Effective Snapshot Compressive-spectral Imaging via Deep Denoising and Total Variation Priors

Yao Wang

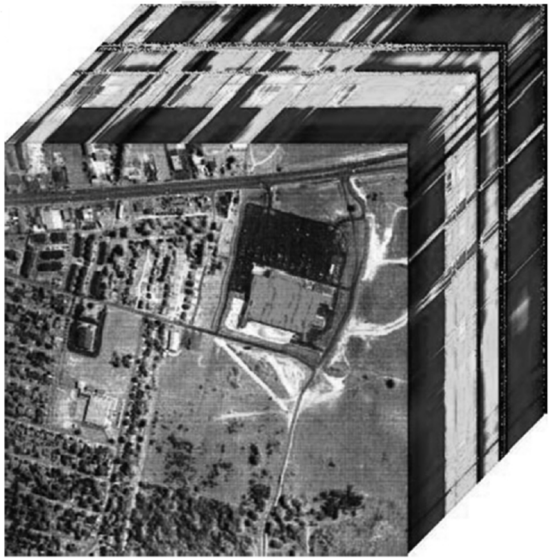
Center for Intelligent Decision-making and Machine Learning

Xi’an Jiaotong University

April 2022

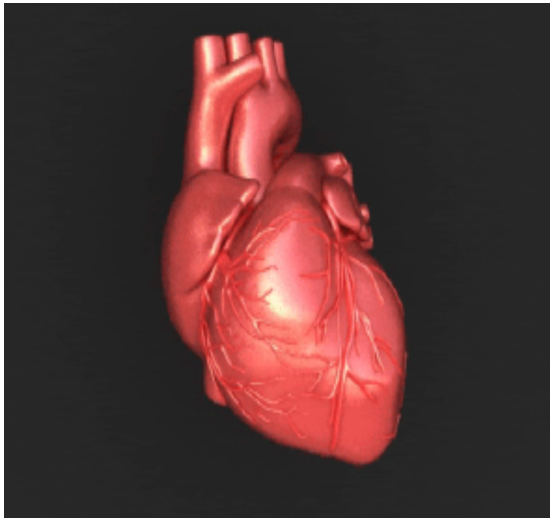
Joint work with Haiquan Qiu and Prof. Deyu Meng

High-dimensional Image Data Everywhere



Hyperspectral

Surveillance



MRI

How to efficiently extract ***compact information*** *from such complex data?*

High-dimensional Sparse Modeling

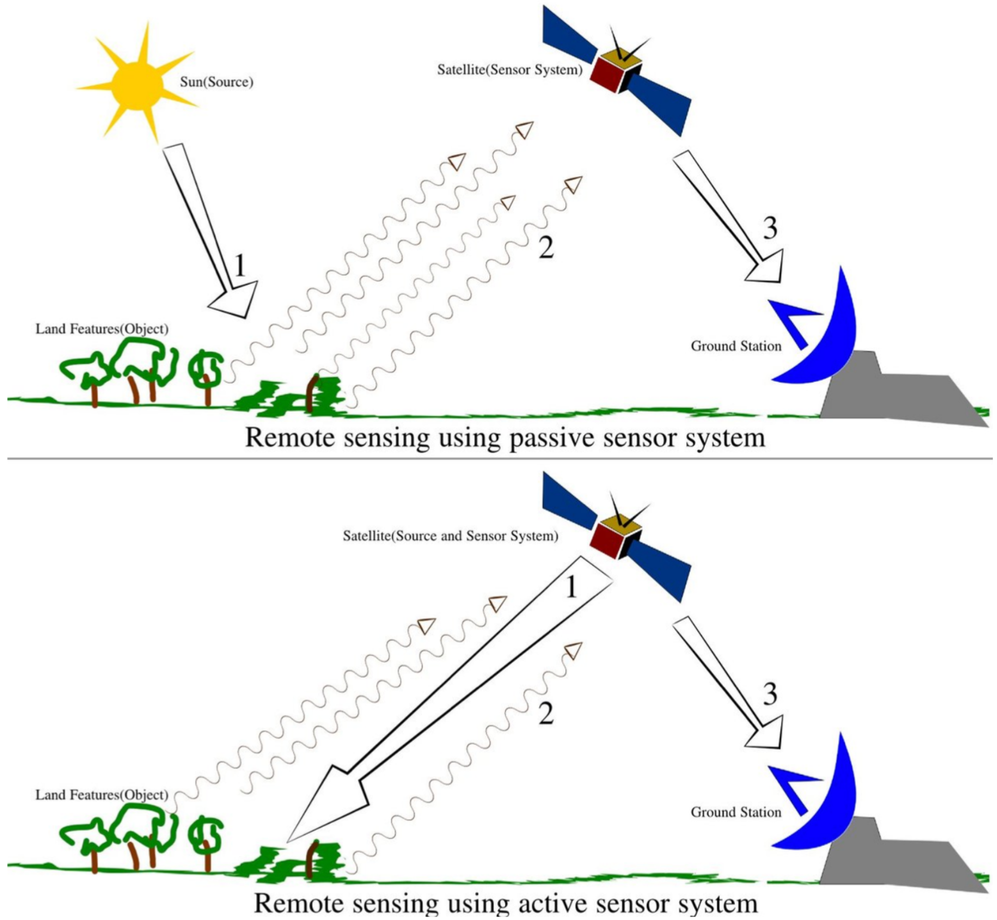
- Applying the *high-dimensional* sparse model

$$\min_{\mathcal{X}} \; S(\mathcal{X})$$
$$s.t. \; \mathbf{y} = \phi(\mathcal{X}) + \mathbf{e}$$

- Main issues
  - Sampling mechanism
  - Structure measure
  - Recovery algorithm
  - Performance assessment
  - Applications

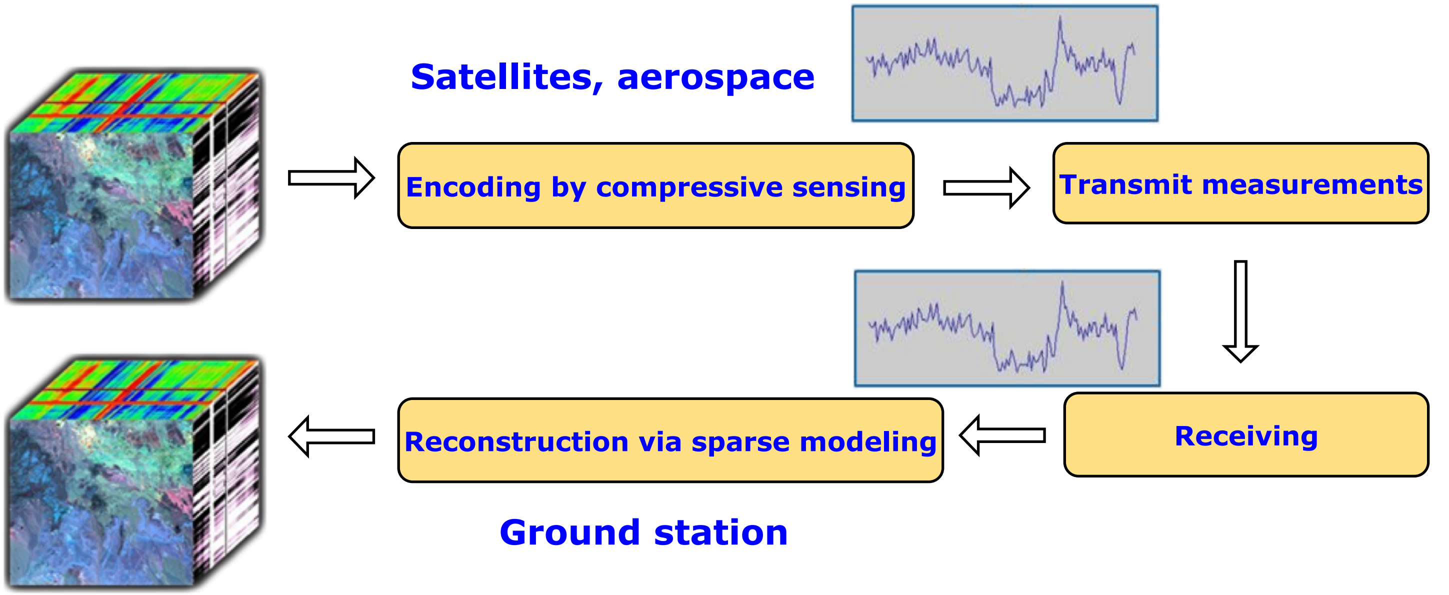
Cao, et al., TIP'16; Wang, et al., NSR' 18; Wang, et al., TGRS'20; Peng, et al., TIP'20; Hou, et al., TPAMI'21

What is Remote Sensing?



The term "remote sensing" generally refers to the use of satellite- or aircraft-based sensor technologies to detect and classify objects on Earth, including on the surface and in the atmosphere and oceans, based on propagated signals.

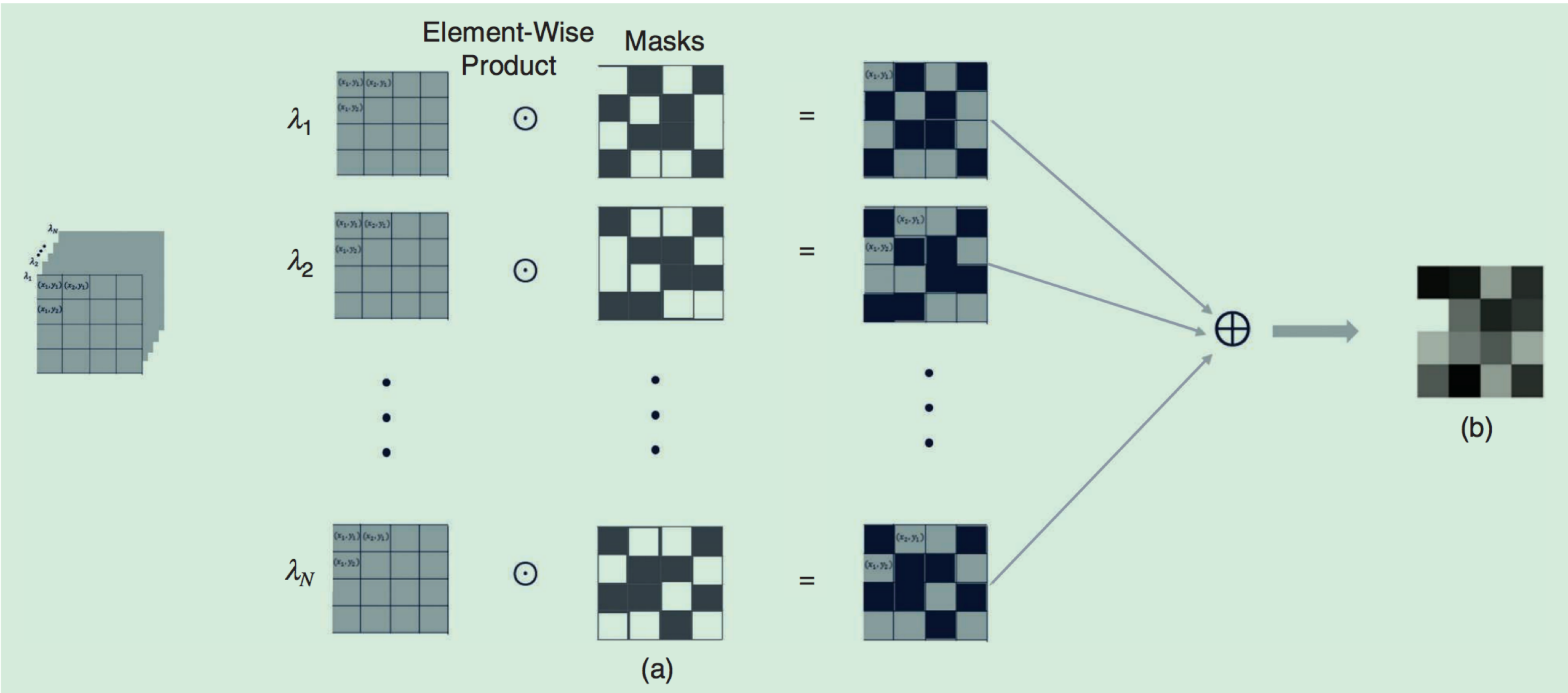
Compressive-spectral Imaging



$$\min_{\mathcal{X}} \; S(\mathcal{X})$$
$$s.t. \; \mathbf{y} = \phi(\mathcal{X}) + \mathbf{e}$$

Spectral image: a *three-dimensional* (x,y,λ) data cube, where x and y represent two spatial dimensions of the scene, and λ represent the spectral of dimension.

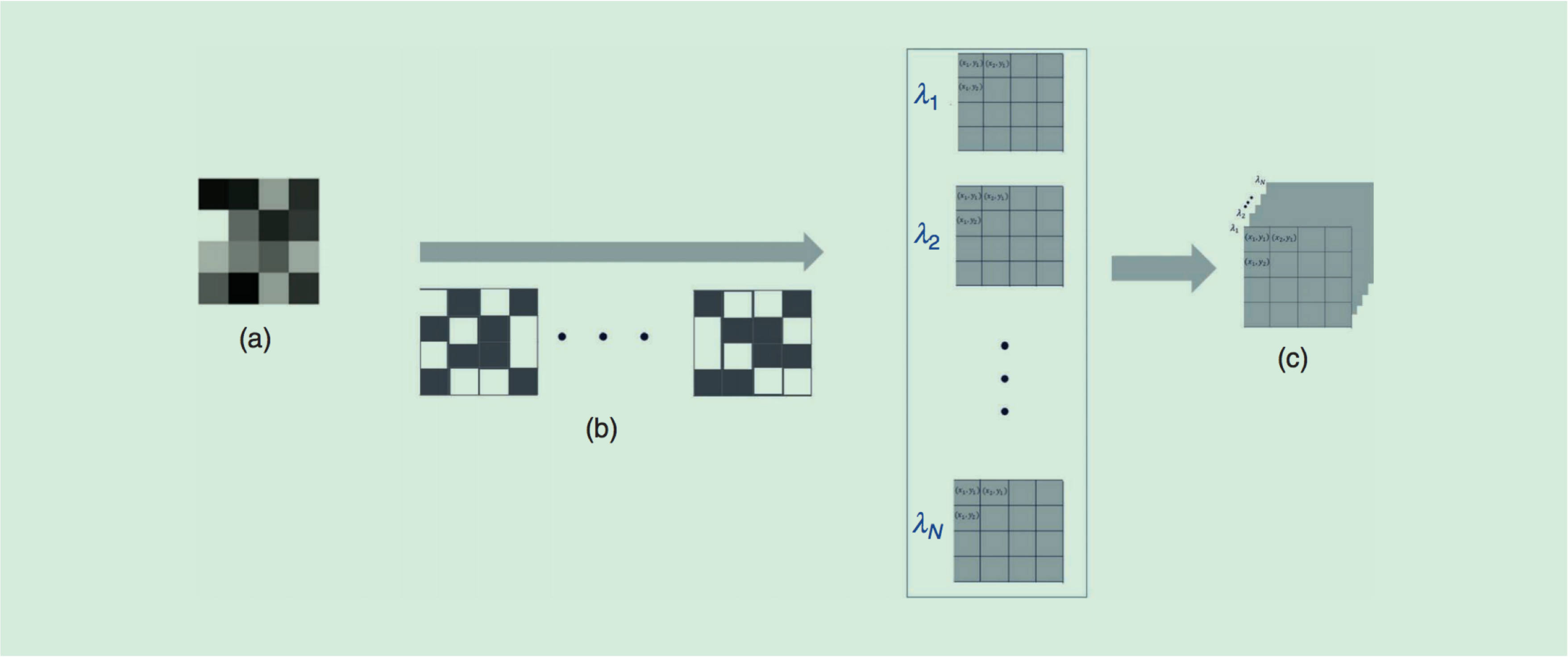
Snapshot Compressive Imaging



The encoding process of SCI

Yuan, Brady and Katsaggelos, IEEE SPM'21

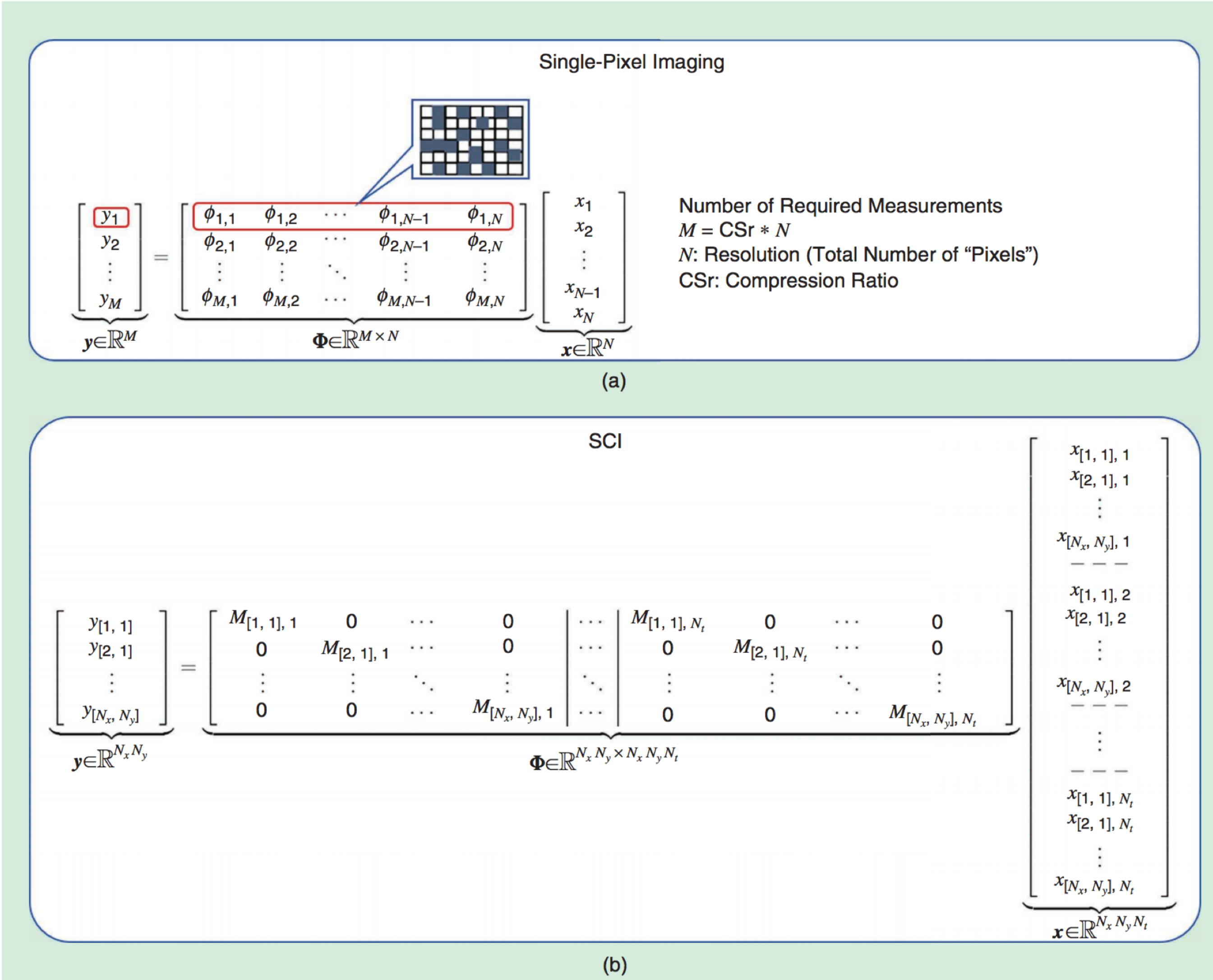
Snapshot Compressive Imaging



The decoding process of SCI

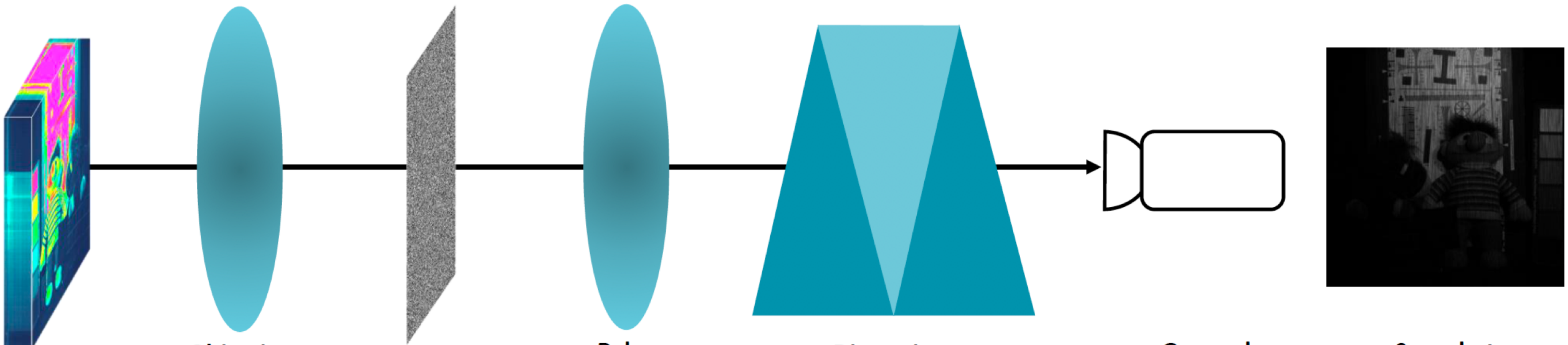
Yuan, Brady and Katsaggelos, IEEE SPM'21

Single-Pixel Imaging vs SCI



The sensing ratio of SCI is $1/N_t$ (in single-pixel imaging, it is M/N)

CASSI System



The sensing process of CASSI (coded aperture compressive spectral imager)

Brady, et al., OE'09; Lin, et al., ACM TOG'14; Yuan, et al., IEEE JSTSP'15

Plug-and-Play Algorithms

- The mathematical model:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2 + \lambda g(\mathbf{x})$$

- PnP-ADMM: *(Ryu, et al., ICML'18; Chan, et al., IEEE TCI'17)*

$$\mathbf{x}^{(k+1)} = (\mathbf{H}^T\mathbf{H} + \gamma\mathbf{I})^{-1}[\mathbf{H}^T\mathbf{y} + \gamma(\mathbf{v}^{(k)} + \mathbf{u}^{(k)})]$$
$$\mathbf{v}^{(k+1)} = \mathcal{D}_\sigma(\mathbf{x}^{(k+1)} - \mathbf{u}^{(k)})$$
$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + (\mathbf{v}^{(k+1)} - \mathbf{x}^{(k+1)})$$

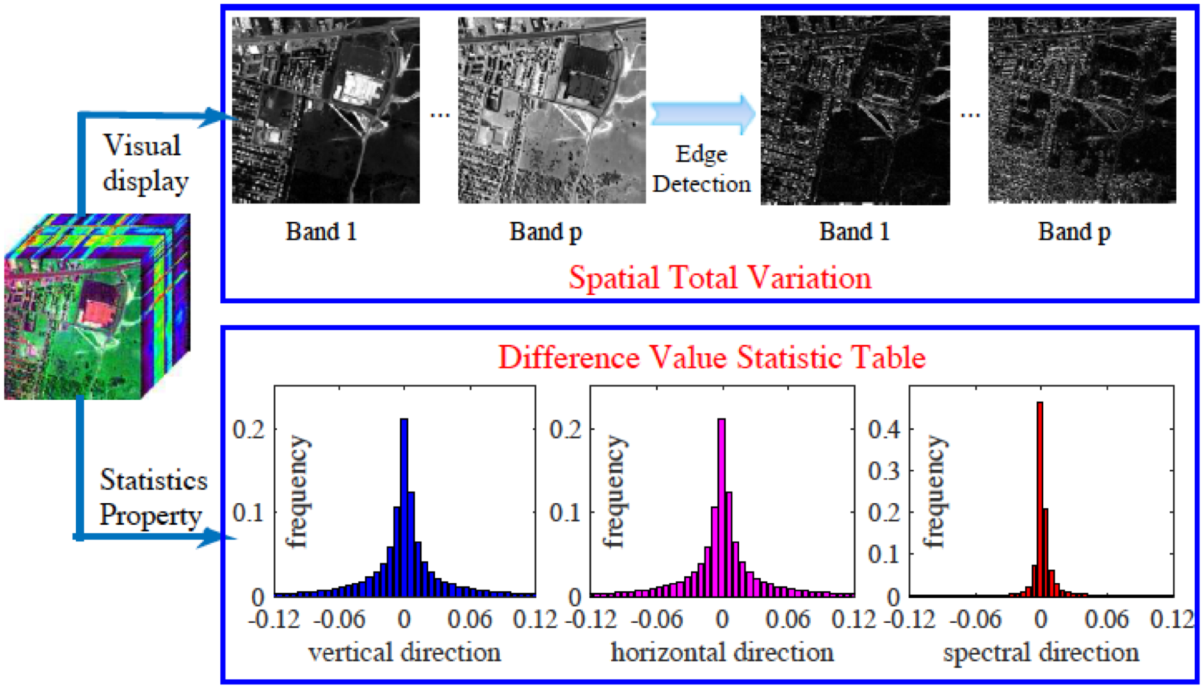
- PnP-GAP: *(Yuan, et al., CVPR'20, TPAMI'21)*

$$\mathbf{x}^{(k+1)} = \mathbf{v}^{(k)} + \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}(\mathbf{y} - \mathbf{H}\mathbf{v}^{(k)}),$$
$$\mathbf{v}^{(k+1)} = \mathcal{D}_\sigma(\mathbf{x}^{(k+1)}).$$

Key idea: replaces the proximal operator with a denoiser

TV+Pretrained FFDNet

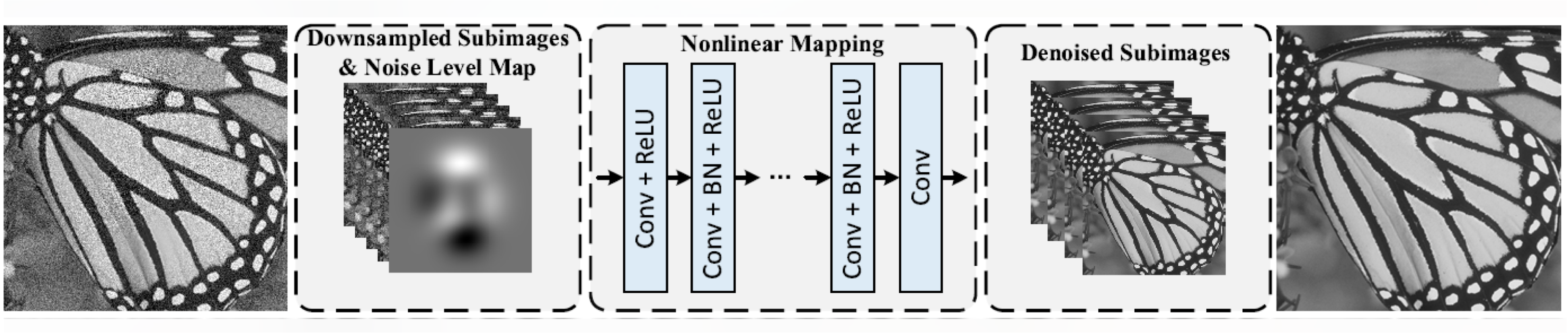
- The variation along the spectral direction is *very small*
- It is typically *piecewise smooth* along the spatial domain



Total Variation Prior

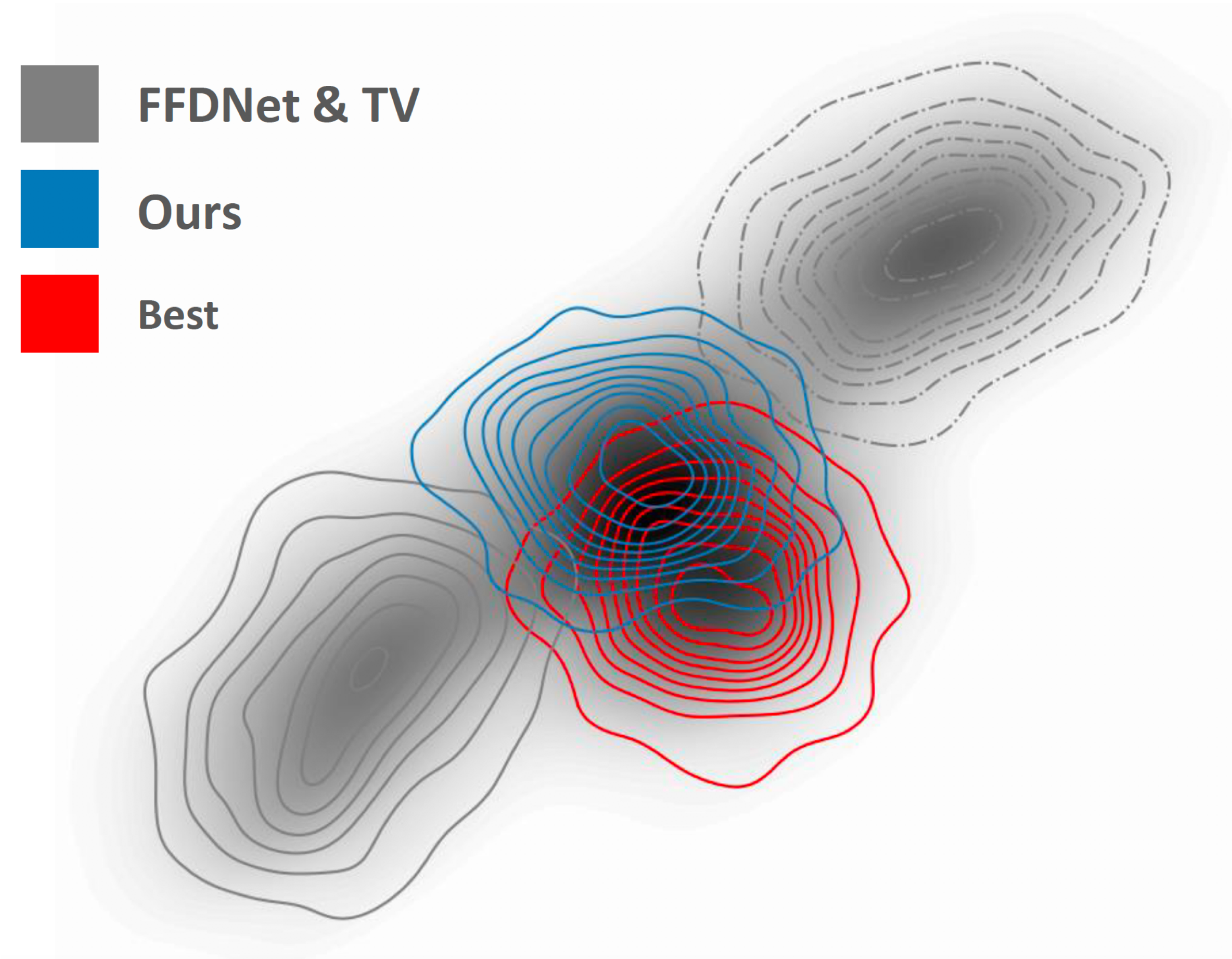
TV+Pretrained FFDNet

- Deep denoiser is efficient while some model-based methods such BM3D and WNNM are time-consuming
- Spatially variant noise can be flexibly handled
- Learn to characterize complex image structures from other data



Deep Denoising Prior

TV+Pretrained FFDNet



The Proposed Procedure

- Treat the denoising step as MAP estimation:

$$\mathbf{v}^{(k+1)} = \arg\min_{\mathbf{v}} \frac{\lambda}{\nu} g(\mathbf{v}) + \frac{1}{2}\|\mathbf{x}^{(k+1)} - \mathbf{v}\|_2^2 = \arg\max_{\mathbf{v}} p(\mathbf{v}|\mathbf{x}^{(k+1)}, \sigma)$$
$$= \mathcal{D}_\sigma(\mathbf{x}^{(k+1)})$$

- Eliminate the hyperparameters as

$$p(\mathbf{v}|\mathbf{x}^{(k+1)}) = \int p(\mathbf{v}|\mathbf{x}^{(k+1)}, \sigma) p(\sigma|\mathbf{x}^{(k+1)}) d\sigma$$
$$q(\mathbf{v}|\mathbf{x}^{(k+1)}) = \int q(\mathbf{v}|\mathbf{x}^{(k+1)}, t) q(t|\mathbf{x}^{(k+1)}) dt$$

- Minimizing the distance:

$$\min_{p(\sigma|\mathbf{x}^{(k+1)}), q(t|\mathbf{x}^{(k+1)})} \mathrm{dist}\left(p(\mathbf{v}|\mathbf{x}^{(k+1)}), q(\mathbf{v}|\mathbf{x}^{(k+1)})\right)$$

Use Gaussian distribution to model the posterior and discrete technique

The Proposed Procedure

Algorithm 1 The proposed PnP-GAP

Require: $\mathbf{H}$, $\mathbf{y}$.

1: Initial $\mathbf{v}^{(0)}$, A, B
2: **while** Not Converge **do**
3: Update $\mathbf{x}$ by Eq. (9).
4: Obtain denoising image set $\{\mathbf{v}_\sigma : \sigma \in A\}$ by $\mathbf{v}_\sigma^{(k+1)} = \mathrm{FFD}_\sigma(\mathbf{x}^{(k+1)})$.
5: Obtain denoising image set $\{\mathbf{v}_t : \sigma \in B\}$ by $\mathbf{v}_t^{(k+1)} = \mathrm{TV}_t(\mathbf{x}^{(k+1)})$.
6: Solve optimization problem (20).
7: Update $\mathbf{v}$ by Eq. (21)
8: **end while**

$$\mathbf{x}^{(k+1)} = \mathbf{v}^{(k)} + \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}(\mathbf{y} - \mathbf{H}\mathbf{v}^{(k)}), \quad (9)$$

$$\min_W W^T \mathbf{P} W \quad (20)$$

$$\text{subject to} \sum_{i=|A|+1}^{|A|+|B|} W_i = 1$$

$$\sum_{i=1}^{|A|} W_i = 1, W \geq 0,$$

$$\mathbf{v}^{(k+1)} = \frac{1}{2}\left(\sum_{\sigma \in A} \hat{w}_\sigma^{ffd} \mathbf{v}_\sigma + \sum_{t \in B} \hat{w}_t^{tv} \mathbf{v}_t\right). \quad (21)$$

More details can be found in our CVPR21's paper

Fixed-point Convergence

Assumption 1. *We assume that all denoisers $\mathcal{D}_\sigma : \mathbb{R}^d \mapsto \mathbb{R}^d$ used in our method satisfy*

$$\|(\mathcal{D}_\sigma - \mathbf{I})(x) - (\mathcal{D}_\sigma - \mathbf{I})(y)\|_2 \leq \epsilon \|x - y\|_2$$

for all $x, y \in \mathbb{R}^d$ for some $\epsilon > 0$.

Assumption 2. *Assume that $\{R_j\}_{j=1}^n > 0$ which means for each spatial location j, the B-frame modulation masks at this location have at least one non-zero entries. We further assume $R_{\max} > R_{\min}$.*

Theorem 1. *Assume $\mathbf{H}$ satisfies Assumption 2. Then the following operator*

$$G = \mathcal{D}_\sigma \circ P$$

is a contraction if $\mathcal{D}_\sigma$ satisfies Assumption 1 and

$$0 < \epsilon < \sqrt{\frac{R_{\max}}{R_{\max} - R_{\min}}} - 1.$$

Theorem 2. *Assume $\mathbf{H}$ satisfies Assumption 2. Let P be a Euclidean projection on linear manifold $\mathbf{y} = \mathbf{H}\mathbf{x}$. Then*

$$T = \frac{1}{2}\mathbf{I} + \frac{1}{2}(2P - \mathbf{I})(2\mathcal{D}_\sigma - \mathbf{I})$$

is a contraction if $\mathcal{D}_\sigma$ satisfies Assumption 1 and

$$0 < \epsilon < 1 - \sqrt{1 - \frac{R_{\min}}{R_{\max}}}.$$

We can also prove the convergence of PnP-ADMM

Experiments

- **Bird** data consists of 24 spectral bands, and the size of each spectral band is 1021 $\times$ 703.
- **Toy** data consists of 31 bands, and the size of each band is 512 $\times$ 512.
- **CAVE** data includes 32 spectral images, and each image contains 31 spectral bands. The image size of each band is 512 $\times$ 512.
- Our code is available at https://github.com/ucker/SCI-TVFFDNet.

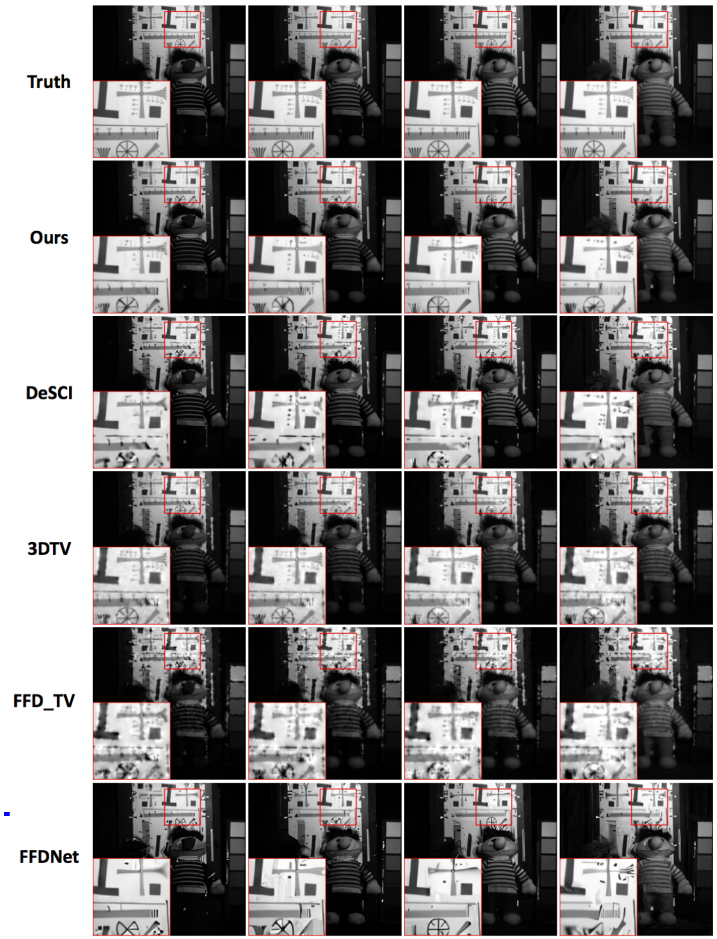
Comparison with Non-DL Methods

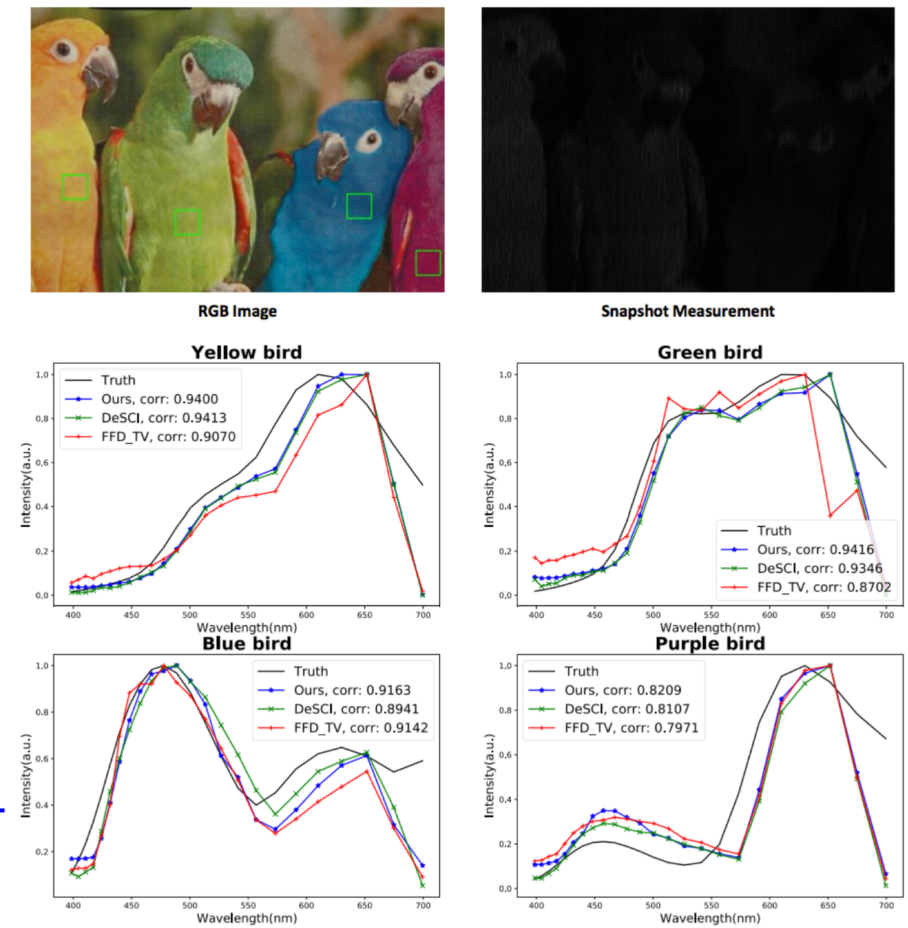
Table 1. The results of PSNR in dB (left entry in each cell) and SSIM (right entry in each cell) by different algorithms on Bird and Toy.

Data	2DTV	3DTV	FFDNet	DeSCI	FFDNet-TV	Ours (2DTV)	Ours (3DTV)
Toy	25.26, 0.8630	28.46, 0.9102	24.28, 0.8298	26.62, 0.9116	25.49, 0.8748	**29.35, 0.9249**	28.86, 0.9225
Bird	37.58, 0.9361	25.84, 0.7919	36.60, 0.9171	38.25, 0.9520	38.21, 0.9383	**39.73, 0.9559**	31.30, 0.9069

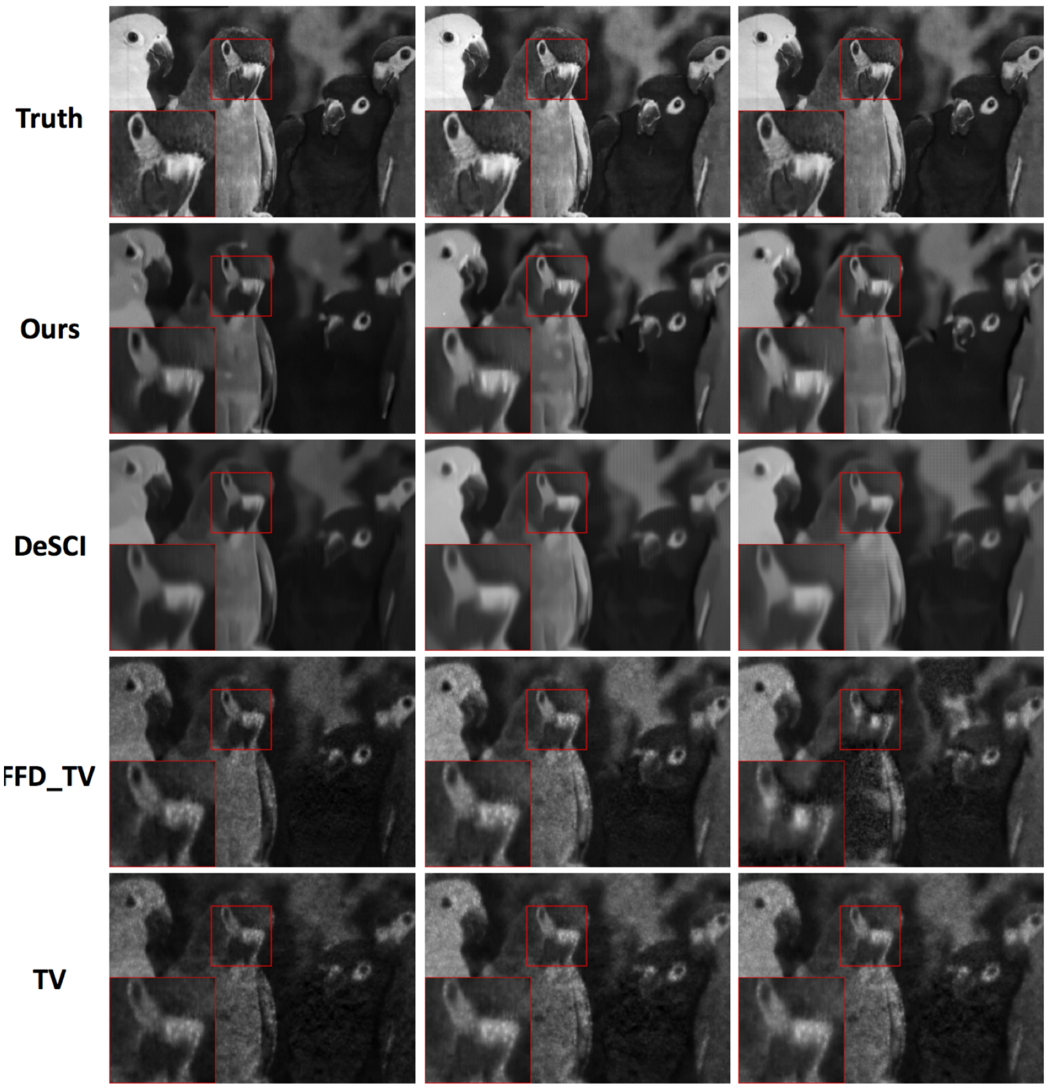
Table 2. The average results of PSNR in dB (left entry in each cell) and SSIM (right entry in each cell) by different algorithms on CAVE.

	2DTV	3DTV	FFDNet	DeSCI	FFDNet-TV	Ours (2DTV)	Ours (3DTV)
Average	30.70, 0.8812	30.15, 0.8906	28.65, 0.8339	31.71, 0.9153	31.26, 0.8867	34.46, 0.9318	**34.79, 0.9347**





Comparison with Non-DL Methods



Comparison with DL Methods

Table 1. Comparison with learned prior method AE [1].

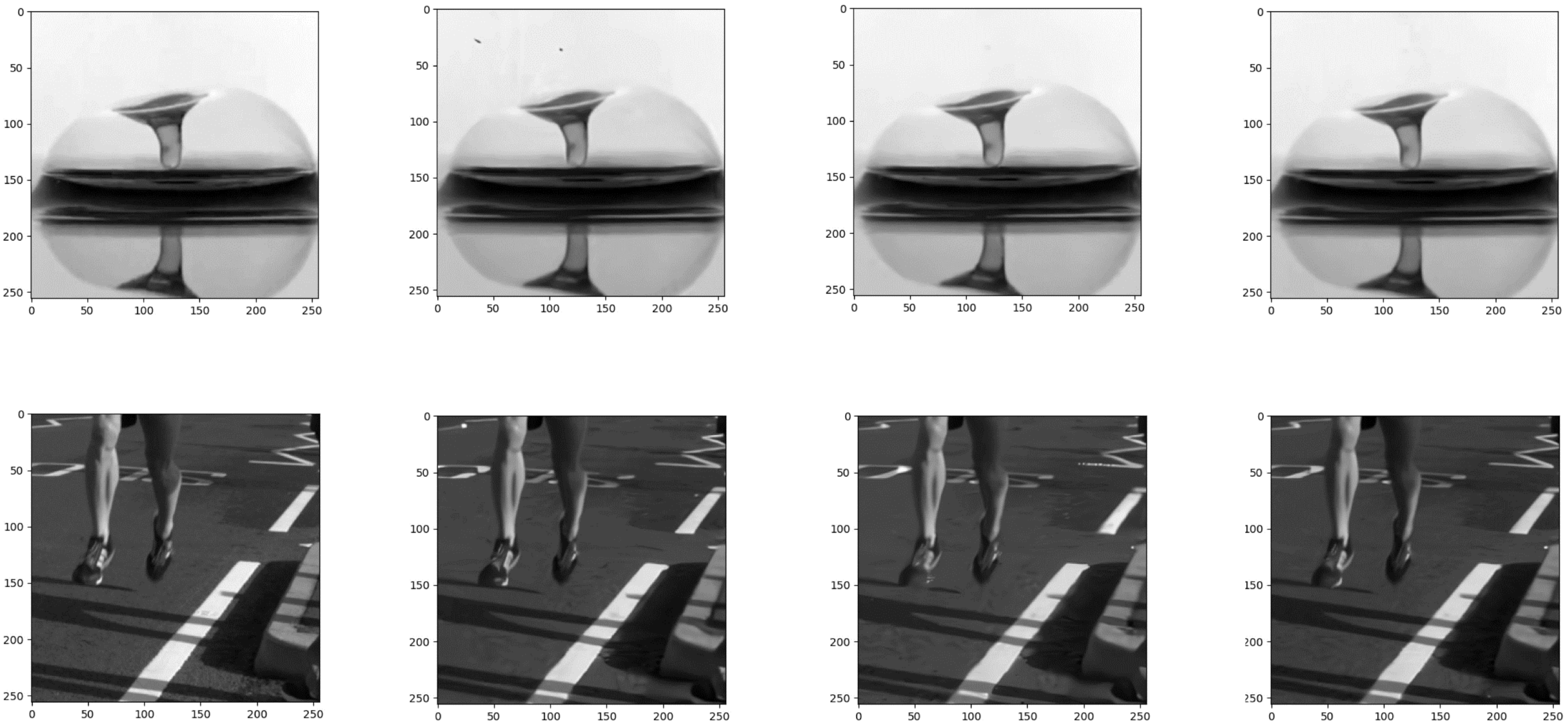
	103	101	73	92	Average
AE	36.75, 0.9726	**38.35, 0.9694**	38.19, 0.9635	32.49, 0.8874	36.45, **0.9482**
3DTV	34.65, 0.9590	35.96, 0.9396	36.36, 0.9490	31.35, 0.8516	34.58, 0.9248
DeSCI	23.98, 0.8053	26.15, 0.8334	28.36, 0.8492	20.76, 0.6865	24.81, 0.7936
Our (3DTV)	**37.05, 0.9735**	38.14, 0.9599	**38.56, 0.9603**	**32.96, 0.8989**	**36.68, 0.9482**

Table 2. Comparison with deep learning method λ-Net.

	3DTV	2DTV	FFDNet	FFDNet-TV	DeSCI	λ-Net	Ours (3DTV)	Ours (2DTV)
Scene 1	31.83, 0.9179	34.68, 0.9321	27.91, 0.8794	35.08, 0.9237	36.07, 0.9505	37.99, 0.8971	37.12, 0.9626	38.16, 0.9630
Scene 2	22.38, 0.7542	27.39, 0.9165	22.55, 0.7463	27.39, 0.9114	30.64, 0.9580	32.70, 0.9465	30.11, 0.9558	30.11, 0.9566
Scene 3	28.66, 0.8851	28.52, 0.8888	21.75, 0.7236	28.04, 0.8883	29.87, 0.9132	34.02, 0.9524	31.70, 0.9156	31.38, 0.9040
Scene 4	25.44, 0.8205	31.72, 0.9309	24.97, 0.8055	32.47, 0.9303	40.35, 0.9780	30.11, 0.9247	31.89, 0.9404	38.25, 0.9691
Scene 5	31.05, 0.8782	31.63, 0.8599	26.17, 0.7643	32.02, 0.8547	33.86, 0.9038	38.10, 0.9330	34.73, 0.9313	34.70, 0.9265
Scene 6	26.47, 0.8517	28.04, 0.8613	26.11, 0.8725	28.79, 0.8525	33.59, 0.9421	30.73, 0.9222	31.44, 0.9234	32.82, 0.9326
Scene 7	29.20, 0.8841	33.90, 0.9287	24.81, 0.8128	34.31, 0.9279	35.76, 0.9515	37.15, 0.9675	34.70, 0.9431	36.17, 0.9476
Scene 8	26.94, 0.8716	30.12, 0.8779	21.64, 0.7117	30.29, 0.8703	31.34, 0.9061	34.35, 0.9454	30.12, 0.9086	31.62, 0.9044
Scene 9	33.31, 0.9350	35.31, 0.9569	37.64, 0.9485	35.90, 0.9541	40.87, 0.9694	36.04, 0.9264	37.07, 0.9705	40.56, 0.9746
Scene 10	25.23, 0.8307	27.59, 0.8431	20.37, 0.6289	27.83, 0.8400	28.96, 0.8746	29.47, 0.9062	28.81, 0.8803	28.99, 0.8731
Average	28.05, 0.8629	30.89, 0.8996	25.39, 0.7893	31.21, 0.8953	34.13, 0.9347	34.07, 0.9321	32.77, 0.9332	**34.28, 0.9352**

Extension to Video CS

Truth | **FFDNet** | **DnCNN** | **Ours**



sensing ratio: 1/8

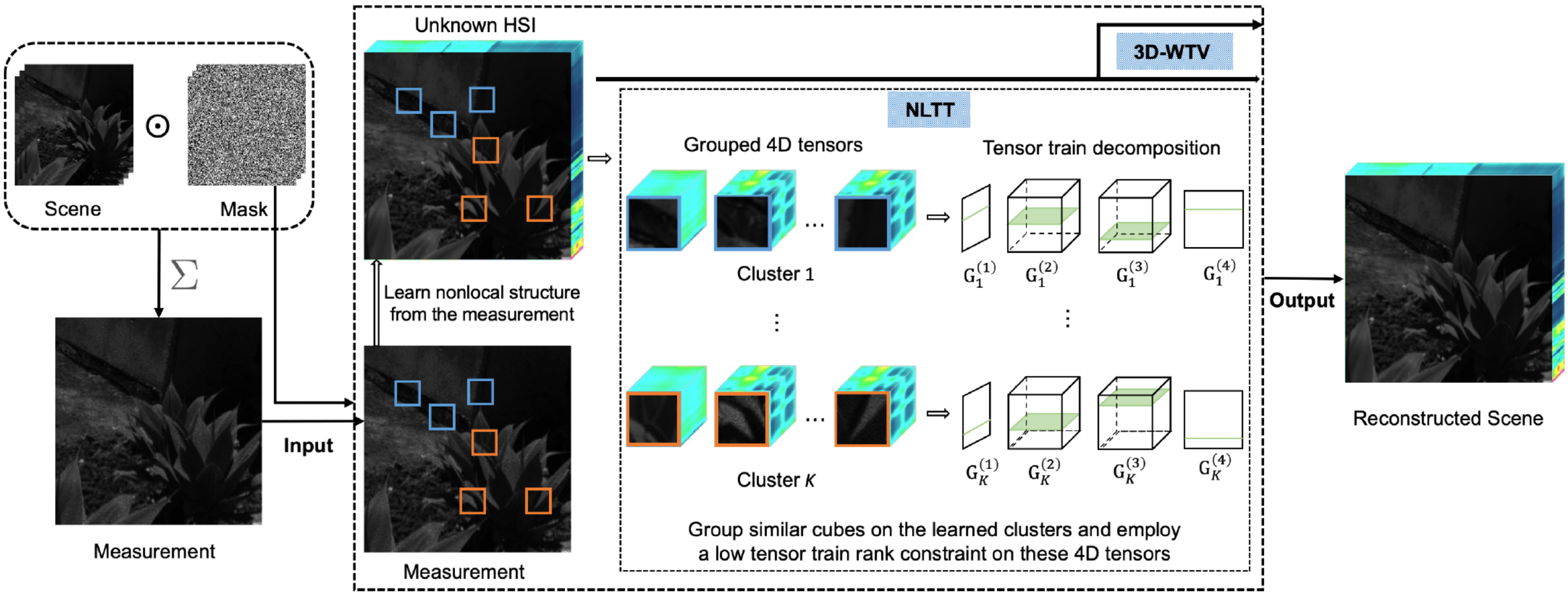
Summary

- Snapshot compressive imaging is an effective way to capture HD image data
- Our PnP algorithms are very flexible
- Several interesting problems need to be further investigated: recovery theory, convergence rate, …

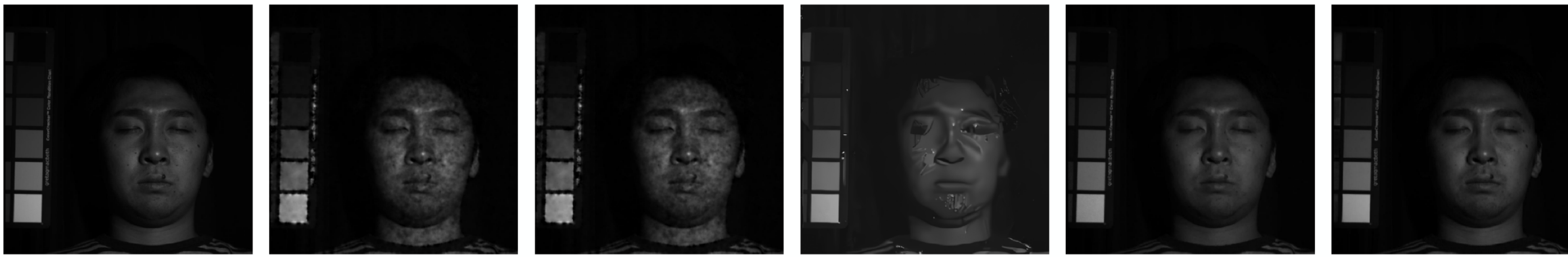
References:

- Xin Yuan, David J. Brady, Aggelos K. Katsaggelos, Snapshot Compressive Imaging: Theory, Algorithms and Applications, IEEE Signal Processing Magazine, 2021
- Haiquan Guo, Yao Wang, Deyu Meng, Effective Snapshot Compressive-spectral Imaging via Deep Denoising and Total Variation Priors, CVPR, 2021
- Shirin Jalali, Xin Yuan, Snapshot Compressed Sensing: Performance Bounds and Algorithms, IEEE TIT, 2019

TV Regularized Nonlocal Low-rank Tensor Train



TV Regularized Nonlocal Low-rank Tensor Train



Truth	GAP–TV	DeSCI	FFDNet	FFDNet+3DTV	NLTT–TV
PSNR/SSIM	30.89/0.93	35.03/0.96	31.28/0.93	36.34/0.96	39.91/0.98
TIME	5m	100m	1m	20m	60m

Face image: 512*512*31

Thank you!