ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446309

SASA: SALIENCY-AWARE SELF-ADAPTIVE SNAPSHOT COMPRESSIVE IMAGING

Yaping Zhao[1,2] , *Edmund Y. Lam*[1,2,*]

[1] Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong SAR
[2]ACCESS — AI Chip Center for Emerging Smart Systems, Hong Kong SAR

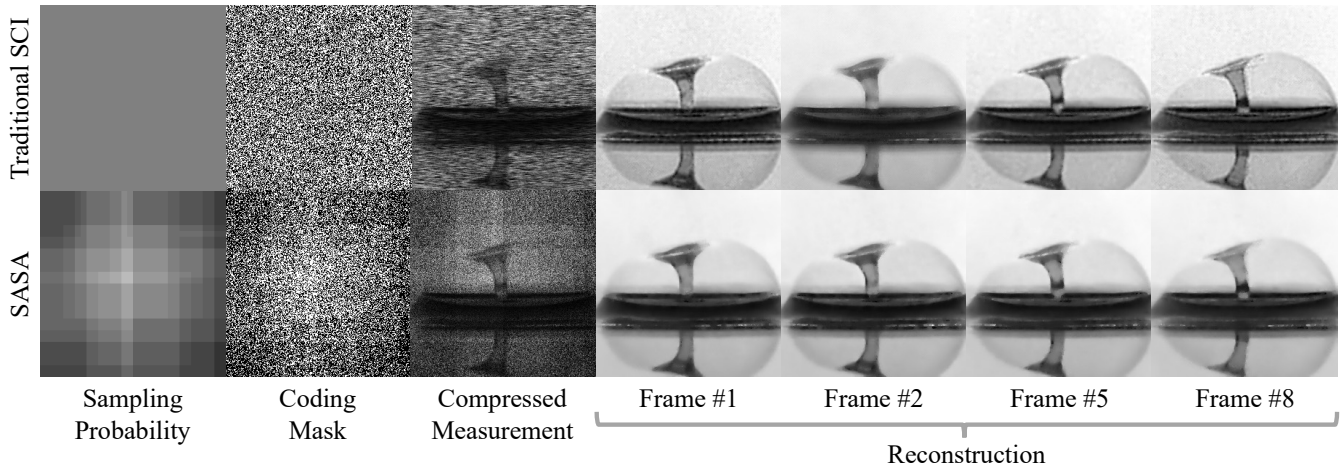

Fig. 1: While traditional snapshot compressive imaging (SCI) directly generates random binary matrices as coding masks to obtain compressed measurements, we propose a framework elegantly with saliency detection to perform self-adaptive sampling, and thus boosting the quality of compressed measurement and the performance of reconstruction result.

ABSTRACT

The ability of snapshot compressive imaging (SCI) systems to efficiently capture high-dimensional (HD) data depends on the advent of novel optical designs to sample the HD data as two-dimensional (2D) compressed measurements. Nonetheless, the traditional SCI scheme is fundamentally limited, due to the complete disregard for high-level information in the sampling process. To tackle this issue, in this paper, we pave the first mile toward the advanced design of adaptive coding masks for SCI. Specifically, we propose an *efficient* and *effective* algorithm to generate coding masks with the assistance of saliency detection, in a *low-cost* and *low-power* fashion. Experiments demonstrate the effectiveness and efficiency of our approach. Code is available at: https://github.com/IndigoPurple/SASA.

Index Terms— snapshot compressive imaging, coded image sensing, compressed sensing, imaging systems, computational imaging

1. INTRODUCTION

Snapshot compressive imaging (SCI) systems address the challenges associated with high-dimensional (HD) visual signal acquisition and analysis [1]. These systems capture HD data as compressed two-dimensional (2D) measurements, and then utilize compressed sensing algorithms to reconstruct the HD data [2, 3, 4, 5]. By employing optical designs, SCI systems enable the acquisition of HD visual signals in a single snapshot, as opposed to conventional methods. This capability proves advantages in scenarios where capturing dynamic or fast-moving scenes is crucial. Therefore, SCI systems offer promising solutions for real-world applications in fields such as surveillance, biomedical imaging, remote sensing

The key principle behind SCI lies in the compression of the acquired HD data. Rather than directly sampling the entire HD signal, SCI systems capture a compressed version of the data, resulting in compressed measurements. However, the existing research in this area has predominantly focused on the reconstruction algorithms, while the study of compression sampling methods remains lacking, let alone optimizing

*Corresponding author.

This work is supported in part by the Research Grants Council (GRF 17201822), by the Research Postgraduate Student Innovation Award (The University of Hong Kong), and by ACCESS — AI Chip Center for Emerging Smart Systems, Hong Kong SAR.

Notation	Description
$\mathbf{X} \in \mathbb{R}^{H \times W \times C}$	The video frames we aim to compress and reconstruct. $\forall c = 1, \ldots, C$, $\mathbf{X}_c = \mathbf{X}(:, :, c) \in \mathbb{R}^{H \times W}$ is the c-th video frame. Here, H, W and C are the image height, image width and compression rate, respectively.
$\mathbf{E} \in \mathbb{R}^{H \times W}$	The measurement noise.
$\mathbf{Y} \in \mathbb{R}^{H \times W}$	The compressed measurement.
$\mathbf{A} \in \{0, 1\}^{H \times W \times C}$	The sensing matrix we use to compress the video frames. $\forall h = 1, \ldots, H, \forall w = 1, \ldots, W, \forall c = 1, \ldots, C$, $\mathbf{A}_c = \mathbf{A}(:, :, c) \in \{0, 1\}^{H \times W}$ is the c-th coding mask, $a_{hwc} = \mathbf{A}(h, w, c) \in \{0, 1\}$ is the element on the h-th row and w-th column of the c-th coding mask.
$Bernoulli(p)$	The Bernoulli distribution, which is the discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability $q = 1 - p$.
$\mathbf{S} \in \mathbb{R}^{H \times W \times D}$	The saliency maps, where $\mathbf{S}_d = \mathbf{S}(:, :, d) \in \{0, 1\}^{H \times W}$, is the d-th saliency map. Here, value 1 indicates saliency, and value 0 indicates no saliency. D is the maximum number of detections to examine.
$\mathbf{P} \in \mathbb{R}^{H \times W}$	The sampling probability, where $p_{hw} = \mathbf{P}(h, w) \in [0, 1]$ is a probability parameter.

Table 1: Key notations and descriptions.

the mask design. Additionally, the current sampling methods treat all pixels equally by employing random binary masks, completely disregarding high-level information such as objects and saliency. As a result, the importance of different regions of the image is not properly considered.

An even more serious issue is that the *"sampling - reconstruction - tasks"* scheme of traditional SCI is fundamentally limited. On one hand, the complete ignorance of high-level information in the sampling process leads to wasteful resource allocation for content-irrelevant computation. On another hand, the independence of subsequent computer vision tasks hampers the sampling and design efficiency of SCI [6].

To solve those problems, as Figure 1 shows, we propose an *efficient* and *effective* algorithm to generate adaptive coding masks with the assistance of saliency detection, in a *low-cost* and *low-power* fashion. Our key insight lies in that the sampling process to acquire compressed measurements should be performed adaptively and intelligently, which assigns a higher sampling probability to salient image regions and a lower probability to non-salient regions.

Our main contributions are as follows:

- To the best of our knowledge, this work is **the first** to design adaptive coding masks specifically for SCI.
- Our proposed method introduces a **low-cost and efficient** approach by performing saliency detection directly on the measurement. This strategy requires only a few atomic operations, making it highly suitable for real-time applications. Notably, the processing speed is on average 250fps on a single laptop CPU.
- Quatitative and qualitative experiments on classical datasets demonstrate the **effectiveness** of our method.

2. RELATED WORK

SCI systems consist of two components: an encoding process that compresses the HD data into a 2D measurement by hardware [7, 8, 1], and a decoding process that reconstructs the desired signal by algorithms [3, 9, 5, 10, 11, 12, 3, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22]. Nevertheless, the existing research in this area focused on the reconstruction algorithms, leaving a gap in the study of compression sampling methods.

Despite recent efforts to bridge the gap between compressive imaging and computer vision, such as directly performing vision tasks on the compressed measurements rather than the reconstructed data [23, 6], the fundamental issue remains unresolved: the complete disregard of high-level information during the sampling process. As a result, there is a pressing need to conduct research on optimizing the mask design. Traditional SCI systems [1] treat all pixels equally, employing random binary masks that fail to consider crucial high-level information, such as objects and saliency. Consequently, the importance of different regions within an image is not appropriately taken into account.

Therefore, in this paper, we aim to pave the first mile towards an adaptive SCI system that leverages high-level information during the sampling process.

3. METHOD

In this section, we first formulate the sampling process in video SCI. Next, we propose an *efficient* and *effective* algorithm to generate adaptive coding masks, in a *low-cost* and *low-power* fashion. Table 1 outlines key notations used in this paper.

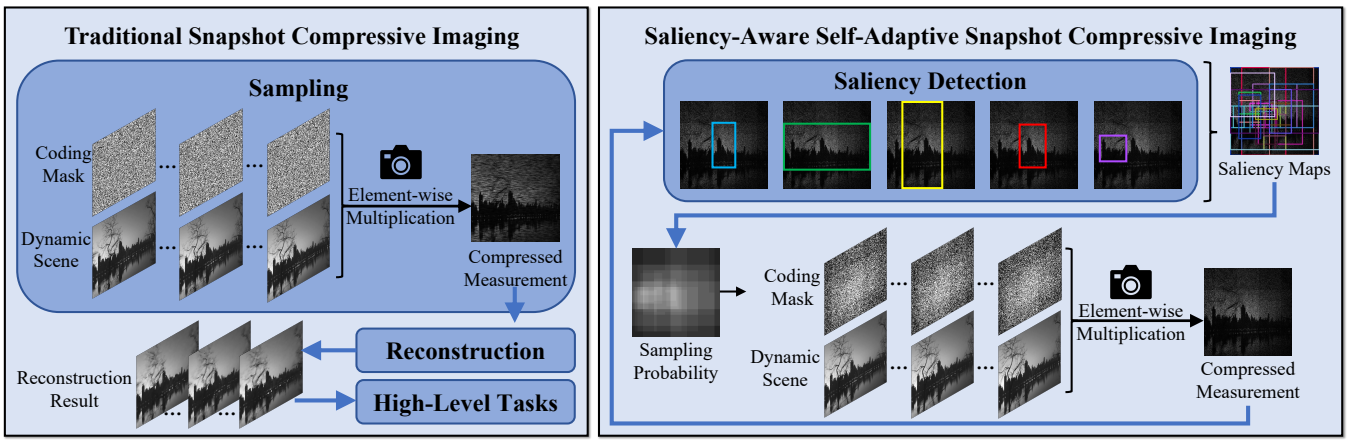

Fig. 2: The pipeline of traditional SCI and our proposed saliency-aware self-adaptive SCI, respectively.

Figure 2 provides an intuitive understanding of our pipeline, jointly with the comparison to traditional SCI. While traditional SCI directly generates random binary matrices as coding masks to obtain compressed measurements, our framework elegantly utilizes saliency detection to perform self-adaptive sampling.

To begin with, the acquisition of compressed measurements in video SCI could be mathematically modeled as:

$$\mathbf{Y} = \sum_{c=1}^{C} \mathbf{A}_c \odot \mathbf{X}_c + \mathbf{E}, \tag{1}$$

where $\mathbf{Y}$ is the compressed measurement, $\mathbf{A}$ is the sensing matrix, $\mathbf{X}$ is the video frames we aim to compress and reconstruct, $\mathbf{E}$ is the measurement noise, and $\odot$ denotes the Hadamard (element-wise) product.

In traditional SCI [1], the sensing matrix $\mathbf{A}$ is always generated by randomly sampling elements from a Bernoulli distribution with probability $p = 0.5$. In our SCI system, we initialize the first sensing matrix by the traditional method:

$$\begin{aligned} \mathbf{A}^{(0)} = [a^{(0)}_{hwc}], \qquad a^{(0)}_{hwc} \sim Bernoulli(0.5), \\ \forall h = 1, \ldots, H,\ \forall w = 1, \ldots, W,\ \forall c = 1, \ldots, C, \end{aligned} \tag{2}$$

which is used to preliminary compress the dynamic scene and capture the first measurement:

$$\mathbf{Y}^{(0)} = \sum_{c=1}^{C} \mathbf{A}^{(0)}_c \odot \mathbf{X}^{(0)}_c + \mathbf{E}^{(0)}. \tag{3}$$

Once the first measurement is obtained, we perform advanced vision processing technology, *i.e.*, saliency detection directly on the measurement in stead of on the reconstructed video, to boost inference speed and reduce bandwidth occupation, memory footprint and energy consumption. Specifically, we adopt the light-weight algorithm from [24], denoted as f, which takes the measurement $\mathbf{Y}^{(t)}$ corresponding to the t-th video sequence as input and outputs saliency maps $\mathbf{S}^{(t)}$:

$$\mathbf{S}^{(t)} = f(\mathbf{Y}^{(t)}; D), \qquad t = 0, 1, 2, \ldots, \tag{4}$$

where D is a parameter that determines the maximum number of detections to examine. Based on the estimated saliency maps, we calculate the sampling probability for the next video sequence by retrieving saliency in the compressed domain with low bandwidth:

$$\mathbf{P}^{(t+1)} = \frac{1}{D} \sum_{d=1}^{D} \mathbf{S}^{(t)}_d, \qquad t = 0, 1, 2, \ldots, \tag{5}$$

where $\mathbf{P}$ is the sampling probability used to guide sensing matrix generation for the next video sequence. It assigns higher sampling probabilities to salient image regions and lower to non-salient, ensuring an adaptive and effective sampling process. To ensure sampling coverage in regions without salient events, we then assign a probability of $1/D$ to areas with zero probability. According to $\mathbf{P} = [p_{hw}]$, the sensing matrix is updated by:

$$\begin{aligned} \mathbf{A}^{(t+1)} = [a^{(t+1)}_{hwc}], \qquad a^{(t+1)}_{hwc} \sim Bernoulli(p^{(t+1)}_{hw}), \\ \forall h = 1, \ldots, H,\ \forall w = 1, \ldots, W,\ \forall c = 1, \ldots, C, \\ t = 0, 1, 2, \ldots. \end{aligned} \tag{6}$$

Here, our key insight lies in that the sampling process of acquiring compressed measurements should be performed adaptively, which assigns a higher sampling probability to salient image regions and lower to non-salient regions.

Finally, the measurement of the next video sequence is captured using the dynamically updated sensing matrix:

$$\begin{aligned} \mathbf{Y}^{(t+1)} = \sum_{c=1}^{C} \mathbf{A}^{(t+1)}_c \odot \mathbf{X}^{(t+1)}_c + \mathbf{E}^{(t+1)}, \\ t = 0, 1, 2, \ldots. \end{aligned} \tag{7}$$

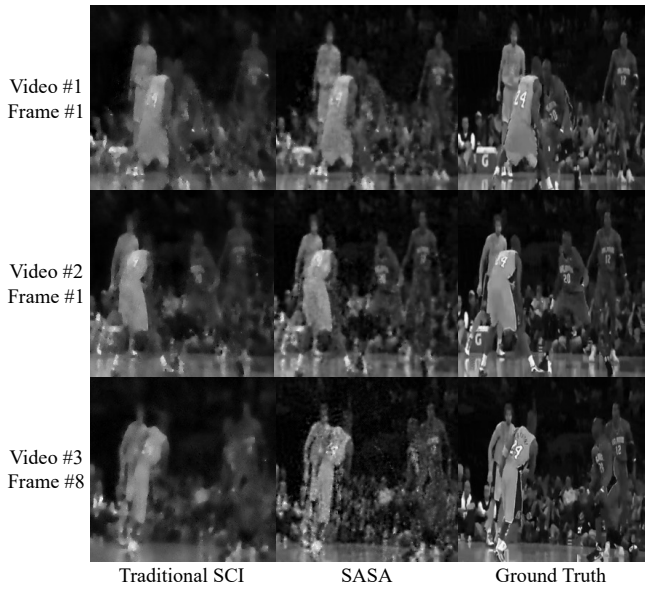

Fig. 3: Comparisons of reconstruction results on the dataset `Kobe`. Reconstruction is by ADMM-TV [2].

It is worth noting that our method performs saliency detection directly on the measurement, and requires only a few atomic operations to generate sensing matrices. In our experiments, the processing speed is on average 250fps on a single laptop CPU, which means that **it enables a low-speed camera to capture high-speed scenes with $C \times 250$ fps** while intelligently and adaptively varying the coding masks.

4. EXPERIMENT

Experimental Setting. We initialize the first sensing matrix and compressed measurement following the traditional SCI [1], and then use the proposed method to generate adaptive sensing matrices for capturing sequential compressed measurements. For evaluations, four of the six classical SCI datasets [9] including `Aerial`, `Vehicle`, `Kobe` and `Traffic` are adopted, while two (`Drop` and `Runner`) are excluded because the videos are too short and they contain only a single measurement. Then two classical SCI reconstruction algorithms ADMM-TV [2] and GAP-TV[3], and a state-of-the-art method DEQSCI [4, 25] are conducted, on different measurements captured by traditional SCI and our proposed framework (SASA), to reconstruct the videos.

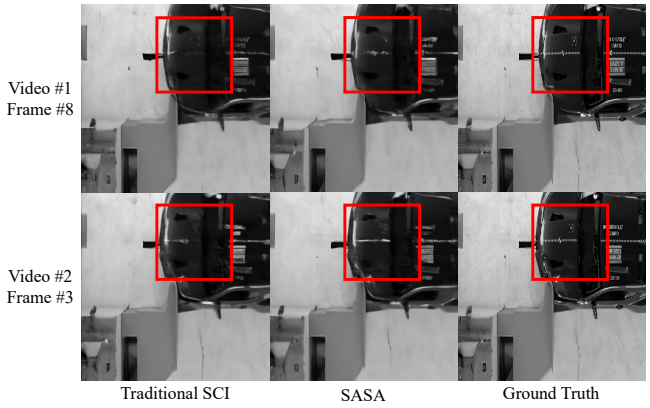

Fig. 4: Comparisons of reconstruction results on the dataset `Vehicle`. Reconstruction is by DEQSCI [4].

Dataset	Aerial	Vehicle	Kobe	Traffic	Average
Recon.	ADMM-TV [2]				
Trad. SCI	24.54, 0.877	23.88, 0.822	25.38, 0.821	20.15, 0.740	23.49, 0.815
SASA	**24.61, 0.882**	**24.44, 0.862**	**26.35, 0.877**	**20.19, 0.746**	**23.90, 0.842**
Recon.	GAP-TV [3]				
Trad. SCI	24.69, 0.861	24.51, 0.867	25.97, 0.863	20.44, 0.763	23.90, 0.839
SASA	**24.71, 0.864**	**24.57, 0.872**	**26.80, 0.895**	**20.48, 0.771**	**24.13, 0.851**

Table 2: The results in terms of PSNR (dB) and SSIM by different compressed measurements on classical four datasets.

D	Aerial	Vehicle	Kobe	Traffic	Average
10	24.77, 0.871	23.69, 0.867	25.16, 0.845	19.92, 0.740	23.39, 0.831
20	24.55, 0.859	24.13, 0.856	25.89, 0.857	20.39, 0.759	23.74, 0.833
30	**24.71, 0.864**	**24.57, 0.871**	**26.80, 0.895**	**20.48, 0.770**	**24.13, 0.850**
40	24.56, 0.865	24.53, 0.873	26.62, 0.896	20.26, 0.754	23.99, 0.847
50	24.42, 0.862	24.41, 0.871	26.48, 0.896	20.16, 0.745	23.87, 0.843

Table 3: Ablation study on different max detection numbers in saliency detection. Reconstruction is by GAP-TV [3].

Quantitative and Qualitative Evaluations. Quantitative comparison results of different compressed measurements are provided in Table 2, where our method achieves approximately 0.2∼0.5 dB improvement in PSNR and 0.01∼0.03 in SSIM [26] on average. The improvement indicates our method could reconstruct images with relatively fine structure, which is confirmed by qualitative evaluations in Figures 3 and 4. As shown in these figures, reconstruction results from traditional SCI have more artifacts and distortions around margins. Our method can maintain a clear and accurate image structure, thus leading to higher performance. Ours adjusts the coding masks during the sampling process, seamlessly integrates with existing reconstruction algorithms, and enhances performance across all baselines. Furthermore, the reconstruction quality can be further improved using various video processing techniques [27, 28, 29, 30, 31].

Ablation Study. In Table 3, we experimented with different values for D, the maximum number of detections to examine in saliency detection, and found $D = 30$ is the best setting.

5. CONCLUSION

This work addresses the limitations of existing SCI systems by an efficient and effective framework to adjust coding masks, paving the first mile towards adaptive SCI systems.

6. REFERENCES

[1] X. Yuan, D. J. Brady, and A. K. Katsaggelos, "Snapshot compressive imaging: Theory, algorithms, and applications," *IEEE Signal Processing Magazine*, vol. 38, no. 2, pp. 65–88, 2021.

[2] S. H. Chan, X. Wang, and O. A. Elgendy, "Plug-and-play admm for image restoration: Fixed-point convergence and applications," *IEEE Transactions on Computational Imaging*, 2016.

[3] X. Yuan, "Generalized alternating projection based total variation minimization for compressive sensing," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 2539–2543.

[4] Y. Zhao, S. Zheng, and X. Yuan, "Deep equilibrium models for snapshot compressive imaging," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 3642–3650.

[5] Q. Yang and Y. Zhao, "Revisit dictionary learning for video compressive sensing under the plug-and-play framework," in *SPIE Proceedings*, vol. 12166. SPIE, 2022, pp. 2018–2025.

[6] Z. Zhang, B. Zhang, X. Yuan, S. Zheng, X. Su, J. Suo, D. J. Brady, and Q. Dai, "From compressive sampling to compressive tasking: retrieving semantics in compressed domain with low bandwidth," *PhotoniX*, vol. 3, no. 1, pp. 1–22, 2022.

[7] 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.

[8] 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. 10 526–10 545, 2013.

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

[10] 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*, 2018.

[11] 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*, vol. 24, no. 1, pp. 106–119, January 2015.

[12] 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*, 2014.

[13] M. Qiao, Z. Meng, J. Ma, and X. Yuan, "Deep learning for video compressive sensing," *APL Photonics*, vol. 5, no. 3, p. 030801, 2020.

[14] S. Zheng, C. Wang, X. Yuan, and H. L. Xin, "Super-compression of large electron microscopy time series by deep compressive sensing learning," *Patterns*, vol. 2, no. 7, p. 100292, 2021.

[15] L. Wang, M. Cao, Y. Zhong, and X. Yuan, "Spatial-temporal transformer for video snapshot compressive imaging," *arXiv preprint arXiv:2209.01578*, 2022.

[16] Z. Cheng, B. Chen, R. Lu, Z. Wang, H. Zhang, Z. Meng, and X. Yuan, "Recurrent neural networks for snapshot compressive imaging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[17] Z. Meng and X. Yuan, "Perception inspired deep neural networks for spectral snapshot compressive imaging," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 2813–2817.

[18] Z. Meng, S. Jalali, and X. Yuan, "Gap-net for snapshot compressive imaging," *arXiv preprint arXiv:2012.08364*, 2020.

[19] Z. Wu, J. Zhang, and C. Mou, "Dense deep unfolding network with 3d-cnn prior for snapshot compressive imaging," *arXiv preprint arXiv:2109.06548*, 2021.

[20] Y. Zhao, "Mathematical cookbook for snapshot compressive imaging," *arXiv preprint arXiv:2202.07437*, 2022.

[21] Y. Zhao, Q. Zeng, and E. Y. Lam, "Adaptive compressed sensing for real-time video compression, transmission, and reconstruction," in *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2023.

[22] Z. Y. Chong, Y. Zhao, Z. Wang, and E. Y. Lam, "Solving inverse problems in compressive imaging with score-based generative models," in *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2023.

[23] S. Lu, X. Yuan, and W. Shi, "Edge compression: An integrated framework for compressive imaging processing on cavs," in *2020 IEEE/ACM Symposium on Edge Computing (SEC)*. IEEE, 2020, pp. 125–138.

[24] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3286–3293.

[25] Y. Zhao, S. Zheng, and X. Yuan, "Deep equilibrium models for video snapshot compressive imaging," *arXiv preprint arXiv:2201.06931*, 2022.

[26] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[27] Y. Zhao, M. Ji, R. Huang, B. Wang, and S. Wang, "EFENet: Reference-based video super-resolution with enhanced flow estimation," in *Artificial Intelligence: First CAAI International Conference, CICAI 2021, Hangzhou, China, June 5–6, 2021, Proceedings, Part I 1*. Springer, 2021, pp. 371–383.

[28] Y. Zhao, H. Zheng, Z. Wang, J. Luo, and E. Y. Lam, "MANet: improving video denoising with a multi-alignment network," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 2036–2040.

[29] Y. Zhao, G. Li, and E. Y. Lam, "Cross-camera human motion transfer by time series analysis," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024.

[30] Y. Zhao, H. Zheng, M. Ji, and R. Huang, "Cross-camera deep colorization," in *CAAI International Conference on Artificial Intelligence*. Springer, 2022, pp. 3–17.

[31] Y. Zhao, H. Zheng, J. Luo, and E. Y. Lam, "Improving video colorization by test-time tuning," in *IEEE International Conference on Image Processing (ICIP)*, 2023.