HIERARCHICAL INTERACTIVE RECONSTRUCTION NETWORK FOR VIDEO COMPRESSIVE SENSING

Tong Zhang, Wenxue Cui, Chen Hui, Feng Jiang*

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China State Key Laboratory of Communication Content Cognition, People's Daily Online, Beijing, China 100733 ChongQing Research Institute of HIT

({tongzhang, wenxuecui, chenhui}@stu.hit.edu.cn; fjiang@hit.edu.cn)

ABSTRACT

Deep network-based image and video Compressive Sensing (CS) has attracted increasing attentions in recent years. However, in the existing deep network-based CS methods, a simple stacked convolutional network is usually adopted, which not only weakens the perception of rich contextual prior knowledge, but also limits the exploration of the correlations between temporal video frames. In this paper, we propose a novel Hierarchical InTeractive Video CS Reconstruction Network(HIT-VCSNet), which can cooperatively exploit the deep priors in both spatial and temporal domains to improve the reconstruction quality. Specifically, in the spatial domain, a novel hierarchical structure is designed, which can hierarchically extract deep features from keyframes and non-keyframes. In the temporal domain, a novel hierarchical interaction mechanism is proposed, which can cooperatively learn the correlations among different frames in the multiscale space. Extensive experiments manifest that the proposed HIT-VCSNet outperforms the existing state-of-the-art video and image CS methods in a large margin.

Index Terms— Image/video compressive sensing, video reconstruction, deep learning, feature fusion, hierarchical interaction, convolutional neural network(CNN)

1. INTRODUCTION

Compressive Sensing (CS) theory [2, 6] expounds that if a signal is sparse in a certain domain, it can be recovered from fewer measurements than prescribed by the Nyquist sampling theorem. Mathematically, given the initial signal $\mathbf{x} \in \mathbb{R}^N$, the CS measurement $\mathbf{y} \in \mathbb{R}^M$ is obtained by:

$$\mathbf{y} = \mathbf{\Phi}\mathbf{x} \tag{1}$$

where $\Phi \in \mathbb{R}^{M \times N}$ is the sampling matrix and the sampling ratio can be defined as $\frac{M}{N}$ ($M \ll N$). CS is widely used in magnetic resonance imaging(MRI) [9], snapshot compressive imaging(SCI) [16] and image/video coding [3].

The core mission of CS is to accurately reconstruct the target signal x from the compressed measurements y. Recently,

many image CS methods are proposed, which can be roughly categorized as the following two groups: optimization-based methods and deep learning-based methods. Specifically, 1) optimization-based methods aim to utilize iterative processes to solve a regularized optimization problem:

$$\min_{\mathbf{x}} \frac{1}{2} \|\Phi \mathbf{x} - \mathbf{y}\|_{2}^{2} + \lambda \psi(\mathbf{x})$$
 (2)

where the former term $\frac{1}{2}\|\Phi\mathbf{x}-\mathbf{y}\|_2^2$ denotes the fidelity term and the latter term $\psi(\mathbf{x})$ comes from the prior knowledge, λ is the regularization parameter. The widely used image priors include local smoothing [18], non-local self-similarity [4] and sparsity [19]. Nevertheless, the high computational complexity limits the practical applications of CS significantly. 2) Deep learning-based methods directly map the measurements to the reconstructed images. However, these methods generally construct a black box network [11] which is not interpretable. In recent years, some deep unfolding networks(DUN) [5, 17] try to embed deep neural networks into optimization algorithms, such as HQS [17] and ISTA [1].

Recently, CS is successfully applied for the video signal. Similar to image CS, video CS reconstruction can also be divided into optimization-based methods and deep learning-based methods. For the optimization-based methods, it can be roughly divided into 3D-sparsity reconstruction [8] and motion-compensation reconstruction [7, 10, 23]. The former assumes joint sparsity in the 3D transform domain to recover frames simultaneously. The latter reconstructs each frame independently by motion compensation. While the complex computation restricts the practical application seriously. For the deep learning-based methods, Shi et al. [14] present VC-SNet based on CNN and explore both intraframe and interframe correlations. It is noted that the existing image CS methods can be directly used for video frame compression.

However, the existing deep learning-based video CS methods still face the following problems: 1) The existing CS networks use simple stacked CNNs, which can not perceive rich spatial contextual prior information effectively. 2) When the motion in the video is fast, it is difficult to capture the temporal correlation efficiently.

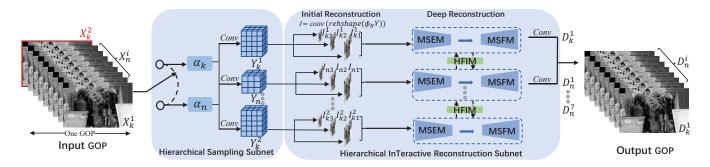


Fig. 1. The architecture of the HIT-VCSNet.

To overcome all these drawbacks, in this paper, we propose a novel Hierarchical InTeractive Video CS Reconstruction Network(HIT-VCSNet), which can exploit the deep priors in both spatial and temporal domains. In the spatial domain, we apply a Hierarchical Feature Fusion Module(HFFM) to hierarchically perceive multi-scale contextual priors. In the temporal domain, we propose a Hierarchical Feature Interaction Module(HFIM) to automatically interact hierarchical interframe information.

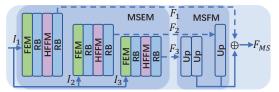
In summary, our main contributions are as followings: 1) We present an end-to-end Hierarchical InTeractive Video CS Reconstruction Network HIT-VCSNet, which can cooperatively exploit the deep priors in both spatial and temporal domains. 2) In the spatial domain, a Hierarchical Feature Fusion Module(HFFM) is presented to hierarchically perceive richer contextual priors in the multi-scale space. 3) In the temporal domain, a Hierarchical Feature Interaction Module(HFIM) is developed to automatically learn the interframe correlations in a hierarchical manner. 4) Extensive experiments manifest that the proposed HIT-VCSNet outperforms the existing state-of-the-art video and image CS networks in a large margin.

2. THE PROPOSED HIT-VCSNET

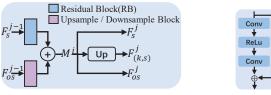
As showed in Fig.1, our HIT-VCSNet composes a hierarchical sampling subnet and a hierarchical interactive reconstruction subnet. Given the GOPs as the input of HIT-VCSNet, the sampling subnet outputs the CS measurements of frames. The initial reconstruction subnet recovers CS measurements into multi-scale initial frames. The deep reconstruction subnet is composed of a Multi-Scale Extraction Module(MSEM) and a Multi-Scale Fusion Module(MSFM). Moreover, a HFIM is applied to interact interframe information among keyframes and non-keyframes. Ultimately, the outputs of MSFM pass a convolutional layer to convert into final frames.

2.1. Hierarchical Sampling Subnet

As depicted in Fig.1, our sampling subnet has two model settings, the high sampling rate mode for keyframes(i.e., α_k) along with the low sampling rate mode for non-keyframes(i.e.,



(a) Hierarchical Deep Reconstruction Subnet



(b) Hierarchical Feature Fusion Module

(c) RB

Fig. 2. The sub-modules of the deep reconstruction subnet.

 α_n). Therefore, the input frames in a GOP flow into corresponding branches separately. We divide the input keyframes \mathbf{X}_k and non-keyframes \mathbf{X}_n into certain numbers of non-overlapping image blocks of size $B \times B$. The CS sampling process $\mathbf{y} = \mathbf{\Phi} \mathbf{x}$ can be simulated by convolution as follows:

$$\mathbf{Y}_k = \mathbf{S}_k * \mathbf{X}_k \tag{3}$$

$$\mathbf{Y}_n = \mathbf{S}_n * \mathbf{X}_n \tag{4}$$

where \mathbf{S}_k and \mathbf{S}_n has $\alpha_k B^2$ and $\alpha_n B^2$ filters of size $B \times B$. In addition, there is no bias term, the stride is B and padding is 0. The shape of \mathbf{Y}_k and \mathbf{Y}_n are $h \times w \times \alpha_k B^2$ and $h \times w \times \alpha_n B^2$.

2.2. Hierarchical Interactive Reconstruction Subnet

Initial reconstruction: The initial reconstruction network consists of multiple branches, and each branch corresponds to the frame of each scale. We set the scale as S. For each branch of scale, the initial reconstruction includes three same steps, namely up-sampling, reshape and concatenation:

$$\mathbf{I} = cat(reshape(\mathbf{\Phi}_B \mathbf{Y})) \tag{5}$$

We utilize a convolutional layer with B^2 filters to learn the up-sampling matrix Φ_B , after which a series of vectors of

GOT 3 OF SIX CIT VI	aco sequi	chees with differ	tent es rados. E	est results are in	bola.		
Sequence	Ratio	KTSLR [8]	MC/ME [10]	VideoMH [7]	RRS [23]	VCSNet2 [14]	HIT-VCSNet
Akiyo		29.49/0.8836	27.15/0.8191	31.16/0.9161	25.09/0.7785	40.10/0.9834	41.59/0.9842
Coastguard		24.02/0.4902	21.65/0.4478	24.83/0.5536	22.25/0.4304	27.93/0.6775	30.75/0.8272
Foreman		24.29/0.7687	26.02/0.7504	27.84/0.8085	20.44/0.6471	28.89/0.8283	33.39/0.8949
Mother_daughter	0.01	30.72/0.8592	29.29/0.8174	32.59/0.8780	25.36/0.7142	39.80/0.9602	42.18/0.9741
Paris		20.60/0.5957	21.55/0.6616	20.64/0.6044	17.10/0.4143	26.55/0.8668	29.06/0.9263
Silent		25.93/0.7179	28.55/0.8327	27.22/0.7316	22.91/0.5908	34.71/0.9331	34.27/0.9300
Average		25.84/0.7192	25.70/0.7215	27.38/ 0.7487	22.19/0.5959	33.00/0.8749	35.21/0.9245
Akiyo		33.50/0.9328	38.77/0.9657	39.48/0.9693	41.45/0.9816	41.47/0.9815	43.59/0.9904
Coastguard		26.33/0.6222	28.09/0.7370	28.89/0.7814	29.35/0.7843	30.24/0.7876	33.13/0.9412
Foreman		28.35/0.8372	32.64/0.8803	33.94/0.8995	35.50/0.9323	34.28/0.9222	37.96/0.9786
Mother_daughter	0.1	33.82/0.9064	38.56/0.9449	38.93/0.9499	42.02/0.9719	41.43/0.9671	44.07/0.9879
Paris		23.04/0.7232	25.83/0.7778	25.58/0.7821	24.20/0.8213	28.23/0.9030	31.85/0.9819
Silent		29.22/0.8043	33.03/0.8841	32.69/0.8784	35.17/0.9162	36.44/0.9439	36.33/0.9428
Average		29.04/0.8043	32.82/0.8650	33.25/0.8767	34.61/0.9013	35.35/0.9176	37.82/0.9705

Table 1. Comparison with state-of-the-art video CS methods. The average results of PSNR in dB and SSIM on the first two GOPs of six CIF video sequences with different CS ratios. Best results are in bold.

size $1 \times 1 \times B^2$ are generated. Then a reshape layer is applied to convert each vector to a $B \times B \times 1$ block. Ultimately, a concatenation layer is applied to generate the whole frame.

Deep reconstruction: Our proposed framework jointly extracts features in both spatial and temporal domains. As depicted in Fig.2(a), MSEM extracts the multi-scale features and MSFM outputs the fused multi-scale feature \mathbf{F}_{MS} . The size of \mathbf{I}_1 , \mathbf{I}_2 and \mathbf{I}_3 are 32×32 , 16×16 and 8×8 . \mathbf{F}_1 , \mathbf{F}_2 and \mathbf{F}_3 correspond to the features of three different scales obtained after MSEM. Above all, a Feature Extraction Module(FEM) composed of a convolutional layer is applied for each branch to extract the features of the initial frames. Afterwards, numbers of Residual Blocks(RB) are employed to extract depth features. Several up-sampling sub-modules consist MSFM, which is composed of a deconvolution layer with stride 2 as well as two sets of ReLU and convolutional layers.

In the spatial domain, a Hierarchical Feature Fusion Module(HFFM) is implemented in MSEM to perceive richer contextual priors. As presented in Fig.2(b), for each fusion operation of HFFM at s-th scale, the fused feature \mathbf{F}_s^{j-1} from upper level and the feature \mathbf{F}_{os}^{j-1} passed among each scale are merged as \mathbf{M}^j (j represents the level of RB). Moreover, upsampling and down-sampling blocks are implemented due to the scale of the transmitting feature. The arrangement of the sub-blocks is equal to RB, which involves a skip connection consisting of convolutional layers with stride 1 and a ReLU shown in Fig.2(c). Specifically, to obtain the keyframe feature $\mathbf{F}_{(k,s)}^j$ for inter-frame fusion, \mathbf{M}^j will be passed to an up-sampling block.

In the temporal domain, HFIM is applied to interact hierarchical temporal information from keyframes in current and adjacent GOPs with non-keyframes:

$$\mathbf{F}_{(n,s)}^{(i,j)} = \mathbf{F}_{(k,s)}^{(1,j)} + \mathbf{F}_{(k,s)}^{(2,j)} + RB\left(\mathbf{F}_{(n,s)}^{(i,j-1)}\right)$$
(6)

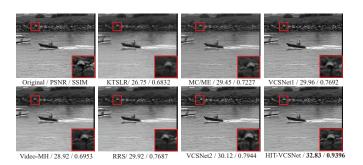


Fig. 3. Visual comparison on the 16^{th} frame of selected video sequence Coastguard when the sampling ratio is 0.1.

2.3. Loss Function

Given the initial and deep reconstruction of the keyframe and the non-keyframe I_k , D_k , I_n and D_n , the ground-truth keyframe and non-keyframe X_k and X_n , we design the end-to-end loss function for HIT-VCSNet as follows:

$$\mathcal{L} = \sum_{i=1}^{N} \begin{pmatrix} \left\| \mathbf{I}_{k}^{(i)} - \mathbf{X}_{k}^{(i)} \right\|_{2}^{2} + \left\| \mathbf{D}_{k}^{(i)} - \mathbf{X}_{k}^{(i)} \right\|_{2}^{2} \\ + \left\| \mathbf{I}_{k}^{(i+1)} - \mathbf{X}_{k}^{(i+1)} \right\|_{2}^{2} + \left\| \mathbf{D}_{k}^{(i+1)} - \mathbf{X}_{k}^{(i+1)} \right\|_{2}^{2} \\ + \sum_{k=1}^{K} \left(\left\| \mathbf{I}_{n}^{(i)} - \mathbf{X}_{n}^{(i)} \right\|_{2}^{2} + \left\| \mathbf{D}_{n}^{(i)} - \mathbf{X}_{n}^{(i)} \right\|_{2}^{2} \right) \end{pmatrix}$$
(7

where K refers to the total number of non-keyframes, and N represents the number of GOPs in the training dataset.

3. EXPERIMENT RESULT

3.1. Dataset and Implementation Details

For fair comparison, we use HEVC test video sequences in [14], which are divided into 128000 groups with data augmentation. Our HIT-VCSNet is trained end-to-end and the model training is performed on 4 GeForce RTX 3090 GPUs

Table 2. Comparison with deep learning-based image CS methods. The average results of PSNR in dB and SSIM on the first two GOPs of six CIF video sequences($\alpha_n = 0.1$). Best results are in bold.

Sequence	ISTA-Net ⁺ [20]	CSNet ⁺ [13]	SCSNet [12]	AMP-Net ⁺ [22]	OPINE-Net ⁺ [21]	HIT-VCSNet
Akiyo	34.83/0.9243	35.36/0.9463	35.56/0.9567	36.57/0.9758	36.83/0.9826	43.59/0.9904
Coastguard	27.23/0.6099	29.13/0.6946	29.33/0.7012	30.33/0.7251	30.61/0.7357	33.13/0.9412
Foreman	32.83/0.8795	32.38/0.9013	32.58/0.9121	33.56/0.9341	33.87/0.9461	37.96/0.9786
Mother_daughter	35.54/0.8956	37.18/0.9251	37.31/0.9310	38.36/0.9547	38.63/0.9663	44.07/0.9879
Paris	24.07/0.6893	24.66/0.7831	24.89/0.7931	25.83/0.8012	26.01/0.8276	31.85/0.9819
Silent	30.23/0.7932	31.82/0.8366	32.07/0.8451	33.06/0.8761	33.39/0.8792	36.33/0.9428
Average	30.79/0.8286	31.76/0.8478	31.96/0.8565	32.95/0.8778	33.22/0.8896	37.82/0.9705

by PyTorch. We set block size B=32, scale S=3. The GOP size and the batch size are set as 8 and 32 for 100 epochs. Specifically, we use Adam optimizer with the initial learning rate being 1×10^{-4} , which is reduced by half after every 30 epochs. α_k is set to 0.5, and α_n is set to 0.01 and 0.1 respectively. As for different sampling rates, the model of non-keyframe is trained independently. With regards to testing, 6 groups of standard CIF vedio sequences Akiyo, Coastguard, Foreman, Mother_daughter, Paris, Silent are applied. We utilize standard metrics (PSNR and SSIM [15]) for evaluation. Particularly, we transform color frames into YCbCr space and conduct operation merely for Y channel (i.e., luminance).

Table 3. The average results of PSNR on the first two GOPs compared with image CS methods ($\alpha_n = 0.01$).

Sequence	SCSNet [12]	AMP-Net ⁺ [22]	OPINE-Net ⁺ [21]	HIT-VCSNet
Average	26.38	27.75	28.83	35.21

Table 4. The average results of PSNR on the non-keyframes of the first two GOPs compared with image CS methods.

Ratio	ISTA-Net ⁺ [20]	CSNet ⁺ [13]	VCSNet2 [14]	HIT-VCSNet
0.01	20.14	24.14	31.95	33.95
0.1	29.34	30.53	34.63	36.56

3.2. Comparisons with State-of-the-Art Methods

We evaluate our HIT-VCSNet with state-of-the-art video CS methods, including KTSLR [8], MC/ME [10], Video-MH [7], RRS [23] and VCSNet2 [14]. Moreover, we compare five deep learning-based image CS methods with HIT-VCSNet, namely ISTA-Net [20], CSNet [13], SCSNet [12], AMPNet [22] and OPINE-Net [21]. As reported in Table 1-Table 4 and Fig.3, one can see that our HIT-VCSNet outperforms all the other methods. In particular, due to the tiny range of motion, the reconstruction effect of HIT-VCSNet is worse than VCSNet2 on the *silent* sequence, while our model owns predominant performance for large scale motion. Moreover, the

corresponding number of network parameters and time consumption of HIT-VCSNet are 4-5 times that of VCSNet2.

Table 5. Ablation study of HFIM and HFFM modules.

Ratio	w/o HFIM	w/o HFFM	HIT-VCSNet
0.01	33.96/0.8917	32.99/0.8792	35.21/0.9240
0.1	36.53/0.9372	35.72/0.9267	37.82/0.9705

3.3. Ablation Study

We retrain our model without HFIM and HFFM respectively, represented as "w/o HFIM" and "w/o HFFM". We evaluate the average PSNR and SSIM from the models for the first two GOPs of the 6 test video sequences. As shown in Table 5, HIT-VCSNet leads to a boost of 1.25dB and 2.22dB on PSNR at the sampling ratio of 0.01 compared with "w/o HFIM" and "w/o HFFM", which reflects the effectiveness of exploiting the deep priors in both spatial and temporal domains.

4. CONCLUSION

In this paper, we propose a novel Hierarchical InTeractive Video CS Reconstruction Network HIT-VCSNet, which can cooperatively exploit the deep priors in both spatial and temporal domains. Moreover, the hierarchical structure enables the proposed framework not only to hierarchically exploit richer contextual priors, but also to capture the interframe correlations more efficiently in the multi-scale space. Extensive experiments manifest that the proposed HIT-VCSNet outperforms the existing state-of-the-art video and image CS methods in a large margin.

Acknowledgments

This work is founded by National Natural Science Foundation of China (No.62076080), Natural Science Foundation of ChongQing CSTB2022NSCQ-MSX0922 and the Postdoctoral Science Foundation of Heilongjiang Province of China (LBH-Z22175).

¹Test videos are available at https://media.xiph.org/video/derf/

5. REFERENCES

- [1] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm with application to wavelet-based image deblurring. In 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 693–696, 2009.
- [2] E.J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.
- [3] Wenxue Cui, Feng Jiang, Xinwei Gao, Shengping Zhang, and Debin Zhao. An efficient deep quantized compressed sensing coding framework of natural images. *Proceedings of the 26th ACM international conference on Multimedia*, 2018. 1
- [4] Wenxue Cui, Shaohui Liu, Feng Jiang, and Debin Zhao. Image compressed sensing using non-local neural network. *IEEE Transactions on Multimedia*, pages 1–1, 2021. 1
- [5] Wenxue Cui, Shaohui Liu, and Debin Zhao. Fast hierarchical deep unfolding network for image compressed sensing. Proceedings of the 30th ACM International Conference on Multimedia, 2022. 1
- [6] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. 1
- [7] Ewtje Fowler. Video compressed sensing with multihypothesis. In *Data Compression Conference*, 2011. 1, 3,
- [8] S. G. Lingala, H. Yue, E. Dibella, and M. Jacob. Accelerated dynamic mri exploiting sparsity and low-rank structure: k-t slr. *IEEE Transactions on Medical Imaging*, 30(5):1042–1054, 2011. 1, 3, 4
- [9] Michael Lustig, David L. Donoho, and John M. Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine*, 58, 2007. 1
- [10] S. Mun and J. E. Fowler. Residual reconstruction for block-based compressed sensing of video. In *Data Com*pression Conference, 2011. 1, 3, 4
- [11] F. Ren, X. Kai, and Z. Zhang. Lapran: A scalable laplacian pyramid reconstructive adversarial network for flexible compressive sensing reconstruction, 2020.
- [12] Wuzhen Shi, Feng Jiang, Shaohui Liu, and Debin Zhao. Scalable convolutional neural network for image compressed sensing. In 2019 IEEE/CVF Conference

- on Computer Vision and Pattern Recognition (CVPR), pages 12282–12291, 2019. 4
- [13] Wuzhen Shi, Feng Jiang, Shengping Zhang, and Debin Zhao. Deep networks for compressed image sensing. In 2017 IEEE International Conference on Multimedia and Expo (ICME), pages 877–882, 2017. 4
- [14] Wuzhen Shi, Shaohui Liu, Feng Jiang, and Debin Zhao. Video compressed sensing using a convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 31, 2021. 1, 3, 4
- [15] Z. Wang. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004. 4
- [16] Zhengjue Wang, Hao Zhang, Ziheng Cheng, Bo Chen, and Xin Yuan. Metasci: Scalable and adaptive reconstruction for video compressive sensing. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2083–2092, 2021. 1
- [17] Zhuoyuan Wu, Jian Zhang, and Chong Mou. Dense deep unfolding network with 3d-cnn prior for snapshot compressive imaging. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 4872– 4881, 2021. 1
- [18] Jinjun Xu and Stanley Osher. Iterative regularization and nonlinear inverse scale space applied to wavelet-based denoising. *IEEE Transactions on Image Processing*, 16(2):534–544, 2007. 1
- [19] Jun Xu, Lei Zhang, and David Dian Zhang. A trilateral weighted sparse coding scheme for real-world image denoising. *ArXiv*, abs/1807.04364, 2018. 1
- [20] J. Zhang and B. Ghanem. Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 4
- [21] Jian Zhang, Chen Zhao, and Wen Gao. Optimizationinspired compact deep compressive sensing. *IEEE Jour*nal of Selected Topics in Signal Processing, 14(4):765– 774, 2020. 4
- [22] Zhonghao Zhang, Y. Liu, Jiani Liu, Fei Wen, and Ce Zhu. Amp-net: Denoising-based deep unfolding for compressive image sensing. *IEEE Transactions on Im*age Processing, 30:1487–1500, 2021. 4
- [23] 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):1–1, 2017. 1, 3, 4