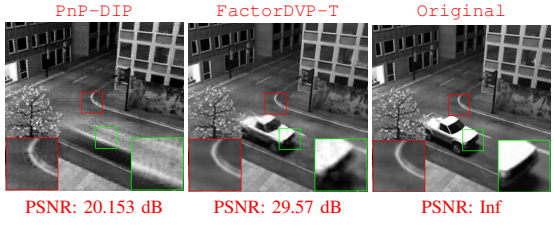

Fig. 9. The reconstructed videos by our method and PnP-DIP.

be more specific, the compared methods are listed as follows:

- **2D ResNet + 3D U-Net**: In this method, 2D ResNet is utilized for background spatial coefficient modeling, and 3D U-Net is used for moving foreground modeling.
- **2D U-Net + 3D ResNet**: In this method, 2D U-Net is utilized for background spatial coefficient modeling and 3D ResNet is used for moving foreground modeling.
- **2D ResNet + 3D ResNet**: In this method, 2D ResNet is utilized for background spatial coefficient modeling and 3D ResNet is used for moving foreground modeling.
- **2D U-Net + 3D U-Net**: This is our FactorDVP-T, in which 2D U-Net is utilized for background spatial coefficient modeling and 3D U-Net is used for moving foreground modeling.

The results are reported in Table II. We can observe that the method with 2D U-Net and 3D U-Net demonstrates the best performance. This observation reveals that U-Net has more powerful ability in encoding image priors as compared with ResNet.

TABLE II
QUANTITATIVE COMPARISON OF DIFFERENT METHODS ON VIDEO *Highway*. THE **BEST** VALUES ARE HIGHLIGHTED IN BOLD.

Method	PSNR	SSIM
2D ResNet + 3D U-Net	32.07	0.93
2D U-Net + 3D ResNet	31.91	0.90
2D ResNet + 3D ResNet	30.27	0.89
2D U-Net + 3D U-Net	**34.06**	**0.97**

C. FactorDVP-T v.s. PnP-DIP

PnP-DIP has shown excellent performance on SCI reconstruction for multi-spectral images. However, for video, the performance of this method would be compromised because it pays less attention to video data's unique traits, e.g., moving foreground and dynamic backgrounds. Distinct from PnP-DIP, in FactorDVP-T, we disentangle background and foreground, and model them individually. The reconstructed videos are shown in Fig. VI-A. We can observe that PnP-DIP shows promising performance in recovering the background while blurring the foreground. This observation reveals that PnP-DIP sometimes fails to capture the foreground information faithfully. In contrast, our FactorDVP-T presents satisfactory performance both in recovering foreground and background.

D. Sensitivity Analysis of the Parameter r

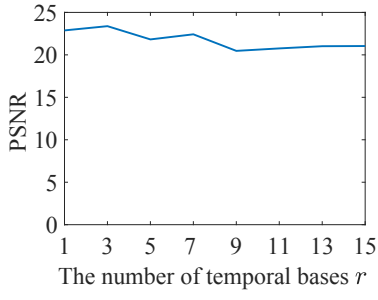

Fig. 10. PSNR values of the results by FactorDVP-T with different r on video *Trees*.

We study the parameter sensitivity of the number of temporal bases r, which mainly depicts the low rankness of background. Fig. 10 displays the PSNR values of the results by FactorDVP-T with different temporal bases number r. We can observe that the proposed FactorDVP-T exhibits stable and superior performance within a certain range of r (r =1, 3). Considering that larger r would lead to unsatisfactory visual quality, we set r to be 3 for video *Trees* in our experiments.

E. Robustness to Noise

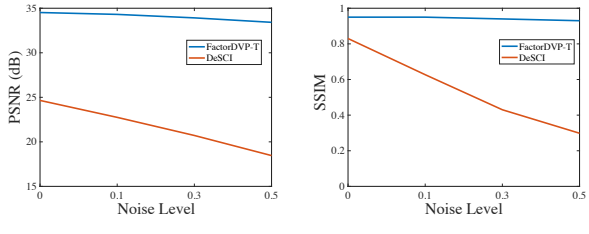

Fig. 11. Comparison of FactorDVP-T and DeSCI with noisy measurements on video *Waterfall*.

In real scenario, the captured videos are inevitably corrupted by noise, which usually leads to the performance degradation for further video processing. To verify the robustness to noise of our method, we perform the experiments on *Waterfall* datasets by adding different levels of white Gaussian noise.

The results produced by `FactorDVP-T` and `DeSCI` are summarized in Fig. 11. As one can see, the performance of `DeSCI` deteriorates significantly as the noise level increases, but our method does not fluctuate greatly. Therefore, our method is recommended in realistic system with noise.

VII. Conclusion

We extend the well-known DIP from natural images to videos and further propose a domain-factorized DVP-based method (i.e., `FactorDVP-T`) for video SCI reconstruction. In this work, inspired by the physicallyly interpreted video decomposition, we first decompose a video into a foreground and a background. And then, the background could be further decomposed into spatial coefficients and temporal bases. In this way, 2D U-Nets and FCNs could be employed to model these factors individually. Besides, 3D U-Net is employed to model the foreground. In addition, we employ affine transformation to overcome the dynamic background challenge in real scenarios. Extensive experiments on various videos verify that the proposed method can better adapt to various videos in real scenarios, as compared with several state-of-the-art methods. Moreover, the limitation of our method is the relatively long running time. The reason is that adaptively learning different CNN parameters for each observed measurement is time-consuming. One of the directions for future research is how to overcome this limitation.

References

[1] P. Llull, X. Liao, X. Yuan, J. Yang, D. Kittle, L. Carin, G. Sapiro, and D. J. Brady, "Coded aperture compressive temporal imaging," *Optics express*, vol. 21, no. 9, pp. 10526–10545, 2013. 1, 3, 8

[2] L. Song, L. Wang, M. H. Kim, and H. Huang, "High-accuracy image formation model for coded aperture snapshot spectral imaging," *IEEE Transactions on Computational Imaging*, vol. 8, pp. 188–200, 2022. 1

[3] A. Wagadarikar, R. John, R. Willett, and D. Brady, "Single disperser design for coded aperture snapshot spectral imaging," *Applied optics*, vol. 47, no. 10, pp. B44–B51, 2008. 1

[4] S. Jalali and X. Yuan, "Snapshot compressed sensing: Performance bounds and algorithms," *IEEE Transactions on Information Theory*, vol. 65, no. 12, pp. 8005–8024, 2019. 1

[5] S. Zhang, Y. Xia, Y. Xia, and J. Wang, "Matrix-Form neural networks for complex-variable basis pursuit problem with application to sparse signal reconstruction," *IEEE Transactions on Cybernetics*, pp. 1–11, 2021. 1

[6] Y. Zhao, X. Liao, X. He, and R. Tang, "Centralized and collective neurodynamic optimization approaches for sparse signal reconstruction via ℓ_1-minimization," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2021. 1

[7] S. Athey, M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi, "Matrix completion methods for causal panel data models," *Journal of the American Statistical Association*, vol. 0, no. 0, pp. 1–15, 2021. 1

[8] O. Lebedeva, A. Osinsky, and S. Petrov, "Low-rank approximation algorithms for matrix completion with random sampling," *Computational Mathematics and Mathematical Physics*, vol. 61, no. 5, pp. 799–815, 2021. 1

[9] L. Wang, S. Zhang, and H. Huang, "Adaptive dimension-discriminative low-rank tensor recovery for computational hyperspectral imaging," *International Journal of Computer Vision*, pp. 1–20, 2021. 1

[10] H. Wang, F. Zhang, J. Wang, T. Huang, J. Huang, and X. Liu, "Generalized nonconvex approach for low-tubal-rank tensor recovery," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2021. 1

[11] J. Yang, X. Yuan, X. Liao, P. Llull, D. J. Brady, G. Sapiro, and L. Carin, "Video compressive sensing using gaussian mixture models," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4863–4878, 2014. 1

[12] J. Yang, X. Liao, X. Yuan, P. Llull, D. J. Brady, G. Sapiro, and L. Carin, "Compressive sensing by learning a gaussian mixture model from measurements," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 106–119, 2015. 1

[13] X. Yuan, "Generalized alternating projection based total variation minimization for compressive sensing," in *ICIP*, pp. 2539–2543, 2016. 1, 6

[14] Y. Liu, X. Yuan, J. Suo, D. J. Brady, and Q. Dai, "Rank minimization for snapshot compressive imaging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 12, pp. 2990–3006, 2018. 1, 5, 6, 8

[15] S. Wang, J. Lv, Z. He, D. Liang, Y. Chen, M. Zhang, and Q. Liu, "Denoising auto-encoding priors in undecimated wavelet domain for MR image reconstruction," *Neurocomputing*, vol. 437, pp. 325–338, 2021. 1

[16] M. Kellman, K. Zhang, E. Markley, J. Tamir, E. Bostan, M. Lustig, and L. Waller, "Memory-efficient learning for large-scale computational imaging," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1403–1414, 2020. 1

[17] C. D. Bahadir, A. Q. Wang, A. V. Dalca, and M. R. Sabuncu, "Deep-learning-based optimization of the under-sampling pattern in MRI," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1139–1152, 2020. 1

[18] G. Oh, B. Sim, H. Chung, L. Sunwoo, and J. C. Ye, "Unpaired deep learning for accelerated MRI using optimal transport driven cyclegan," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1285–1296, 2020. 1

[19] Y. Sanghvi, Y. Kalepu, and U. K. Khankhoje, "Embedding deep learning in inverse scattering problems," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 46–56, 2019. 1

[20] Y. Zhang, T. Lv, R. Ge, Q. Zhao, D. Hu, L. Zhang, J. Liu, Y. Zhang, Q. Liu, W. Zhao, and Y. Chen, "CD-Net: Comprehensive domain network with spectral complementary for dect sparse-view reconstruction," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 436–447, 2021. 1

[21] H. Zhou, Y. Wang, Q. Liu, and Y. Wang, "Rnmf-guided deep network for signal separation of GPR without labeled data," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021. 1

[22] G. Tsagkatakis, M. Bloemen, B. Geelen, M. Jayapala, and P. Tsakalides, "Graph and rank regularized matrix recovery for snapshot spectral image demosaicing," *IEEE Transactions on Computational Imaging*, vol. 5, no. 2, pp. 301–316, 2018. 1

[23] M. Iliadis, L. Spinoulas, and A. K. Katsaggelos, "Deep fully-connected networks for video compressive sensing," *Digital Signal Processing*, vol. 72, pp. 9–18, 2018. 1

[24] Z. Cheng, B. Chen, G. Liu, H. Zhang, R. Lu, Z. Wang, and X. Yuan, "Memory-efficient network for large-scale video compressive sensing," in *CVPR*, pp. 16246–16255, June 2021. 1, 5, 6

[25] Z. Wang, H. Zhang, Z. Cheng, B. Chen, and X. Yuan, "Metasci: Scalable and adaptive reconstruction for video compressive sensing," in *CVPR*, pp. 2083–2092, 2021. 1

[26] Y. Liu, Q. Liu, M. Zhang, Q. Yang, S. Wang, and D. Liang, "IFR-Net: Iterative feature refinement network for compressed sensing MRI," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 434–446, 2019. 1

[27] Z. Cheng, R. Lu, Z. Wang, H. Zhang, B. Chen, Z. Meng, and X. Yuan, "Birnat: Bidirectional recurrent neural networks with adversarial training for video snapshot compressive imaging," in *ECCV*, pp. 258–275, Springer, 2020. 1

[28] R. Lu, Z. Cheng, B. Chen, and X. Yuan, "Motion-aware dynamic graph neural network for video compressive sensing," *arXiv preprint arXiv:2203.00387*, 2022. 1

[29] J. Ma, X.-Y. Liu, Z. Shou, and X. Yuan, "Deep tensor admm-net for snapshot compressive imaging," in *ICCV*, pp. 10222–10231, 2019. 1

[30] X. Han, B. Wu, Z. Shou, X.-Y. Liu, Y. Zhang, and L. Kong, "Tensor FISTA-Net for real-time snapshot compressive imaging," in *AAAI*, vol. 34, pp. 10933–10940, 2020. 1

[31] Z. Meng, S. Jalali, and X. Yuan, "GAP-Net for snapshot compressive imaging," *arXiv preprint arXiv:2012.08364*, 2020. 1

[32] Y. Li, M. Qi, R. Gulve, M. Wei, R. Genov, K. N. Kutulakos, and W. Heidrich, "End-to-end video compressive sensing using anderson-accelerated unrolled networks," in *ICCP*, pp. 1–12, 2020. 1

[33] Z. Wu, Z. Zhang, J. Song, and M. Zhang, "Spatial-temporal synergic prior driven unfolding network for snapshot compressive imaging," in *ICME*, pp. 1–6, 2021. 1

[34] Z. Wu, J. Zhang, and C. Mou, "Dense deep unfolding network with 3d-cnn prior for snapshot compressive imaging," in *ICCV*, pp. 4892–4901, 2021. 1

[35] K. Zhang, W. Zuo, and L. Zhang, "FFDNet: Toward a fast and flexible solution for CNN-based image denoising," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4608–4622, 2018. 1

[36] X. Yuan, Y. Liu, J. Suo, and Q. Dai, "Plug-and-play algorithms for large-scale snapshot compressive imaging," in *CVPR*, pp. 1447–1457, 2020. 1, 5, 6

[37] H. Qiu, Y. Wang, and D. Meng, "Effective snapshot compressive-spectral imaging via deep denoising and total variation priors," in *CVPR*, pp. 9127–9136, 2021. 1, 6

[38] E. Cands, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM*, vol. 58, no. 3, pp. 1–37, 2011. 2

[39] O. Sidorov and J. Y. Hardeberg, "Deep hyperspectral prior: Single-image denoising, inpainting, super-resolution," in *ICCVW*, pp. 3844–3851, 2019. 2, 3, 4

[40] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006. 2, 3

[41] N. Saha, M. S. Ifthekhar, N. T. Le, and Y. M. Jang, "Survey on optical camera communications: challenges and opportunities," *IET Optoelectronics*, vol. 9, no. 5, pp. 172–183, 2015. 2

[42] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *CVPR*, pp. 9446–9454, 2018. 3

[43] A. Shocher, N. Cohen, and M. Irani, "zero-shot super-resolution using deep internal learning," in *CVPR*, pp. 3118–3126, 2018. 3

[44] D. Bau, H. Strobelt, W. Peebles, J. Wulff, B. Zhou, J.-Y. Zhu, and A. Torralba, "Semantic photo manipulation with a generative image prior," *ACM Transactions on Graphics*, vol. 38, no. 4, pp. 1–11, 2019. 3

[45] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *ICCV*, pp. 5933–5942, 2019. 3

[46] Y. Gandelsman, A. Shocher, and M. Irani, "Double-DIP: Unsupervised image decomposition via coupled deep-image-priors," in *CVPR*, pp. 11026–11035, 2019. 3

[47] Y.-C. Miao, X.-L. Zhao, X. Fu, J.-L. Wang, and Y.-B. Zheng, "Hyperspectral denoising using unsupervised disentangled spatiospectral deep priors," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–16, 2021. 3, 4, 6

[48] Z. Meng, Z. Yu, K. Xu, and X. Yuan, "Self-supervised neural networks for spectral snapshot compressive imaging," in *ICCV*, pp. 2622–2631, 2021. 3

[49] S. Shrestha, X. Fu, and M. Hong, "Deep generative model learning for blind spectrum cartography with nmf-based radio map disaggregation," in *ICASSP*, pp. 4920–4924, 2021. 4

[50] M. Ding, X. Fu, T.-Z. Huang, J. Wang, and X.-L. Zhao, "Hyperspectral super-resolution via interpretable block-term tensor modeling," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 3, pp. 641–656, 2021. 4

[51] G. Zhang, X. Fu, J. Wang, X.-L. Zhao, and M. Hong, "Spectrum cartography via coupled block-term tensor decomposition," *IEEE Transactions on Signal Processing*, vol. 68, pp. 3660–3675, 2020. 4

[52] Y. Qian, F. Xiong, S. Zeng, J. Zhou, and Y. Y. Tang, "Matrix-vector nonnegative tensor factorization for blind unmixing of hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 3, pp. 1776–1792, 2017. 4

[53] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013. 4

[54] Z. Lv, C. Beall, P. F. Alcantarilla, F. Li, Z. Kira, and F. Dellaert, "A continuous optimization approach for efficient and accurate scene flow," in *ECCV*, pp. 757–773, Springer, 2016. 5

[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015. 5

[56] R. Rojas, *The Backpropagation Algorithm*, pp. 149–182. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996. 5

[57] Y. Sun, X. Yuan, and S. Pang, "Compressive high-speed stereo imaging," *Optics Express*, vol. 25, no. 15, pp. 18182–18190, 2017. 6

[58] C. Lei, Y. Xing, and Q. Chen, "Blind video temporal consistency via deep video prior," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1083–1093, 2020. 6

[59] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool, "The 2017 davis challenge on video object segmentation," *arXiv:1704.00675*, 2017. 6