Unfolding Framework with Prior of Convolution-Transformer Mixture and Uncertainty Estimation for Video Snapshot Compressive Imaging

Siming Zheng
Computer Network Information Center,Chinese Academy of Science,
Beijing, 100190, China
University of Chinese Academy of Sciences,
Beijing, 100049, China
zhengsiming@cnic.cn

Xin Yuan
School of Enginering, Westlake Universiy
Hangzhou, 310024, China
xylab@westlake.edu.cn

Abstract

We consider the problem of video snapshot compressive imaging (SCI), where sequential high-speed frames are modulated by different masks and captured by a single measurement. The underlying principle of reconstructing multi-frame images from only one single measurement is to solve an ill-posed problem. By combining optimization algorithms and neural networks, deep unfolding networks (DUNs) score tremendous achievements in solving inverse problems. In this paper, our proposed model is under the DUN framework and we propose a 3D Convolution-Transformer Mixture (CTM) module with a 3D efficient and scalable attention model plugged in, which helps fully learn the correlation between temporal and spatial dimensions by virtue of Transformer. To our best knowledge, this is the first time that Transformer is employed to video SCI reconstruction. Besides, to further investigate the high-frequency information during the reconstruction process which are neglected in previous studies, we introduce ***variance estimation*** *characterizing the* ***uncertainty*** *on a pixel-by-pixel basis. Extensive experimental results demonstrate that our proposed method achieves state-of-the-art (SOTA) (with a* ***1.2dB*** *gain in PSNR over previous SOTA algorithm) results. Code can be found on* https://github.com/zsm1211/CTM-SCI.

1. Introduction

Nowadays, due to the ability of capturing high-dimensional data in an efficient way, Snapshot Compressive Imaging (SCI) [32, 59] has attracted much attention. SCI system just employs a low-speed 2D camera to capture 3D sequential video frames, hyperspectral data, etc. [66, 3], where a digital micro-mirror device [17, 39] or a shifting mask [61] is utilized to modulate consequent frames. With

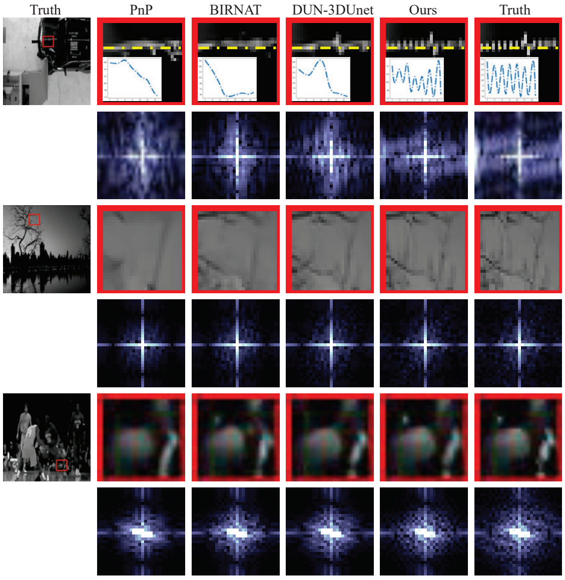

Figure 1. Illustration of comparison between the reconstruction results in the high-frequency details of previous SOTA algorithms and Ours. We present the details both in image domain (first line) and frequency domain (second line). To be more visually clearly, we also present the intensity profiles extracted from the cross-section yellow lines in the first example. Our proposed algorithm can reconstruct better high-frequency details.

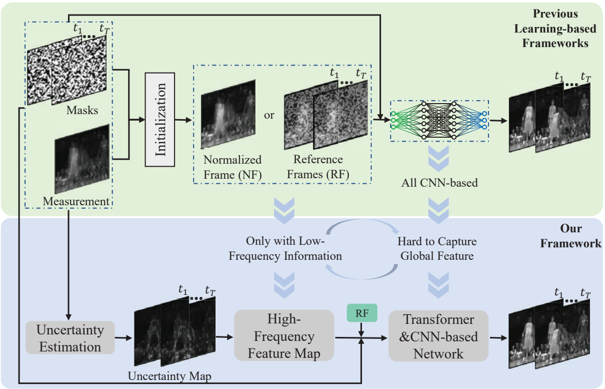
Previous Learning-based Frameworks
t_1 ... t_T
Masks
Measurement
Initialization
Normalized Frame (NF)
or
Reference Frames (RF)
All CNN-based
Only with Low-Frequency Information
Hard to Capture Global Feature
Our Framework
Uncertainty Estimation
Uncertainty Map
High-Frequency Feature Map
RF
Transformer &CNN-based Network

Figure 2. Illustration of the design motivation.

the knowledge of modulation, the captured single 2D measurement can be reconstructed to original sequential frames by algorithms [19, 50, 26, 51, 65, 62, 56]. In this paper, we focus on the reconstruction problem of video SCI systems.

To be concrete, the reconstruction process can be regard as solving an ill-posed inverse problem where the number of pixels to be reconstructed is much higher than the number of known parameters. With the development of deep learning, deep neural networks has been employed to conduct the reconstruction in recent years, where convolutional neural networks (CNN) are dominating. Compared to optimization-based algorithms, learning-based algorithms can directly map the measurement and the target images which makes it easier and faster to bring reconstruction results up to the mark. To improve learning-based algorithms' defect of lacking interpretability, recently proposed deep unfolding networks (DUNs) [50, 26, 51] combine the merits of both optimization-based and learning-based algorithms, and achieve the best results so far.

Motivation: As shown in Fig. 1, previous SOTA learning-based algorithms [50, 5, 60] are not ideal for high-frequency detail reconstruction. In their initialization (details in Fig. 3), as shown in the top-middle of Fig. 2, the NF and RF both have a relatively clean background and clear stationary areas, which can directly feed the low-frequency information to the following network. However, **the high-frequency features such as edges and textures can not be directly obtained from the measurement and are neglected by previous studies for video SCI**. For the network structure, convolution-based backbone architectures have long dominated visual modeling in computer vision [25, 40, 18, 16]. It is the same in the video SCI tasks, all previous SOTA learning-based algorithms are CNN-based. Although CNN has many advantages, its receptive field is usually small and relies on deeper layers or larger convolution kernels, which is not conducive to capture global features such as contour features and texture features which are also the high-frequency features. By contrast, Transformer can well capture long-distance dependencies and global inter-dependence between different regions, yet few researches of applying Transformer to video SCI are carried out.

To sum up, previous learning-based frameworks mainly suffer from following two problems: 1) High-frequency information is not taken into consideration. 2) Compared to Transformer, CNNs are weak in capturing global features, some of which are also high-frequency features like contour features and texture features. Due to the mutual influence of these two aspects in the reconstruction process, the fidelity of the high-frequency details is compromised.

Contributions: Towards this end, hereby, we propose a *Transformer enabled deep unfolding framework* for video SCI and we further introduce *uncertainty estimation to take high-frequency information as regularized prior under the unfolding framework into consideration* for better reconstruction. Our contributions can be summarized as follows:

1) We propose a novel video **Convolution-Transformer module**, dubbed CTM, for video SCI that can well capture local and global spatial-temporal interactions which is composed of 3D CNN, 3D scalable *blocked dense and dilated sparse attention*. Note that the attention modules take both local and global information into consideration with only a linear complexity.
2) Unlike previous studies that only consider the low-frequency information such as the information of stationary areas or backgrounds [5, 50, 4], we first bring **high-frequency information as regularized prior under the unfolding framework in video SCI for focusing on areas with high reconstruction uncertainty and improving the fidelity of reconstruction**, which is achieved by the variance estimation characterizing the *uncertainty* on a pixel-by-pixel basis.
3) We first introduce Transformer for video SCI reconstruction. Both real and simulation experiments demonstrate that **our proposed framework outperform previous SOTA algorithms with a large margin of PSNR over 1.2dB**.

2. Related Work

Snapshot compressive imaging: In terms of hardware, except capturing high-speed video frames [17, 11, 32], SCI has demonstrated promising results on spectral[12, 65, 28], spectral-temporal[42], polarization[43], and coherent diffraction imaging[3], etc. The underlying principle of these systems is to modulate the high dimensional signals and capture the measurement compressively.

From the software perspective, the reconstruction algorithms can be broadly divided into two categories, *i.e.*, optimization-based and learning-based algorithms. Optimization-based methods utilize various priors [58, 64, 29, 53, 54] during reconstruction. However, the inference time is limited by the iterative solution process. As the development of deep learning, learning-based algorithms achieve impressive success in solving inverse problems. Different network backbones, such as CNN[38, 4] and recurrent neural network (RNN)[5] have been employed for video SCI reconstruction. Though these learning-based methods can achieve more decent results, their reconstruction process lacks interpretability. Combining the merits of above both kind of methods, DUNs[50, 26, 51] have been developed. However, *pixels with high reconstruction variance* have not attracted enough attention.

Uncertainty: Uncertainty has been widely studied to help solve the reliability assessment, regression, risk-based decision making problems[8, 10, 36, 13, 49]. Recently, uncertainty has been introduced into deep learning to improve the robustness and performance of deep neural networks for computer vision tasks such as semantic segmentation[20, 22], image classification[14], object detection[6], etc. For uncertainties in deep learning, they can be roughly classified into model uncertainty capturing the noise of the network's parameters, and data uncertainty referring to the noise inherent in the training data. Ning *et al*. investigated the data uncertainty with estimated mean and variance in low-level vision task such as super-resolution[34], which focuses on the areas with higher variance and achieved better result. **In image or video restoration tasks, high-frequency information is hard to be reconstructed [52] due to the corresponding high reconstruction uncertainty.** To introduce the high-frequency information from the measurement into the network during training process, we first estimate the uncertainty and extract the feature maps of high-frequency information which are finally fed into the unfolding framework as regularized prior.

Transformer for vision: Recently, Transformer has achieved impressive success in the field of natural language processing due to the powerful self-attention mechanism, which inspires numerous researchers to introduce attention mechanism into vision. Many works[46, 55] provide a complementary component (Self-attention/Transformers) to CNNs for modeling long range dependency. Vision Transformer (ViT)[9] and its follow-ups[15, 41, 45, 57, 44] start the trend of that backbone architectures for computer vision shift from CNNs to Transformers. Swin Transformer[30] is a typical representative and the key design is its shift of the window partition between consecutive self-attention layers, which enables it to serve as a general backbone for various tasks. Video Swin Transformer[31] extends the scope of local attention computation from only the spatial domain to the spatiotemporal domain through spatiotemporal adaptation of Swin Transformer. In this paper, our proposed CTM takes both spatiotemporal globally and locally into account by **integrating Transformer and 3D-CNN**, and outperforms all previous SOTA methods.

3. Review the Forward Model of Video SCI

The top-left of Fig. 3 depicts the principle of video SCI, where multiple high-speed frames $\mathbf{X} \in \mathbb{R}^{W\times H\times T}$ are modulated by different masks $\mathbf{M} \in \mathbb{R}^{W\times H\times T}$ and then the measurement $\mathbf{Y} \in \mathbb{R}^{W\times H}$ is captured by a 2D camera, where W, H, and T denote the width, height, and the number of frames, respectively. The 2D measurement is

$$\mathbf{Y} = \textstyle\sum_{t=1}^{T} \mathbf{X}_t \odot \mathbf{M}_t + \mathbf{N}, \tag{1}$$

where $\mathbf{N} \in \mathbb{R}^{WH}$ denotes the measurement noise and $\odot$ represents the Hadamard (element-wise) multiplication. Eq. (1) can be rewritten as the following linear from:

$$y = \Phi x + n, \tag{2}$$

where $x = \mathtt{vec}(\mathbf{X}') \in \mathbb{R}^{WHT}$, $y = \mathtt{vec}(\mathbf{Y}) \in \mathbb{R}^{WH}$, and $n = \mathtt{vec}(\mathbf{N}) \in \mathbb{R}^{WH}$. $\mathtt{vec}()$ here denotes vectorization. The sensing matrix $\Phi \in \mathbb{R}^{WH\times WHT}$ can be expressed as

$$\Phi = [\mathtt{Diag}(\mathtt{vec}(\mathbf{M}_1)), \ldots, \mathtt{Diag}(\mathtt{vec}(\mathbf{M}_t))]. \tag{3}$$

$\mathtt{Diag}()$ here means diagonalizing the vector. Note that Φ is a very sparse matrix and the reconstruction error is bounded when $T > 1$ [21].

4. Proposed Method

DUN Framework: SCI reconstruction is an ill-posed problem which can be modeled as:

$$x = \arg\min_x \|y - \Phi x\|_2^2 + \lambda \psi(x), \tag{4}$$

where $\psi(x)$ denotes the regularization term to confine the solutions, λ balances the two terms. Here we unfold the iterations utilizing the framework of generalized alternating projection (GAP) [27], which solves:

$$\{\hat{x}, \hat{v}\} = \arg\min_x \|x - v\|_2^2 + \lambda \psi(v), \;\; s.t. \;\; y = \Phi x. \tag{5}$$

The solution can be derived by the following two steps:

- Given v, x is updated by the following projection:

$$x^{(j)} = v^{(j-1)} + \Phi^\top (\Phi\Phi^\top)^{-1} (y - \Phi v^{(j-1)}). \tag{6}$$

  Recall Eq. (3), we have $\Phi\Phi^\top = \mathtt{Diag}(R_1, \ldots, R_{WH})$ is a diagonal matrix where $R_i = \sum_{t=1}^{T} \mathbf{M}_{t,i}^2$, $\forall i = 1, \ldots, WH$. Thus Eq. (6) can be efficiently solved.
- Given x, v is achieved by:

$$v^{(j)} = \Theta([x^{(j)}, \Gamma]), \tag{7}$$

  where $v^{(j)}$ denotes the j-th phase's estimate of the target signal, $[\cdot]$ denotes the concatenation, Γ represents other inputs of different phases, and Θ symbolizes the proposed prior module in each phase. To balance the trade-off

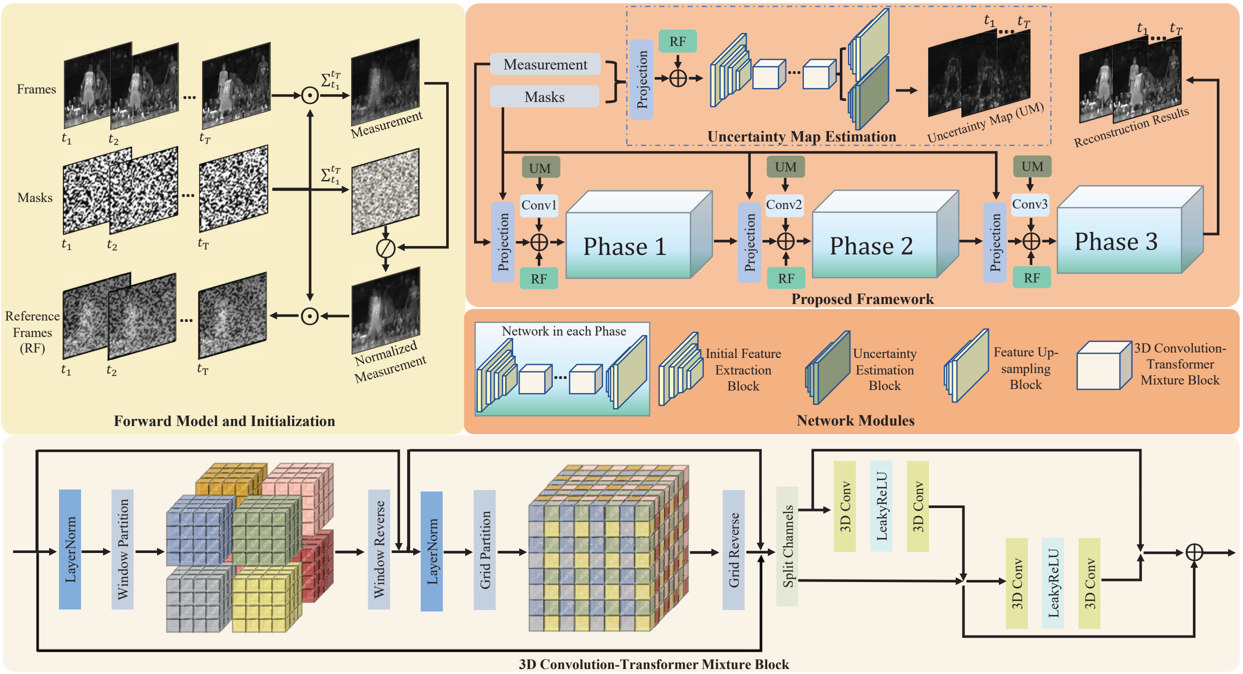

Figure 3. Illustration of video SCI and our proposed model. Top-Left: Sequential video frames are modulated by dynamic masks and then compressed to the measurement. Normalized Measurement is achieved by element-divide the sum of dynamic masks. Reference Frames are acquired by element-wise multiplication. Top-Right: Architecture of our proposed uncertainty guided DUN for video SCI. Bottom: Details of CTM blocks, composed of 3D scalabe blocked local and dilated global attention combining 3D-CNN. $\bigoplus$ here denotes concatenation. More details are in Supplementary Material (SM).

between reconstruction performance and model size, we only utilize 3 phases which will be disscused in Sec.5.3. Unlike conventional optimization-based algorithms utilizing various denoisers, in most unfolding-based algorithms deep networks are used to learn a more appropriate prior to constrain the signal domain. **Differently, we do not just let the network to learn a prior and we further introduce a regularized prior input into our unfolding framework by uncertainty estimation which focuses on the pixels with higher reconstruction uncertainty.** More network structure details in the following sections.

Uncertainty Estimation for SCI: As mentioned above, uncertainty could be roughly classified into model uncertainty capturing the noise of the network's parameters, and data uncertainty referring to the noise inherent in given training data[23]. We investigate the data uncertainty estimation for SCI. Let $f(\cdot)$ denotes the reconstruction algorithm, the data uncertainty can be formed as an additive term σ. In this way, the observation model can be formulated as:

$$x = f(y) + \varepsilon\sigma, \tag{8}$$

where $\varepsilon \sim \mathscr{N}(0,1)$. We assume a Gaussian distribution to characterize the likelihood function:

$$p(x,\sigma|y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\|x-f(y)\|_2^2}{2\sigma^2}\right), \tag{9}$$

the log-likelihood function is naturally represented as:

$$\ln p(x,\sigma|y) = -\frac{\|x-f(y)\|_2^2}{2\sigma^2} - \frac{1}{2}\ln\sigma^2 - \frac{1}{2}\ln 2\pi. \tag{10}$$

As shown in the top-right of Fig. 3, we learn the target estimation (mean value, $f(y)$) and uncertainty (variance, σ^2) respectively by two decoding branches sharing the same encoder. Note that the network $f(\cdot)$ here has the same structure as the network in each phase except the additional decoding branch for uncertainty estimation, we will talk about this in the Sec.5.3. For more stable training, we estimate the log variance $\beta = \ln\sigma^2$ rather than directly estimate σ^2 due to the high dynamic range. Maximizing the likelihood in Eq.(10) is same as minimizing the following loss function for learning the uncertainty (variance) of SCI reconstruction:

$$\mathscr{L}_U = \exp(-\beta)\|x-f(y)\|_2^2 + \beta. \tag{11}$$

The uncertainty estimation results is shown in Fig. 4. To visually highlight the pixel with high variance, we utilize thresholding method for the binarization processing in Fig. 4, and the threshold is the mean of the intensity. **We can observe that pixels with high variance are distributed around the high-frequency details, such as edges and textures.** In previous researches [5, 50, 4], Reference Frames (RF) are utilized in the initialization part for introducing the low-frequency information to improve the

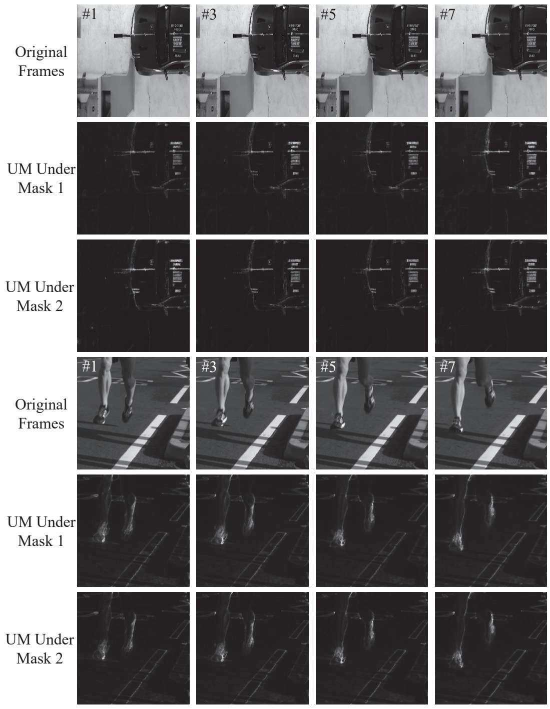

Figure 4. Visualization of the estimated uncertainty of two selected scenes under two different masks.

reconstruction performance. As shown in the top-left of Fig. 3, the Normalized Measurement can achieve more visually clear background and stationary areas *but with blurry edges and textures.* Hence, at the initialization of each phase, we not only focus on the low-frequency information but also take the high-frequency details into consideration. The feature maps being fed into each phase are extracted from the estimated uncertainty map (UM) by three 3D-CNN blocks. The 3D-CNN blocks have the same structure but without sharing parameters. Fig. 4 also shows the uncertainty estimation module's adaptability to different masks. Under different masks, the uncertainty estimation is unaffected.

Convolution-Transformer Mixture: Multi-head self-attention modules (MSA) are widely used in Transformers. Most of traditional MSAs of Transformers for video perform global spatial interactions by utilizing all tokens extracted from the whole feature map, which requires quadratic complexity. Compared to images, videos need to take the correlation of temporal dimension into consideration. Inspired by previous studies [30, 31, 44], we propose a novel attention module. As shown in the bottom of Fig. 3, CTM is composed of three sequential stacked parts, *i.e.*, 3D blocked dense attention (BDA) for local interaction, 3D dilated sparse attention (DSA) for global interaction, and 3D-CNN based feature fusion (FF) module for further exploring spatiotemporal correlations.

Let $\mathbf{X}_f \in \mathbb{R}^{W\times H\times T\times C}$denote input feature map. In BDA, given a 3D window size of $P\times P\times M$, the input tokens are partitioned into $\frac{W}{P}\times\frac{H}{P}\times\frac{T}{M}$ non-overlapping 3D windows. As shown in the bottom of Fig. 3, given an input with the size of $8\times8\times8$ and the 3D window size $4\times4\times4$, we achieve 8 3D windows. And we conduct MSA on each window:

$$\mathrm{MSA}(\mathbf{X}_f) = \mathrm{Softmax}(QK^T/\sqrt{d}+B)V, \tag{12}$$

where Q,K,V denotes the query, key, and value matrix respectively, the number of each head's channels $d=\frac{C}{N}$ and N is the number of heads. B represents the 3D relative position bias. compared to full self-attention (FSA),

$$\Omega(\mathrm{FSA}) = 4\mathrm{WHTC}^2 + 2(\mathrm{WHT})^2\mathrm{C}. \tag{13}$$

BDA allows local spatiotemporal interactions with only a linear complexity,

$$\begin{aligned}\Omega(\mathrm{BDA}) &= [4P^2MC^2 + 2(P^2M)^2C]\tfrac{WHT}{P^2M},\\ &= 4WHTC^2 + 2WHTP^2MC.\end{aligned}$$

For practical applications of SCI, large-scale scenarios are very common. However, local-attention models do not adapt well to large scales[7, 9]. Inspired by [63], we propose 3D DSA for global interaction. Unlike BDA where the input tokens are partitioned into non-overlapping 3D windows, in DSA, to keep the fixed group size of $S\times S\times B$, the tokens are selected from the sparse positions with the interval of $\frac{W}{S}\times\frac{H}{S}\times\frac{T}{B}$. As shown in the bottom of Fig. 3, given an input with the size of $8\times8\times8$ and the interval size of $2\times2\times2$, we achieve 8 groups with the size of $4\times4\times4$ and employ MSA as well. Note that the complexity of DSA for global interaction is also linear,

$$\begin{aligned}\Omega(\mathrm{DSA}) &= [4S^2BC^2 + 2(S^2B)^2C]\tfrac{WHT}{S^2B}\\ &= 4WHTC^2 + 2WHTS^2BC.\end{aligned}$$

Recall video Swin[31] where the mechanism of 3D shifted windows is employed to bridge the connections across different windows, our proposed 3D local and global attention achieves this in a more implementation friendly way and is scalable.

We propose an initialization feature extraction block at the beginning of each phase to increase the generalization and trainability of the network. In each CTM block, to further explore the correlation of spatiotemporal dimensions, we plug FF into each CTM block. In FF, the feature map is first divided into two parts according to the channels. Then the two parts with skip connection are respectively sent into two Resnet modules with the same structure but not sharing parameters. Finally the features are fused to keep the original dimensions. We utilize 3D-CNN for all the convolutional layers.

Training: Prior to the training of uncertainty estimation, we first train the whole network without uncertainty estimation to ensure the convergence. Given the training pairs $(y_i, x_i)_{i=1}^N$, where N is training data number (52000 cropped pairs used here), the mean square error (MSE) loss is selected as the loss function. After 20 epoch training, $\mathscr{L}_U$ loss function is utilized to estimate the uncertainty. The initial learning rate is $5e^{-5}$ for the first 10 epochs and decays to $1e^{-5}$ for the last 10 epochs. After the training of UM estimation, we fix all the parameters of uncertainty estimation network and train the proposed framework with the initialization of corresponding parameters from the same network modules, *i.e.*, duplicating the corresponding parameters from the uncertainty estimation network to each phase, which will lead to faster convergence of training.

The network is trained on 2 NVIDIA A40 GPUs utilizing PyTorch [35]. Adam [24] is employed as the optimizer. Note that, for the training of uncertainty estimation, if we directly use $\mathscr{L}_U$ for training, the training is easy to diverge. Therefore, we used MSE loss for training at the first. $S \times S \times B$ and $P \times P \times M$ we set in the experiments are the same, *i.e.*, $7 \times 7 \times 2$. The setting of the spatial parameters, *i.e.*, P and S, follows Swin Transformer[30]. And the chosen number of B and M, *i.e.*, two, echoes the two divided parts in the FF module.

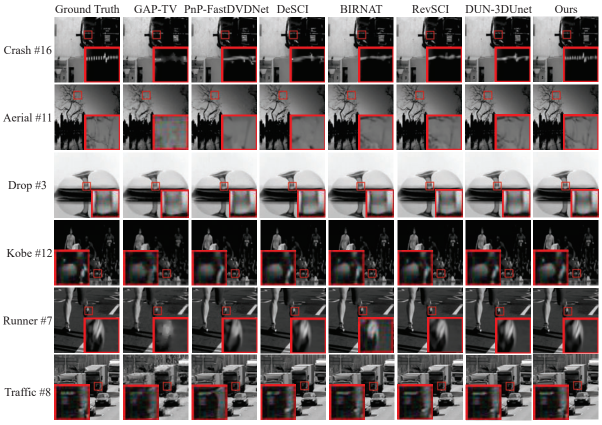

Figure 5. Selected multiple reconstruction frames of simulated benchmark dataset.

5. Experiments

Dataset: We choose **DAVIS 2017** [37] as our training dataset following previous studies. It contains 90 scenes with two resolutions: 480P and 1080P. We conduct data augmentation by random cropping, rotation and flip.

5.1. Benchmark Simulation of SCI

The testing synthetic datasets of Benchmark follow previous study [29] including **Kobe, Traffic, Runner, Drop, Crash** and **Aerial** with the size of $256 \times 256 \times 8$. We compare our model with previous SOTA algorithms, *i.e.*, GAP-TV [58], PnP-FFDnet [60], PnP-FastDVDnet [61], DeSCI [29], E2E-CNN [38], GAP-Unet-S12 [33], BIRNAT [5], MetaSCI [48], RevSCI [4] and DUN-3DUnet [50]. The quantitative comparison is summarized in Tab. 1. Both PSNR and structured similarity (SSIM) [47] are selected to evaluate the reconstruction quality. It can be observed that our method substantially outperforms (by a large margin of nearly 1.2dB in PSNR) all previous SOTA algorithms. Selected reconstructed frames are shown in Fig. 5. As we can see, the optimization-based algorithms, such as GAP-TV and PnP, usually lead to oversmooth (Crash,Kobe,Runner, and Traffic) artifacts. DeSCI is with poor restoration of the irregular textures (Aerial). When the object is with large motion, other learning-based methods do not work well. Obviously, our proposed method achieves much better visually results on the areas with high uncertainty (variance), such as the edges, textures, and other high-frequency details. The inference time is on par with previous SOTA DUN-3DUnet.

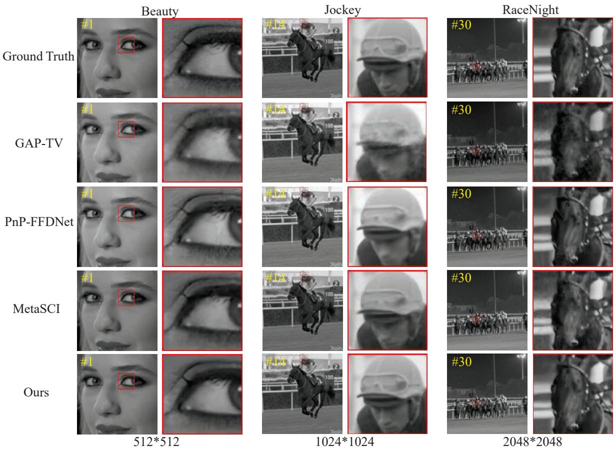

Figure 6. Selected reconstruction frames of different scales. Zoom in for better view.

Adaptability: We test our uncertainty estimation module under different masks. As shown in Fig. 4, it can well adapt to different masks. We further test the adaptability of the reconstruction, the results are presented in Tab. 3. Note that all the experiments are directly conducted without training with other masks, which is never achieved by previous learning-based methods. (Other methods' results are in SM.)

5.2. Scalability of Transformer on Large-scale Data

As mentioned in the preceding part of the paper, the ability to cope with large-scale data is crucial for reconstruction algorithms. Our proposed scalable Transform module (BDA and DSA) facilitates the practical applications of SCI. We test the proposed model on the large-scale benchmark dataset [48]. The quantitative comparison is summarized in Tab. 2. As we can see, few algorithms can be applied to large scale data due to GPU memory limit while

Dataset	Kobe	Traffic	Runner	Drop	Aerial	Crash	Average	Running time
GAP-TV [58]	26.45 0.845	20.90 0.715	28.48 0.899	33.81 0.963	25.03 0.828	24.82 0.838	26.58 0.848	4.2
E2E-CNN [38]	27.79 0.807	24.62 0.840	34.12 0.947	36.56 0.949	27.18 0.869	26.43 0.882	29.45 0.882	0.0312
DeSCI [29]	33.25 0.952	28.72 0.925	38.76 0.969	43.22 0.993	25.33 0.860	27.04 0.909	32.72 0.935	6180
PnP-FFDNet [60]	30.47 0.926	24.08 0.833	32.88 0.938	40.87 0.988	24.02 0.814	24.32 0.836	29.44 0.889	3.0
PnP-FastDVDNet [61]	32.73 0.946	27.95 0.932	36.29 0.962	41.82 0.989	27.98 0.897	27.32 0.925	32.35 0.942	18
BIRNAT [5]	32.71 0.950	29.33 0.942	38.70 0.976	42.28 0.992	28.99 0.927	27.84 0.927	33.31 0.951	0.16
GAP-Unet-S12 [33]	32.09 0.944	28.19 0.929	38.12 0.975	42.02 0.992	28.88 0.914	27.83 0.931	32.86 0.947	0.0072
MetaSCI [48]	30.12 0.907	26.95 0.888	37.02 0.967	40.61 0.985	28.31 0.904	27.33 0.906	31.72 0.926	0.025
RevSCI [4]	33.72 0.957	30.02 0.949	39.40 0.977	42.93 0.992	29.35 0.924	28.12 0.937	33.92 0.956	0.19
DUN-3DUnet [50]	35.00 0.969	31.76 0.966	40.90 0.983	44.46 0.994	30.46 0.943	29.35 0.955	35.32 0.968	1.35
Ours	35.77 0.984	32.40 0.979	41.82 0.993	45.25 0.996	31.41 0.968	31.08 0.978	36.29 0.983	1.26
Ours-with-Uncertainty	**35.97 0.986**	**32.59 0.981**	**42.10 0.995**	**45.49 0.998**	**31.64 0.970**	**31.33 0.980**	**36.52 0.985**	1.58

Table 1. The quantitative comparison of different algorithms. The average results of PSNR in dB (left entry), SSIM (right entry) and running time per measurement in seconds. Note that GAP-TV and DeSCI are running on CPU while others are on GPU. The best results are bold, and the second best are underlined. Full results are in SM.

Size	Algorithm	Beauty	Bosphorus	HoneyBee	Jockey	ShakeNDry	Average	Running Time
512x512	GAP-TV [58]	32.13 0.857	29.18 0.934	31.40 0.887	31.01 0.940	32.52 0.882	31.25 0.900	44.67
	PnP-FFDNet [60]	30.70 0.855	35.36 0.952	31.94 0.872	34.88 0.955	30.72 0.875	32.72 0.902	14.22
	MetaSCI [48]	35.10 0.901	38.37 0.950	34.27 0.913	36.45 0.962	33.16 0.901	35.47 0.925	0.12
	Ours	41.22 0.983	42.39 0.990	43.63 0.990	41.81 0.988	37.09 0.966	41.23 0.983	4.97
	Ours-with-Uncertainty	**41.36 0.984**	**42.59 0.990**	**43.71 0.991**	**42.10 0.989**	**37.40 0.966**	**41.41 0.984**	6.32
Size	Algorithm	Beauty	Jockey	ShakeNDry	ReadyGo	YachtRide	Average	Test Time
1024x1024	GAP-TV [58]	33.59 0.852	33,27 0.971	33.86 0.913	27.49 0.948	24.39 0.937	30.52 0.924	178.11
	PnP-FFDNet [60]	32.36 0.857	35.25 0.976	32.21 0.902	31.87 0.965	30.77 0.967	32.49 0.933	52.47
	MetaSCI [48]	35.23 0.929	37.15 0.978	36.06 0.939	33,34 0.973	32.68 0.955	34.89 0.955	0.59
	Ours	40.11 0.978	42.28 0.988	38.95 0.978	40.39 0.989	37.76 0.982	39.90 0.983	23.76
	Ours-with-Uncertainty	**40.40 0.979**	**42.46 0.990**	**39.22 0.979**	**40.60 0.989**	**37.96 0.983**	**40.13 0.984**	31.78
Size	Algorithm	City	Kids	Lips	RaceNight	RiverBank	Average	Test Time
2048x2048	GAP-TV [58]	21.27 0.902	26.05 0.956	26.46 0.890	26.81 0.875	27.74 0.848	25.67 0.894	764.75
	PnP-FFDNet [60]	29.31 0.926	30.01 0.966	27.99 0.902	31.18 0.891	30.38 0.888	29.77 0.915	205.62
	MetaSCI [48]	32.63 0.930	32.31 0.965	30.90 0.895	33.86 0.893	32.77 0.902	32.49 0.917	2.38
	Ours	40.31 0.981	40.22 0.984	35.26 0.933	36.36 0.924	36.87 0.970	37.81 0.964	95.06
	Ours-with-Uncertainty	**40.54 0.983**	**40.45 0.985**	**35.49 0.934**	**36.59 0.956**	**37.10 0.971**	**38.04 0.966**	120.09

Table 2. Large-scale results (CR: 8): quantitative comparison of existing algorithms that can be applied to large-scale data. The best results are in bold, and the second best results are underlined. PSNR and SSIM are selected as the evaluation metrics.

Evaluation metrics	Trained mask	New mask 1	New mask 2
PSNR SSIM	36.52 0.985	36.47 0.985	36.48 0.985

Table 3. Quantitative comparison with different masks.

training, our proposed method far exceeds (nearly 6dB in PSNR) all previous SOTA algorithms with competitive inference time, which verifies our proposed Transformer module is with enough scalability to large-scale data. Details of selected reconstruction frames of different scales are shown in Fig. 6. It can be observed that we can achieve much better visual performance especially in the details.

5.3. Ablation Study

Effectiveness of modules: To validate the effectiveness of each part of our proposed CTM module, we conduct ablation experiments on the benchmark dataset for each submodules, *i.e.*, BDA, DSA and FF. To reduce the effects of uncontrollable factors on the experiments, the above ablation experiments are conducted without uncertainty estimation with quantitative result shown in Tab. 4, where ✓ denotes the corresponding components are preserved, × is on the contrary. As we can observe, each of the modules is essential for the whole framework.

As described in the above, to test the efficiency of each module, we directly remove each part of the module separately. However, **we should not ignore the effect brought by the reduction of parameter count.** In order to measure the effectiveness of the Transformer module more accurately, we conduct experiments utilizing BDA to replace DSA and utilizing DSA to replace DSA respectively, which all maintain the same parameter count and FLOPs. Block attention mechanism' efficiency has been verified in many other computer vision tasks [30]. However when we use BDA to replace DSA, PSNR decreases by 0.33dB on the benchmark dataset ($256 * 256 * 8$). When we use DSA to replace BDA, PSNR decreases by 0.87 dB on the benchmark dataset ($256 * 256 * 8$). The results demonstrate that local attention plays a more important role, yet the combination of both local and global attention leads to higher performance. We also test **different order of the sub-modules,** *i.e.*, BDA, DSA and FF in CTM block. Because the blocks are sequentially arranged, the change of the order of the submodules does not affect the performance.

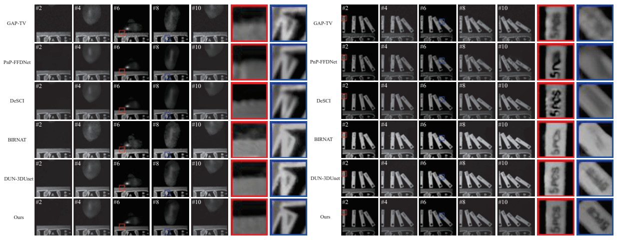

Figure 7. Selected reconstruction frames of real data **Water Balloon** and **Domino**. More results are in SM.

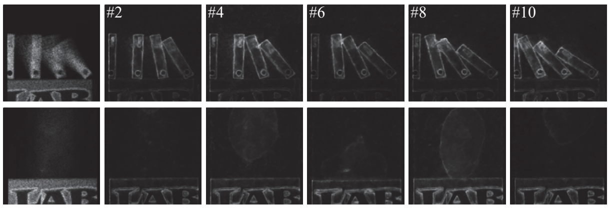

Figure 8. Selected estimated uncertainty map of real data **Water Balloon** and **Domino**.

BDA	DSA	FF	PSNR SSIM
×	✓	✓	31.11 0.960
✓	×	✓	31.04 0.955
✓	✓	×	27.81 0.912
✓	✓	✓	36.29 0.983

Table 4. Ablation study of CTM on benchmark dataset. The quantitative effects (PSNR in dB and SSIM) are shown.

Effectiveness of uncertainty estimation: We also verify the effect of uncertainty estimation. Directly applying $\mathscr{L}_U$ loss for reconstruction brings a direct drop of nearly 1.1dB in PSNR. Recall Eq. (11), the attention of pixels with high variance will be impaired by the division. Though paying less attention to pixels with high variance (uncertainty) helps promote performance in high level vision tasks[23, 1, 2], it does not work in low level vision tasks. As shown in Tab. 1 and Tab. 2 , when the high-frequency information is not introduced, the PSNR and SSIM decline nearly 0.2-0.3dB and 0.002-0.003, respectively. Although we can further improve the reconstruction performance, the inference speed of the model is sacrificed. There is a trade-off between the benefits of reconstruction quality and the sacrifice of the inference speed. As mentioned above, we utilize the network of only one-phase instead of three-phases to estimate the uncertainty, which can reduce the inference time. We conducted experiments with the uncertainty map estimated by two different network, *i.e.*, one-phase and three-phases networks, the reconstruction quality is almost the same. Obviously, one-phase uncertainty estimation has higher inference speed. Considering the memory cost, the phase number we chose is three in this paper. **The three-phases inference model with uncertainty estimation is basically with the same number of parameters as four-phases model.** Hence we test different phase numbers, *i.e.*, 1, 2, 3 and 4, under our proposed framework without uncertainty estimation utilizing the same benchmark dataset $(256 * 256 * 8)$. As shown in Tab. 5, as the phase number increases, the reconstruction quality improvement is slowing down. Compared with three-phases, four-phases model only gain an increase of less than 0.1dB in PSNR, which is why we only use 3 phases and also illustrates the effectiveness of introducing uncertainty estimation.

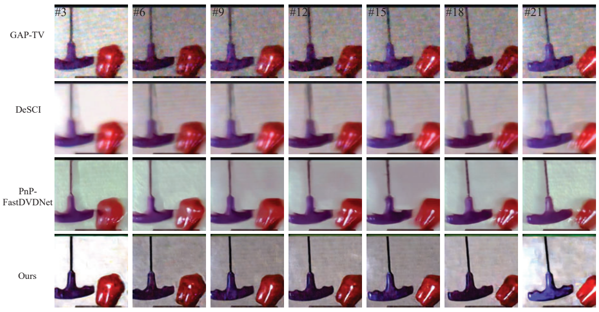

Figure 9. Comparison of selected reconstruction video frames of real color data **Hammer**.

Phase Number	One	Two	Three	Four	Three with Uncertainty
PSNR SSIM	35.51 0.970	36.12 0.981	36.29 0.983	36.37 0.983	36.52 0.985

Table 5. Reconstruction with different phase numbers.

5.4. Real Data Benchmark

We test our model on the real data **Water Balloon** and **Domino** with the size of 512 × 512 × 10 [38]. Due to the uncontrollable noise during capturing, it is more challenging to reconstruct real measurements. Note that *we do not add any noise to the training data during the training with real masks, which demonstrates the generalization ability of our model to a certain extent*. The selected results are presented in Fig. 7. In the areas with higher uncertainty (variance), such as edges and textures, our proposed method outperforms all existing algorithms, which is shown in the left part of Fig. 7. Even when the water balloon collides with box, the edge of the box is still sharp in our reconstruction. Besides, falling dominoes are with higher speed, which further increases the difficulty of reconstruction. As we can observe in right part of Fig. 7, all previous SOTA algorithms can not recover the legible letters except our proposed method. Our results are with sharper edges, more details, and cleaner background, which indicates our proposed method is more powerful in practical applications. The estimated uncertainty maps of real data are shown in Fig. 8, the edge and texture features can be directly obtained from the real measurement.

We also test our model on real color dataset **Hammer** with the size of 512 × 512 × 22. Few learning based algorithms conducted experiments on the color video SCI task. We compare our model with previous SOTA algorithms which are iteration-based. As we can see in Fig. 9, GAP-TV has noisy results, DeSCI and PnP-FastDVDNet are blurry in the areas of background and edges, our results are cleaner and have sharp edges than other methods. The implementation details are in the SM.

6. Conclusions and Future Work

We have proposed a Transformer and 3D-CNN based network for video SCI reconstruction and introduced high-frequency information by uncertainty estimation. The design of the backbone with Transformer and 3D-CNN helps explore the correlation across the spatio-temporal dimensions. More importantly, our proposed method achieved SOTA results with a competitive inference time.

Although we have achieved the best results so far, the introduction of high-frequency information is time-inefficient and when the model is applied to large-scale data, the inference time is still long for real-time applications. In the future, we will reduce the parameters for high inference speed by knowledge distillation and employ the high-frequency information in a more time-efficient way. Besides video, our proposed framework can also be used in other inverse problems such as image compressive sensing and spectral compressive imaging.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 8

[2] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5710–5719, 2020. 8

[3] Ziyang Chen, Siming Zheng, Zhishen Tong, and Xin Yuan. Physics-driven deep learning enables temporal compressive coherent diffraction imaging. *Optica*, 9(6):677–680, Jun 2022. 1, 2

[4] Ziheng Cheng, Bo Chen, Guanliang Liu, Hao Zhang, Ruiying Lu, Zhengjue Wang, and Xin Yuan. Memory-efficient network for large-scale video compressive sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16246–16255, 2021. 2, 3, 4, 6, 7

[5] Ziheng Cheng, Ruiying Lu, Zhengjue Wang, Hao Zhang, Bo Chen, Ziyi Meng, and Xin Yuan. Birnat: Bidirectional recurrent neural networks with adversarial training for video snapshot compressive imaging. In *European Conference on Computer Vision*, pages 258–275. Springer, 2020. 2, 3, 4, 6, 7

[6] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee. Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 502–511, 2019. 3

[7] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021. 5

[8] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009. 3

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 5

[10] Michael Havbro Faber. On the treatment of uncertainties and probabilities in engineering decision analysis. 2005. 3

[11] Liang Gao, Jinyang Liang, Chiye Li, and Lihong V Wang. Single-shot compressed ultrafast photography at one hundred billion frames per second. *Nature*, 516(7529):74–77, 2014. 2

[12] M. E. Gehm, R. John, D. J. Brady, R. M. Willett, and T. J. Schulz. Single-shot compressive spectral imaging with a dual-disperser architecture. *Opt. Express*, 15(21):14013–14027, Oct 2007. 2

[13] Paul Goldberg, Christopher Williams, and Christopher Bishop. Regression with input-dependent noise: A gaussian process treatment. *Advances in neural information processing systems*, 10, 1997. 3

[14] Yingjie Gu, Zhong Jin, and Steve C Chiu. Active learning combining uncertainty and diversity for multi-class image classification. *IET Computer Vision*, 9(3):400–407, 2015. 3

[15] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[17] Yasunobu Hitomi, Jinwei Gu, Mohit Gupta, Tomoo Mitsunaga, and Shree K Nayar. Video from a single coded exposure photograph using a learned over-complete dictionary. In *2011 International Conference on Computer Vision*, pages 287–294. IEEE, 2011. 1, 2

[18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[19] Michael Iliadis, Leonidas Spinoulas, and Aggelos K Katsaggelos. Deepbinarymask: Learning a binary mask for video compressive sensing. *Digital Signal Processing*, 96:102591, 2020. 2

[20] Shuya Isobe and Shuichi Arai. Deep convolutional encoder-decoder network with model uncertainty for semantic segmentation. In *2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 365–370. IEEE, 2017. 3

[21] Shirin Jalali and Xin Yuan. Snapshot compressed sensing: Performance bounds and algorithms. *IEEE Transactions on Information Theory*, 65(12):8005–8024, 2019. 3

[22] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015. 3

[23] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 4, 8

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[25] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[26] Yuqi Li, Miao Qi, Rahul Gulve, Mian Wei, Roman Genov, Kiriakos N Kutulakos, and Wolfgang Heidrich. End-to-end video compressive sensing using anderson-accelerated unrolled networks. In *2020 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2020. 2, 3

[27] X. Liao, H. Li, and L. Carin. Generalized alternating projection for weighted-$\ell_{2,1}$ minimization with applications to model-based compressive sensing. *SIAM Journal on Imaging Sciences*, 7(2):797–823, 2014. 3

[28] Xing Lin, Yebin Liu, Jiamin Wu, and Qionghai Dai. Spatial-spectral encoded compressive hyperspectral imaging. *ACM Transactions on Graphics (TOG)*, 33(6):1–11, 2014. 2

[29] Yang Liu and etc. Rank minimization for snapshot compressive imaging. *IEEE TPAMI*. 2, 6, 7

[30] Ze Liu, Lin, and etc. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 3, 5, 6, 7

[31] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 3, 5

[32] Patrick Llull, Xuejun Liao, Xin Yuan, Jianbo Yang, David Kittle, Lawrence Carin, Guillermo Sapiro, and David J. Brady. Coded aperture compressive temporal imaging. *Opt. Express*, 21(9):10526–10545, May 2013. 1, 2

[33] Ziyi Meng, Shirin Jalali, and Xin Yuan. Gap-net for snapshot compressive imaging. *arXiv preprint arXiv:2012.08364*, 2020. 6, 7

[34] Qian Ning, Weisheng Dong, Xin Li, Jinjian Wu, and Guangming Shi. Uncertainty-driven loss for single image super-resolution. *Advances in Neural Information Processing Systems*, 34:16398–16409, 2021. 3

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 6

[36] M Elisabeth Paté-Cornell. Uncertainties in risk analysis: Six levels of treatment. *Reliability Engineering & System Safety*, 54(2-3):95–111, 1996. 3

[37] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 6

[38] Mu Qiao, Ziyi Meng, Jiawei Ma, and Xin Yuan. Deep learning for video compressive sensing. *APL Photonics*, 5(3):030801, 2020. 3, 6, 7, 8

[39] D. Reddy, A. Veeraraghavan, and R. Chellappa. P2c2: Programmable pixel compressive camera for high speed imaging. In *CVPR 2011*, pages 329–336, June 2011. 1

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[41] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 3

[42] Tsung-Han Tsai, Patrick Llull, Xin Yuan, Lawrence Carin, and David J Brady. Spectral-temporal compressive imaging. *Optics letters*, 40(17):4054–4057, 2015. 2

[43] Tsung-Han Tsai, Xin Yuan, and David J Brady. Spatial light modulator based color polarization imaging. *Optics express*, 23(9):11912–11926, 2015. 2

[44] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. *arXiv preprint arXiv:2204.01697*, 2022. 3, 5

[45] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 3

[46] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 3

[47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[48] Zhengjue Wang, Hao Zhang, Ziheng Cheng, Bo Chen, and Xin Yuan. Metasci: Scalable and adaptive reconstruction for video compressive sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2083–2092, 2021. 6, 7

[49] WA Wright. Bayesian approach to neural-network modeling with input uncertainty. *IEEE Transactions on Neural Networks*, 10(6):1261–1270, 1999. 3

[50] Zhuoyuan Wu and etc. Dense deep unfolding network with 3d-cnn prior for snapshot compressive imaging. In *ICCV 2021*. 2, 3, 4, 6, 7

[51] Zhuoyuan Wu, Zhenyu Zhang, Jiechong Song, and Man Zhang. Spatial-temporal synergic prior driven unfolding network for snapshot compressive imaging. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021. 2, 3

[52] Ke Xu, Xin Yang, Baocai Yin, and Rynson W.H. Lau. Learning to restore low-light images via decomposition-and-enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[53] 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 Transaction on Image Processing*, 24(1):106–119, January 2015. 2

[54] J. Yang, X. Yuan, X. Liao, P. Llull, G. Sapiro, D. J. Brady, and L. Carin. Video compressive sensing using Gaussian mixture models. *IEEE Transaction on Image Processing*, 23(11):4863–4878, November 2014. 2

[55] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. In *European Conference on Computer Vision*, pages 191–207. Springer, 2020. 3

[56] Michitaka Yoshida, Akihiko Torii, Masatoshi Okutomi, Kenta Endo, Yukinobu Sugiyama, Rin-ichiro Taniguchi, and Hajime Nagahara. Joint optimization for compressive video sensing and reconstruction under hardware constraints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 634–649, 2018. 2

[57] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021. 3

[58] Xin Yuan. Generalized alternating projection based total variation minimization for compressive sensing. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2539–2543. IEEE, 2016. 2, 6, 7

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

[60] Xin Yuan and etc. Plug-and-play algorithms for large-scale snapshot compressive imaging. In *IEEE CVPR 2020*. 2, 6, 7

[61] Xin Yuan, Patrick Llull, Xuejun Liao, Jianbo Yang, David J. Brady, Guillermo Sapiro, and Lawrence Carin. Low-cost compressive sensing for color video and depth. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3318–3325, 2014. 1, 6, 7

[62] Xuanyu Zhang, Yongbing Zhang, Ruiqin Xiong, Qilin Sun, and Jian Zhang. Herosnet: Hyperspectral explicable reconstruction and optimal sampling deep network for snapshot compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17532–17541, 2022. 2

[63] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018. 5

[64] Chen Zhao, Siwei Ma, Jian Zhang, Ruiqin Xiong, and Wen Gao. Video compressive sensing reconstruction via reweighted residual sparsity. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(6):1182–1195, 2016. 2

[65] Siming Zheng, Yang Liu, Ziyi Meng, Mu Qiao, Zhishen Tong, Xiaoyu Yang, Shensheng Han, and Xin Yuan. Deep plug-and-play priors for spectral snapshot compressive imaging. *Photonics Research*, 9(2):B18–B29, 2021. 2

[66] Siming Zheng, Chunyang Wang, Xin Yuan, and Huolin L Xin. Super-compression of large electron microscopy time series by deep compressive sensing learning. *Patterns*, 2(7):100292, 2021. 1