Supplementary Material of Joint Spectral Image Reconstruction and Semantic Segmentation with Cooperative Unfolding

Zijun He[1,2] Ping Wang[2*] Xiaodong Wang[1,2] Chang Chen[2,3] Xin Yuan[2*]
[1]Zhejiang University [2]Westlake University [3]Westlake Intelligent Vision
{hezijun, wangping, wangxiaodong, chenchang, xyuan}@westlake.edu.cn

1. Detailed derivation of Eq. (13) in the Main Paper

In the main paper, we solve the $\boldsymbol{z}$-problem:

$$\boldsymbol{z}_{k+1} = \text{argmin}_{\boldsymbol{z}} \frac{1}{2}||\boldsymbol{y}-\mathbf{A}\boldsymbol{z}||_2^2+\frac{1}{2}||\boldsymbol{z}-\boldsymbol{\Phi\theta}_k||_2^2+\frac{\beta}{2}||\boldsymbol{z}-\boldsymbol{x}_k||_2^2, \tag{1}$$

by:

$$\boldsymbol{z}_{k+1} = \tilde{\boldsymbol{x}}_k+\mathbf{A}^\top(\boldsymbol{y}-\mathbf{A}\tilde{\boldsymbol{x}}_k)\oslash(1+\beta+\text{Diag}(\mathbf{AA}^\top)), \tag{2}$$

where $\tilde{\boldsymbol{x}}_k = \frac{\beta\boldsymbol{x}_k+\boldsymbol{\Phi\theta}_k}{1+\beta}$. Next, we provide the detailed derivation of Eq. (2).

Let $\boldsymbol{g}(\boldsymbol{z}) = \frac{1}{2}||\boldsymbol{y}-\mathbf{A}\boldsymbol{z}||_2^2+\frac{1}{2}||\boldsymbol{z}-\boldsymbol{\Phi\theta}_k||_2^2+\frac{\beta}{2}||\boldsymbol{z}-\boldsymbol{x}_k||_2^2$, we have:

$$\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{z}} = -\mathbf{A}^\top(\boldsymbol{y}-\mathbf{A}\boldsymbol{z})+(\boldsymbol{z}-\boldsymbol{\Phi\theta}_k)+\beta(\boldsymbol{z}-\boldsymbol{x}_k). \tag{3}$$

Setting Eq. (3) to zero yields the closed-form solution:

$$\begin{aligned}\boldsymbol{z}_{k+1} &= (\mathbf{A}^\top\mathbf{A}+(1+\beta)\mathbf{I})^{-1}(\mathbf{A}^\top\boldsymbol{y}+\boldsymbol{\Phi\theta}_k+\beta\boldsymbol{x}_k)\\ &= (\mathbf{A}^\top\mathbf{A}+\hat{\beta}\mathbf{I})^{-1}(\mathbf{A}^\top\boldsymbol{y}+\hat{\beta}\tilde{\boldsymbol{x}}_k),\end{aligned} \tag{4}$$

where $\mathbf{I}$ denotes the identity matrix with desired dimensions and $\hat{\beta} = 1+\beta$. Since $\mathbf{A}$ is a fat matrix, $\mathbf{A}^\top\mathbf{A}+\hat{\beta}\mathbf{I}$ will be large and thus we can simplify the solution by the matrix inversion formula as:

$$(\mathbf{A}^\top\mathbf{A}+\hat{\beta}\mathbf{I})^{-1} = \hat{\beta}^{-1}\mathbf{I}-\hat{\beta}^{-1}\mathbf{A}^\top(\mathbf{I}+\mathbf{A}\hat{\beta}^{-1}\mathbf{A}^\top)^{-1}\mathbf{A}\hat{\beta}^{-1}. \tag{5}$$

By plugging Eq. (5) into Eq. (4), we can reformulate Eq. (4) as:

$$\begin{aligned}\boldsymbol{z}_{k+1} =& \frac{\mathbf{A}^\top\boldsymbol{y}+\hat{\beta}\tilde{\boldsymbol{x}}_k}{\hat{\beta}}-\frac{\mathbf{A}^\top(\mathbf{I}+\mathbf{A}\hat{\beta}^{-1}\mathbf{A}^\top)\mathbf{A}\mathbf{A}^\top\boldsymbol{y}}{\hat{\beta}^2}\\ &-\frac{\mathbf{A}^\top(\mathbf{I}+\mathbf{A}\hat{\beta}^{-1}\mathbf{A}^\top)^{-1}\mathbf{A}\tilde{\boldsymbol{x}}_k}{\hat{\beta}}\end{aligned} \tag{6}$$

In CASSI systems, $\mathbf{AA}^\top$ is a diagonal matrix which can be defined as $\mathbf{AA}^\top = \text{Diag}\{a_1,\dots,a_n\}$. By plugging $\mathbf{AA}^\top$ into $(\mathbf{I}+\mathbf{A}\hat{\beta}^{-1}\mathbf{A}^\top)^{-1}$ and $(\mathbf{I}+\mathbf{A}\hat{\beta}^{-1}\mathbf{A}^\top)^{-1}\mathbf{AA}^\top$, we obtain:

$$(\mathbf{I}+\mathbf{A}\hat{\beta}^{-1}\mathbf{A}^\top)^{-1} = \text{Diag}\{\frac{\hat{\beta}}{\hat{\beta}+a_1},\dots,\frac{\hat{\beta}}{\hat{\beta}+a_n}\}, \tag{7}$$

$$(\mathbf{I}+\mathbf{A}\hat{\beta}^{-1}\mathbf{A}^\top)^{-1}\mathbf{AA}^\top = \text{Diag}\{\frac{\hat{\beta}a_n}{\hat{\beta}+a_1},\dots,\frac{\hat{\beta}a_n}{\hat{\beta}+a_n}\}, \tag{8}$$

Let $\boldsymbol{y} = [y_1,\dots,y_n]^\top$ and $[\mathbf{A}\boldsymbol{z}_k]_i$ denotes the i-th element of $\mathbf{A}\boldsymbol{z}_k$. We plug Eq. (7) and Eq. (8) into Eq. (6) as:

$$\begin{aligned}\boldsymbol{z}_{k+1} &= \frac{\mathbf{A}^\top y}{\hat{\beta}}+\tilde{\boldsymbol{x}}_k-\frac{1}{\hat{\beta}}\mathbf{A}^\top[\frac{y_1a_1+\hat{\beta}[\mathbf{A}\tilde{\boldsymbol{x}}_k]_1}{\hat{\beta}+a_1},\dots,\frac{y_na_n-\hat{\beta}[\mathbf{A}\tilde{\boldsymbol{x}}_k]_n}{\hat{\beta}+a_n}]\\ &= \tilde{\boldsymbol{x}}_k+\mathbf{A}^\top[\frac{y_1-[\mathbf{A}\tilde{\boldsymbol{x}}_k]_1}{\hat{\beta}+a_1},\dots,\frac{y_n-[\mathbf{A}\tilde{\boldsymbol{x}}_k]_n}{\hat{\beta}+a_n}]\\ &= \tilde{\boldsymbol{x}}_k+\mathbf{A}^\top(\boldsymbol{y}-\mathbf{A}\tilde{\boldsymbol{x}}_k)\oslash(1+\beta+\text{Diag}(\mathbf{AA}^\top)),\end{aligned} \tag{9}$$

and the Eq. (2) is proved.

2. Details of FVgNET

2.1. Statistics of FVgNET

As shown in Tab. 1, we report the number of samples in each category. We ensured, as much as possible, that the proportions of each category in the training and test sets were relatively balanced.

2.2. More Sample Visualization

As shown in Fig. 1, we present 24 samples in FVgNET [4]. FVgNET includes scenarios with only one type of true or false item, as well as scenarios with multiple types of items.

To make it easier to distinguish real and fake objects across categories, we list the corresponding label colors in Tab. 2.

3. Additional Ablation Study

Ablation Study on the Attention Scheme. We conducted an ablation study on the attention scheme of the decoder in the reconstruction and segmentation modules of CRSDUN. All the experiments are conducted in CRSDUN-3stg, and the results are shown in Tab. 3 (The left and right sides

*Corresponding authors.

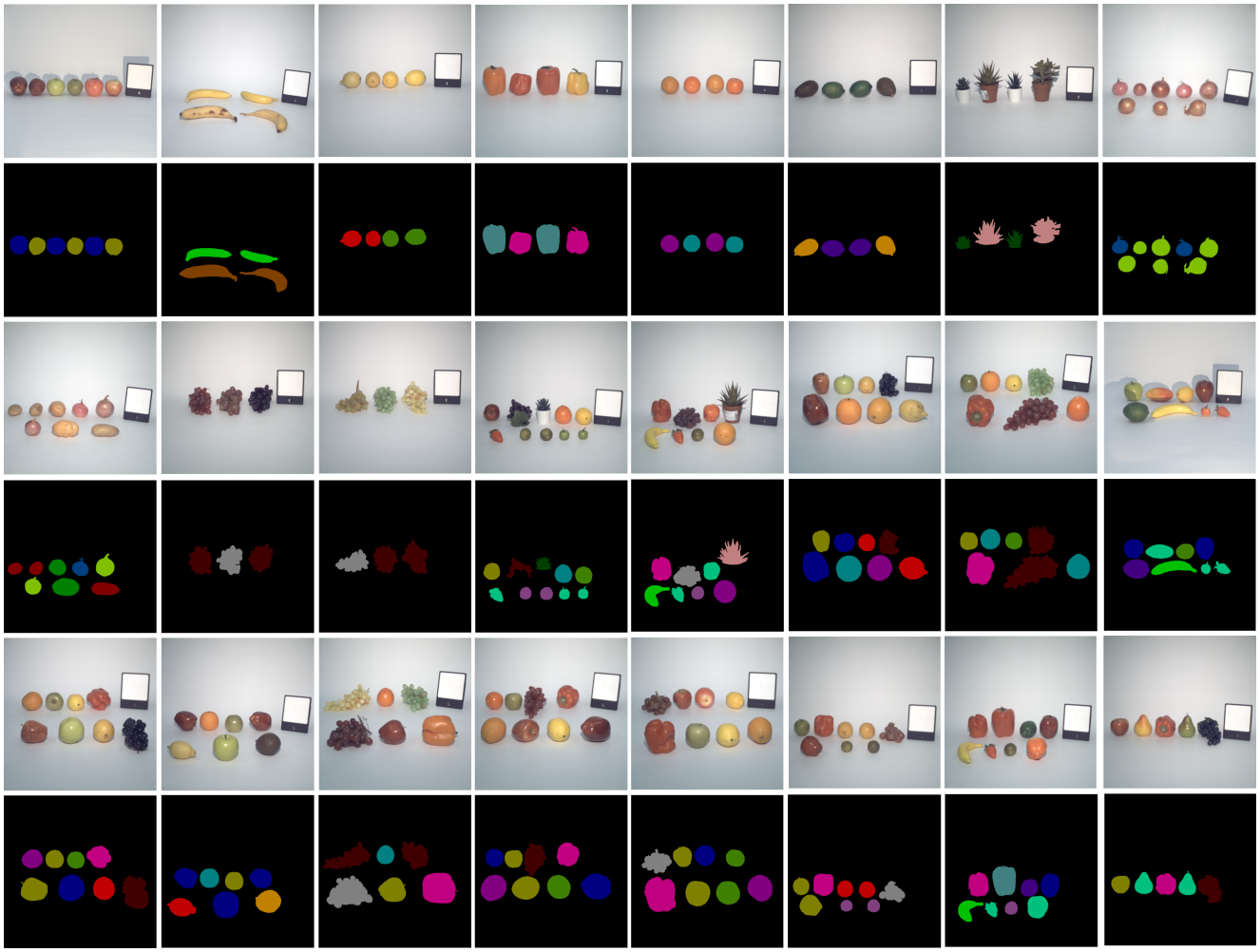

Figure 1. Examples of the FVgNET dataset.

Table 1. Category count statistics in the FVgNET Dataset.

Item	Training Set		Testing Set	
	Real	Fake	Real	Fake
Apple	110	85	24	20
Avocado	25	22	8	6
Banana	37	25	6	5
Grape	64	45	7	7
Lemon	64	50	11	9
Onion	41	30	13	6
Orange	53	43	12	7
Pepper	95	50	17	11
Plant	56	55	12	21
Potato	40	29	10	9
Unknown	42	9	5	3

Table 2. Colors of the labels corresponding to each category in the FVgNET dataset

Item	Real		Fake	
	Color	RGB value	Color	RGB value
Apple		(128, 128, 0)		(0, 0, 128)
Avocado		(192, 128, 0)		(64, 0, 128)
Banana		(128, 64, 0)		(0, 192, 0)
Grape		(128, 128, 128)		(64, 0, 0)
Lemon		(192, 0, 0)		(64, 128, 0)
Onion		(128, 192, 0)		(0, 64, 128)
Orange		(128, 0, 128)		(0, 128, 128)
Pepper		(192, 0, 128)		(64, 128, 128)
Plant		(192, 128, 128)		(0, 64, 0)
Potato		(128, 0, 0)		(0, 128, 0)
Unknown		(128, 64, 128)		(0, 192, 128)

of + represent the attention mechanisms used in the reconstruction and segmentation modules, respectively.). It can be observed that, compared to Local-Window Self-Attention (LWSA) [3] or Window Spectral Self-Attention

(WSSA) [6], CASTA can effectively improve reconstruction and segmentation performance.

Table 3. Comparison of different attention schemes.

Method	PSNR (dB)	mIoU (%)	Params (M)	FLOPs (G)
WSSA+LWSA	38.03	86.55	**3.96**	60.91
WSSA+CASTA	37.98	87.73	3.97	**59.45**
LWSA+CASTA	38.81	89.93	4.04	59.60
CASTA+CASTA	**39.35**	**90.11**	4.02	59.95

Ablation Study on the Loss Function. We first perform an ablation study on the multi-stage loss strategy. The results are shown in Tab. 4. If we only constrain the output of the last stage of our CRSDUN-3stg, this results in a 1.24 dB degradation in PSNR and a 1.90% drop in mIoU.

Table 4. Ablation study on the multi-stage loss strategy.

loss	multi-stage loss	single-stage loss
PSNR (dB)	39.35	38.11
mIoU (%)	90.11	88.21

In addition, we performed an ablation study on λ_{ce}, *i.e.*, the coefficient used to balance the reconstruction and segmentation losses. The results are shown in Tab. 5. When λ_{ce} is 10^{-3}, the segmentation performance is the best, but it leads to suboptimal reconstruction performance, when λ_{ce} is 10^{-5}, the segmentation performance is poor, and when λ_{ce} is 10^{-4}, the performance of reconstruction and segmentation is relatively good.

Table 5. Ablation study on the λ_{ce}.

λ_{ce}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
PSNR (dB)	36.02	38.47	39.35	**39.65**
mIoU (%)	87.64	**91.03**	90.11	82.52

4. Additional Results

4.1. Comparison with more SOTA Algorithms

Table 6 reports the quantitative comparison of our CRSDUN and several additional state-of-the-art (SOTA) methods: RDULF-9stg+Seg [1], PADUT-9stg-Plus+Seg [2], SPECAT+Seg [5]. Our CRSDUN-5stg surpasses RDLUF-9stg+Seg, PADUT-9stg+Seg, SPECAT+Seg by 3.31 dB, 4.83 dB and 7.64 dB in PSNR, 6.30%, 18.24% and 14.49% in mIoU.

4.2. Comparison with Segmentation from Clean HSI

We further compare our pipeline that segmentation from CASSI measurements with segmentation directly from the clean HSIs. As shown in Tab. 7, when using CRSDUN, our pipeline achieves segmentation performance comparable to that obtained from clean HSIs. However, it is worth noting that our pipeline does not require time-consuming scanning.

Table 6. Comparison of reconstruction and segmentation performance of our CRSDUN and three other two-stage algorithms in 50 testing scenes of FVgNET [4] ("+Seg" is omitted).

Algorithms Reference	RDLUF-9stg CVPR' 23 [1]	PADUT-9stg ICCV' 23 [2]	SPECAT CVPR' 24 [5]	CRSDUN-5stg Ours
PSNR (dB)	36.57	35.05	32.24	**39.88**
SSIM	0.963	0.957	0.886	**0.977**
Precision (%)	91.99	84.56	87.04	**95.65**
Recall (%)	92.77	85.46	86.91	**96.11**
F1-score (%)	92.10	84.07	86.44	**95.86**
mIoU (%)	86.03	74.09	77.84	**92.33**
Paras (M)	6.74	9.33	**3.28**	6.73
FLOPs (G)	200.1	135.0	**49.79**	99.07

Table 7. Comparison of segmentation performance of coded measurement and clean HSI.

Input	Algorithm	Precision (%)	Recall (%)	F1-score (%)	mIoU (%)
CASSI Mea	CRSDUN-5stg	95.65	96.11	95.86	92.33
Clean HSI	SwinTransformer	96.01	96.03	95.92	92.54

4.3. More Visual Results

Figs. 2-4 present additional visual comparisons against several two-stage methods built upon SOTA HSI reconstruction algorithms, including SPECAT [5], PADUT [2], RDLUF [1], RCUMP [7] and SSR [6]. As shown in Figs. 2-3. It can be clearly observed that CRSDUN-5stg yields cleaner reconstructions, fewer artifacts, and more accurate spectra. Besides, Fig. 4, demonstrate that our CRSDUN-5stg achieve the best segmentation accuracy.

References

[1] Yubo Dong, Dahua Gao, Tian Qiu, Yuyan Li, Minxi Yang, and Guangming Shi. Residual degradation learning unfolding framework with mixing priors across spectral and spatial for compressive spectral imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22262–22271, 2023. 3

[2] Miaoyu Li, Ying Fu, Ji Liu, and Yulun Zhang. Pixel adaptive deep unfolding transformer for hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12968, 2023. 3

[3] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[4] Maksim Makarenko, Arturo Burguete-Lopez, Qizhou Wang, Fedor Getman, Silvio Giancola, Bernard Ghanem, and Andrea Fratalocchi. Real-time hyperspectral imaging in hardware via trained metasurface encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12692–12702, 2022. 1, 3

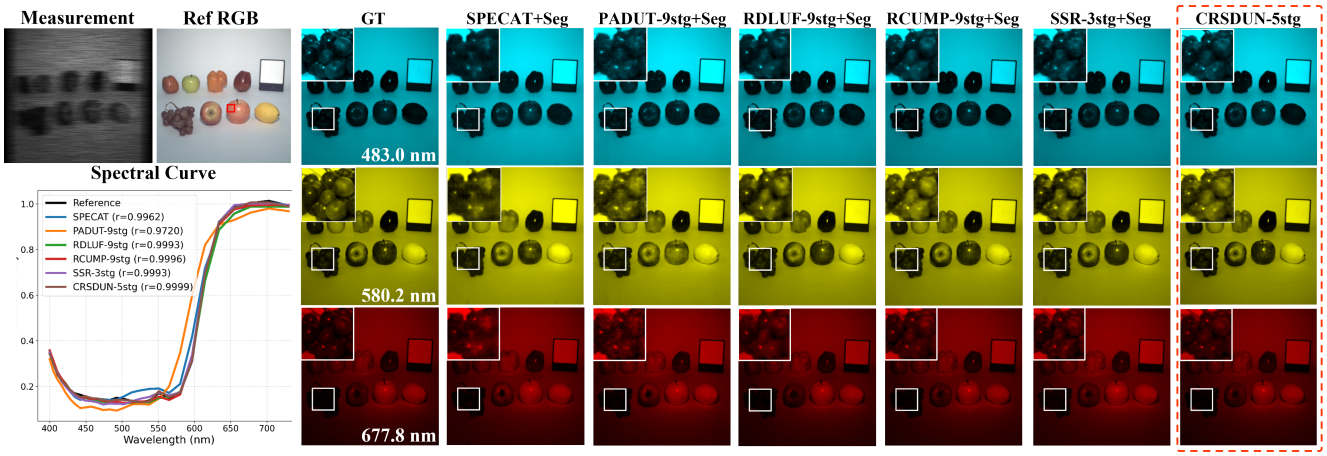

Figure 2. Visualization of reconstruction result obtained by different algorithms in one of the testing data. Zoom in for a better view.

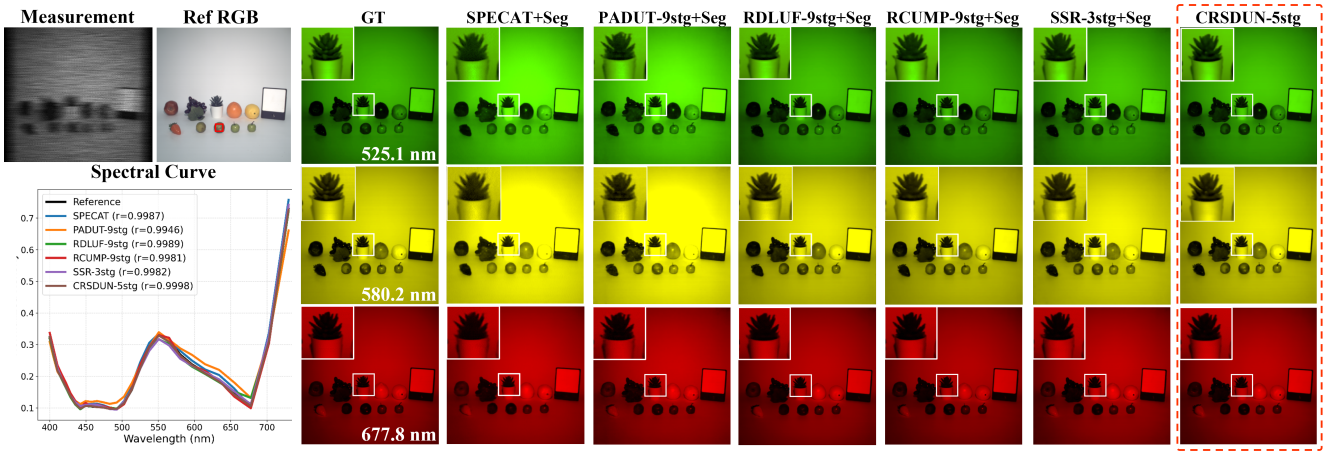

Figure 3. Visualization of reconstruction result obtained by different algorithms in one of the testing data. Zoom in for a better view.

[5] Zhiyang Yao, Shuyang Liu, Xiaoyun Yuan, and Lu Fang. Specat: Spatial-spectral cumulative-attention transformer for high-resolution hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25368–25377, 2024. 3

[6] Jiancheng Zhang, Haijin Zeng, Yongyong Chen, Dengxiu Yu, and Yin-Ping Zhao. Improving spectral snapshot reconstruction with spectral-spatial rectification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25817–25826, 2024. 3

[7] Yin-Ping Zhao, Jiancheng Zhang, Yongyong Chen, Zhen Wang, and Xuelong Li. Rcump: Residual completion unrolling with mixed priors for snapshot compressive imaging. *IEEE Transactions on Image Processing*, 33:2347–2360, 2024. 3

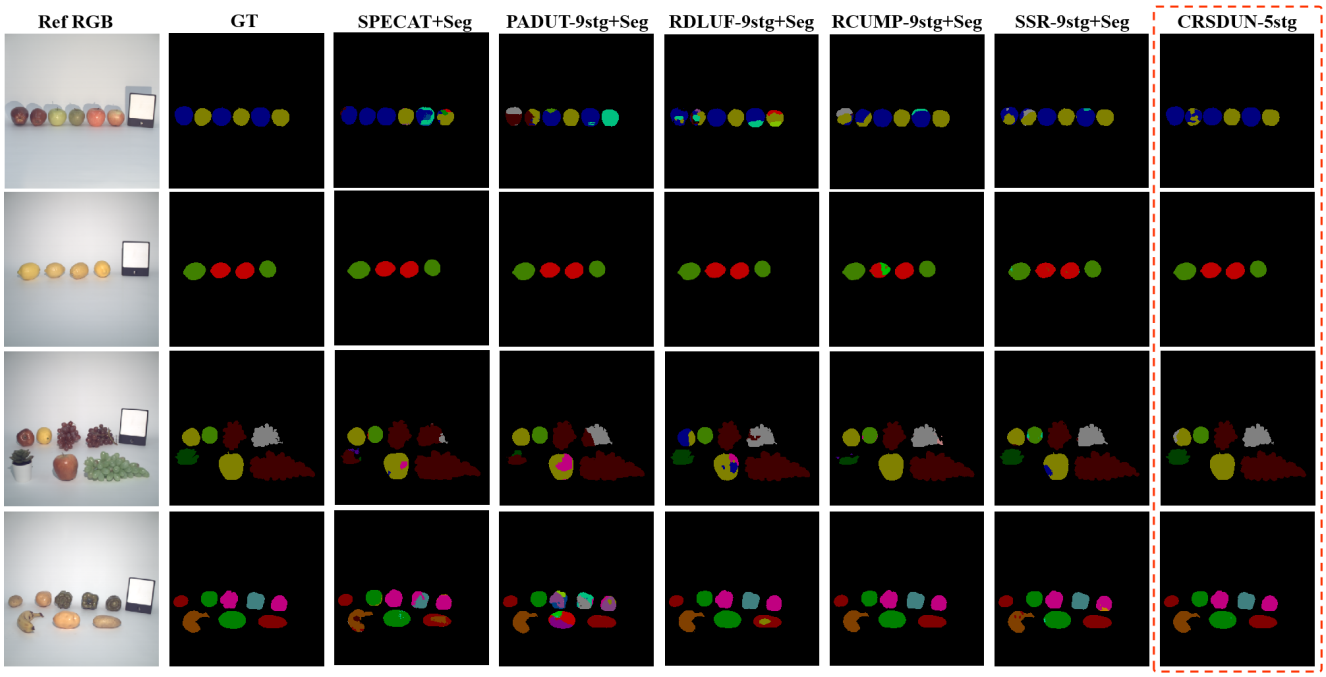

Figure 4. Visualization of segmentation maps obtained by different algorithms in 4 scenes of the testing dataset. Zoom in for a better view.