Exploring Spatiotemporal Feature Propagation for Video-Level Compressive Spectral Reconstruction: Dataset, Model and Benchmark

Supplementary Material

This supplementary material presents the principles of the DD-CASSI system along with prototype hardware specifications, additional quantitative comparisons, reconstruction results under real-world conditions, additional ablation studies and a discussion of method limitations. The content is organized as follows:

- S1 **DD-CASSI and Prototype Configuration**;
- S2 **Generalization Stress Tests**;
- S3 **Quantitative Comparisons with SOTA Methods**;
- S4 **Real-World Scenes and Reconstruction Results**;
- S5 **Ablation Study;**
- S6 **Limitations**;
- S7 **Comparison of Video Reconstruction Quality**.

S1. Prototype Configuration

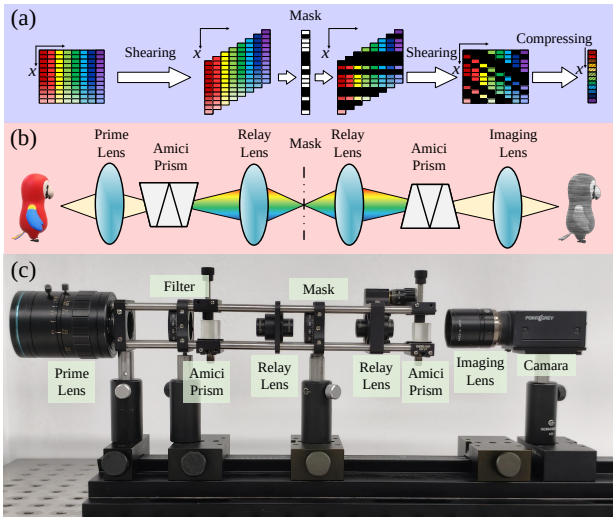


Figure S1. Principle and prototype system of DD-CASSI. (a) Encoding process of the spatial-spectral signal. (b) Optical layout of the DD-CASSI. (c) The prototype system.

Principle. The encoding principle is illustrated in Fig. S1(a), and the optical path layout is shown in Fig. S1(b). The system first disperses the incoming light using a dispersive element, introducing spectral shearing across spatial positions. A random coded mask is then applied to modulate the dispersed signal. Subsequently, another dispersive element performs inverse dispersion, aligning the spectral components back to their original spatial locations. The camera captures the resulting two-dimensional compressed measurement, which represents the spatial-spectral encoded signal. Due to the use of symmetric dispersive elements, the spatial dimensions of the measurement remain consistent with those of the original scene.

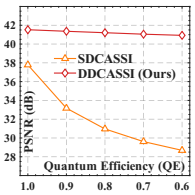
Hardware. As illustrated in Fig. S1(c), the hardware system consists of a 25 mm prime lens (Myutron HF5018V F1.8), a custom-designed random coded aperture (mask), two 30 mm relay lenses (Edmund 45762), a band-pass filter composed of a 500 nm long-pass and a 650 nm short-pass filter (LBTEK MEFH10-500LP and MEFH10-650SP), two custom-designed Amici prisms, and a detector (Point Grey GS3-U3-41S4M). The detector has a resolution of 2016×2016 pixels with the pixel pitch of $3.1\,\mu m$ and supports 14-bit depth.

Wavelength. Two edgepass filters are used to limit the spectral range of the system to $500 - 650\ nm$. This spectral range, under the dispersion introduced by the Amici prisms, results in 58-pixels dispersion on the detector. By calibrating the wavelength-dependent dispersion function, we determine the 30 spectral bands with channels with wavelengths: {500.0, 503.2, 506.5, 510.0, 513.5, 517.1, 520.7, 524.5, 528.4, 532.3, 536.4, 540.6, 544.9, 549.3, 553.9, 558.6, 563.5, 568.5, 573.8, 579.2, 584.8, 590.6, 596.7, 603.0, 609.6, 616.5, 623.6, 631.1, 639.0, 647.2} nm.

Mask capture. The mask patterns corresponding to each spectral band are recorded using monochromatic illumination produced by a monochromator (Zolix Omni-λ3005i) at the previously defined wavelengths. To eliminate the influence of source intensity variation, we further calibrate the optical power at each wavelength using a power meter (Thorlabs S120C). This ensures that the resulting mask matrix reflects only system-dependent modulation. The physical size of the minimum mask feature is designed to be twice the detector pixel pitch (i.e., $6.2\,\mu m$). However, due to the diffraction limit and off-axis aberrations, the smallest effective feature projected onto the detector plane slightly exceeds 2×2 pixels.

S2. Generalization Stress Tests

Architectural Selection. To address concerns regarding the generalization of our architectural selection, we conduct comprehensive stress tests on SD-CASSI and DD-CASSI under varying physical and environmental conditions. These experiments evaluate whether the superiority of DD-CASSI holds under different noise levels and spectral configurations. We evaluated the reconstruction performance under varying Quantum Efficiency (QE) levels and extended spectral ranges. As illustrated in Fig. S2, while the performance of



Method	Metric	450-650nm	Scaling	Rotation
DD-CASSI	PSNR↑	39.31	41.38	41.25
	SSIM↑	0.9828	0.9891	0.9889
SD-CASSI	PSNR↑	36.09	37.61	37.24
	SSIM↑	0.9551	0.9692	0.9670

Figure S2. Generalization stress test under different noise, non-rigid motion and spectral settings.

SD-CASSI drops significantly as noise increases (lower QE), DD-CASSI exhibits remarkable robustness with minimal performance degradation. Specifically, DD-CASSI maintains a PSNR above 40 dB even under challenging noise conditions, whereas SD-CASSI's performance falls below 30 dB. This demonstrates that DD-CASSI's advantage in spatial structural preservation is particularly beneficial in low-SNR regimes.

Generalization to Unseen Non-Rigid Motions. To address the concern that the idealized motion statistics in DynaSpec might limit real-world applicability, we challenged the trained model with unseen non-rigid motions, including complex Scaling and Rotation patterns. As illustrated in Fig. S2, our model demonstrates strong generalization and effectively mitigates domain shift concerns by achieving impressive PSNR results of 41.38 dB for scaling and 41.25 dB for rotation, respectively. These results prove that the spatiotemporal feature propagation mechanism in PG-SVRT successfully learns robust motion representations that extend to unscripted, real-world dynamics.

S3. Quantitative Comparisons

Video Restoration Methods. At present, the spectral reconstruction literature lacks algorithms specifically tailored for video-level tasks. To this end, we select two representative general-purpose video restoration methods TempFormer [1] and VRT [2] as video baselines for comparative analysis. TempFormer transforms features into the frequency domain and leverages its priors for spectral reconstruction. However, the essence of SCI is spatial modulation rather than frequency-domain encoding, leading to a principled mismatch that limits its ability to model this task. By contrast, VRT operates in the spatiotemporal domain and accelerates cross-frame feature fusion via optical-flow mechanisms, yielding competitive results on generic video restoration; nevertheless, lacking mechanisms to decouple compressively encoded data, its performance on spectral reconstruction remains constrained.

From a quantitative perspective, as shown in Fig. S3(Left), PG-SVRT significantly outperforms TempFormer and VRT across all metrics, including PSNR, SSIM, SAM, and STRRED. On the KAIST test set, PG-SVRT attains a PSNR of 41.23 dB, clearly surpassing VRT (39.18 dB) and TempFormer (37.43 dB). On the DynaSpec test set, it likewise leads with a PSNR of 41.82 dB, approximately 1.8 dB higher than VRT and over 2.5 dB higher than TempFormer. In terms of SAM, PG-SVRT achieves the lowest spectral reconstruction error, indicating more accurate recovery of spectral details. For temporal consistency measured by STRRED, PG-SVRT also yields markedly better scores than competing methods, demonstrating superior cross-frame stability and mitigating the flickering artifacts commonly observed in video-level reconstruction.

Additionally, Fig. S3(Right) presents a comparison of PSNR, Params, and GFLOPs between all algorithms evaluated in the main paper and in this section. Notably, PG-SVRT delivers superior performance with substantially lower FLOPs and parameter counts, especially relative to video restoration methods, making it a more cost-effective solution for video-level compressive spectral reconstruction.

Motion Generalization Analysis. To further assess the model's generalization ability under unknown motion patterns, we evaluate the performance of PG-SVRT under transformations such as rotation and scaling, which were not included in the training strategy. Such transformations are common in real-world dynamic scenes, such as camera rotations or gradual approach to a scene. Although these motion patterns were not explicitly included during training, PG-SVRT implicitly develops some adaptability during the learning process, thanks to the rich motion characteristics in the DynaSpec dataset. We applied all trained models to test scenes with additional rotation and scaling transformations and compared their performance across various metrics. Table S1 presents the test results of each method. From the results, it is evident that, compared to image-level methods, PG-SVRT still demonstrates significant advantages in both types of tests: the PSNR in the rotation scenario reaches 41.25 dB, and in the scaling scenario, it reaches 41.38 dB, outperforming all comparison methods and even surpassing the DPU* method, which incorporates temporal information. This indicates that PG-SVRT can maintain high reconstruction quality even under unseen motion patterns.

Therefore, PG-SVRT, while capturing cross-frame spatiotemporal dependencies, effectively leverages the feature redundancy in dynamic scenes, showing good adaptability to complex motions such as rotation and scaling. This demonstrates that, even when faced with motion patterns not encountered during training, the model retains strong generalization ability, ensuring the stability and reliability of spectral video reconstruction.

S4. Real-World Reconstruction Results

As shown in the Fig. S4, we capture five real-world sequences using the constructed DD-CASSI prototype system to validate the proposed method under practical conditions. The scenes are arranged in order of increasing complexity:

1. Translation of a gray book.

Method	TempFormer ECCV 2022	VRT TIP 2024	PG-SVRT Ours
PSNR-K ↑	37.43	39.18	**41.23**
SSIM-K ↑	0.9782	0.9804	**0.9882**
SAM-K ↓	5.4748	6.0422	**3.805**
ST-RRED-K ↓	58.73	40.44	**19.35**
PSNR-D ↑	39.27	40.00	**41.82**
SSIM-D ↑	0.9840	0.9874	**0.9904**
SAM-D ↓	5.0714	5.5497	**4.0118**
ST-RRED-D ↓	67.33	35.08	**27.14**
Params	68.63 M	5.89 M	**2.48 M**
GFLOPs	376.94	161.27	**28.18**
Infer. Time(ms)	47.79	52.76	**34.86**

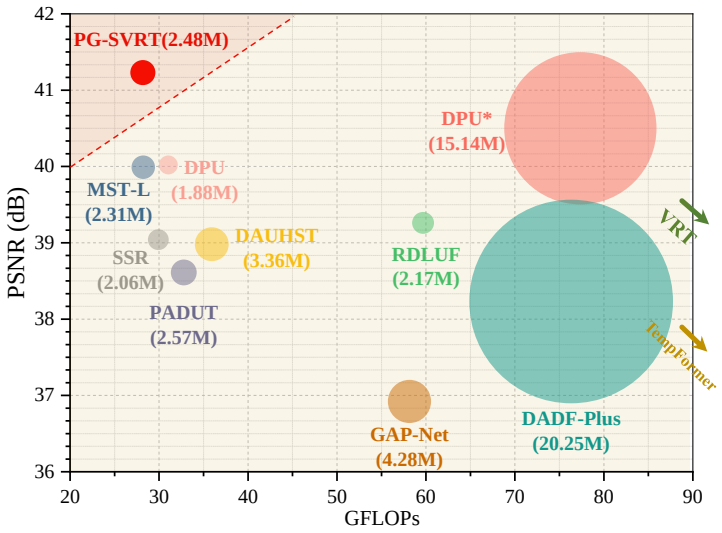


Figure S3. **(Left)** Quantitative comparisons between video restoration methods and PG-SVRT. The suffix "-K" denotes results on KAIST, whereas "-D" indicates evaluations on the DynaSpec test set. **(Right)** PSNR–FLOPs–Params trade-off comparing the proposed PG-SVRT with recent state-of-the-art methods for HSIs reconstruction and video restoration.

Motion	Method	MST-L CVPR 2022	CST-L ECCV 2022	DAUHST NeurIPS 2022	GAP-Net IJCV 2023	DADF-Plus-3 TMI 2023	RDLUF CVPR 2023	PADUT ICCV 2023	S^2-Transfor. TPAMI 2024	SSR CVPR 2024	DPU CVPR 2024	DPU* CVPR 2024	PG-SVRT Ours
Rotation	PSNR-AVE ↑	39.27	39.61	39.45	38.06	38.27	38.82	39.23	34.97	39.05	40.28	**40.66**	**41.25**
	SSIM-AVE ↑	0.9865	0.9863	0.9852	0.9800	0.9837	0.9851	0.9849	0.9694	0.9852	**0.9870**	0.9868	**0.9889**
	SAM-AVE ↓	**4.1555**	4.3804	5.2142	5.7556	4.8157	4.4967	4.6662	7.1241	5.0250	4.8473	4.8271	**3.9537**
	ST-RRED-AVE ↓	47.46	42.81	45.25	71.53	62.66	56.08	53.06	129.10	48.95	31.89	**31.32**	**24.50**
Scalling	PSNR-AVE ↑	39.64	39.76	39.48	38.10	38.41	39.04	39.34	35.22	39.21	40.40	**40.73**	**41.37**
	SSIM-AVE ↑	0.9872	0.9866	0.9853	0.9802	0.9844	0.9855	0.9852	0.9703	0.9856	**0.9872**	0.9868	**0.9891**
	SAM-AVE ↓	**4.1026**	4.3812	5.2410	5.8019	4.8222	4.5067	4.6303	6.9664	5.0250	4.8935	4.9306	**3.9712**
	ST-RRED-AVE ↓	52.09	54.05	54.22	84.06	77.60	66.13	56.49	125.65	58.76	40.30	**35.58**	**31.90**

Table S1. Generalization analysis of the method: A comparison of reconstruction performance between PG-SVRT and spectral reconstruction methods is conducted. The test scenes involve rotation and scaling transformations that were not encountered during training. The suffix "-AVE" denotes the average results on the KAIST and DynaSpec test sets.

Scene 1 Scene 2 Scene 3 Scene 4 Scene 5

Figure S4. Visualization of real-world measurements captured by the DD-CASSI prototype system.

2. Rotation and translation of three toys.
3. Movement of a mechanical crab with high-degree-of-freedom leg motion.
4. Rotation of a LEGO-assembled flower.
5. A toy-simulated complex street scene with two vehicles moving toward each other.

The reconstructed spectral video sequences by PG-SVRT are shown in Figures S6 to S10 (see end of the paper).

S5. Ablation Study

Table S3. Ablation on the number of frames

N Frames	PSNR	SSIM	SAM	ST-RRED	GFLOPs	Params
1	40.096	0.9879	3.9575	46.70	**27.24**	**2.42 M**
2	41.518	0.9892	3.9798	24.89	28.12	2.48 M
3	**41.524**	**0.9893**	**3.9084**	**23.25**	28.18	2.48 M
4	41.030	0.9883	4.2329	33.05	28.24	2.48 M
5	41.003	0.9886	4.0766	33.21	28.30	2.48 M

Comparison of Attention Mechanisms. As shown in Tab. S2, we compare several attention mechanisms in terms of computational complexity and performance in video-level spectral reconstruction. G-MSA [3] and F-MSA [4], which lack any windowing mechanism, exhibit prohibitively high computational costs and are infeasible to run on our GPU due to memory constraints. For W-MSA [2], we set the temporal window size to 2. In addition, we consider a variant in which the spatiotemporal window is reduced to a purely spatial one, referred to as S-W-MSA. Both W-MSA and S-W-MSA employ joint spatiotemporal attention, capturing

Table S2. Comparison of different attention mechanisms in terms of computational complexity and reconstruction performance.

	Computational Complexity	PSNR	SSIM	SAM	ST-RRED	GFLOPs	Params
G-MSA	$4THWC^2 + 2(THW)^2C$	-	-	-	-	-	-
F-MSA	$8THWC^2 + 2T^2HWC + 2T(HW)^2C$	-	-	-	-	-	-
W-MSA	$4THWC^2 + 2THW(T_{win}H_{win}W_{win})C$	41.36	0.9887	4.0562	26.68	51.34	**1.85 M**
S-W-MSA	$4THWC^2 + 2T^2HW(H_{win}W_{win})C$	41.44	0.9889	4.0511	23.93	52.57	1.89 M
W-F-MSA	$8THWC^2 + 2T^2HWC + 2THW(H_{win}W_{win})C$	41.08	0.9880	4.4128	26.40	30.16	2.60 M
CDPA	$4THWC^2 + 2T^2HWC + 4THWN_BC$	**41.52**	**0.9893**	**3.9084**	**23.25**	**28.18**	2.48 M

spatial and temporal dependencies simultaneously. While effective, this design substantially increases computation. Furthermore, we introduce spatial windowing into F-MSA, forming W-F-MSA. It processes spatial and temporal features in a discretized sequential order, which compromises the ability to effectively capture cross-domain feature dependencies. In contrast, our proposed CDPA facilitates effective multi-domain feature interaction while maintaining low computational complexity, thereby achieving superior performance in video-level spectral reconstruction.

Ablation on the Number of Processed Frames. As shown in Table S3, we analyze the impact of the number of input frames T on reconstruction performance. When $T = 1$, the model degenerates into an image-based reconstruction method, which inevitably leads to high temporal inconsistency, manifested as deteriorated ST-RRED metrics. When $T = 2$ or 3, the measurements across short time intervals exhibit strong correlations, allowing the model to effectively exploit complementary information between adjacent frames and substantially improve reconstruction quality. However, as T increases further, the temporal correlation gradually weakens, making it more challenging to extract useful information from an abundance of irrelevant features. As a result, the reconstruction performance shows a downward trend.

S6. Limitations

Our method shows less pronounced improvement on measurements with severely degraded spatial structures, such as those captured by SD-CASSI. This is primarily because the proposed propagation strategy benefits from distinguishable spatial features, which serve as reliable guidance during inter-frame modeling. When the spatial information is heavily aliased, the fusion of complementary spatiotemporal features is hindered, and the model is prone to relying on spatial information rather than leveraging spatiotemporal redundancy.

Fortunately, LADE-DUN [5] provides valuable inspiration. By employing a generative model to recover high-fidelity spatial structures, one can establish a more informative reference for temporal propagation, thereby facilitating inter-frame spectral feature completion and enhancing temporal consistency. However, directly incorporating such a strategy without optimization may incur significant computational overhead, making it impractical for video-level reconstruction tasks.

In future work, we plan to explore lightweight generative priors or structure-aware regularization to enrich the spatial context. Such enhancements may further improve the generalizability of our method across diverse architectures, including those with severely degraded spatial details.

S7. Comparison of Video Quality

To further compare the quality of spectral video reconstruction between PG-SVRT and image-based methods (DPU), we present the reconstruction results of adjacent frames from real-world scene 3 in Fig. S5. As observed, the spectral video reconstructed by DPU exhibits sudden brightening, causing flickering that severely disrupts the temporal continuity of the signal, which may potentially impact downstream tasks such as tracking. In contrast, our method incorporates temporal continuity priors during reconstruction, ensuring consistent information across frames. A supplementary video is provided to offer a more intuitive comparison.

References

[1] Mingyang Song, Yang Zhang, and Tunç O Aydın. Tempformer: Temporally consistent transformer for video denoising. In *European conference on computer vision*, pages 481–496. Springer, 2022. 2

[2] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *IEEE Transactions on Image Processing*, 2024. 2, 3

[3] Ziheng Cheng, Bo Chen, Ruiying Lu, Zhengjue Wang, Hao Zhang, Ziyi Meng, and Xin Yuan. Recurrent neural networks for snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2264–2281, 2023. 3

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 3

[5] Zongliang Wu, Ruiying Lu, Ying Fu, and Xin Yuan. Latent diffusion prior enhanced deep unfolding for snapshot spectral compressive imaging. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 164–181, Cham, 2025. Springer Nature Switzerland. 4

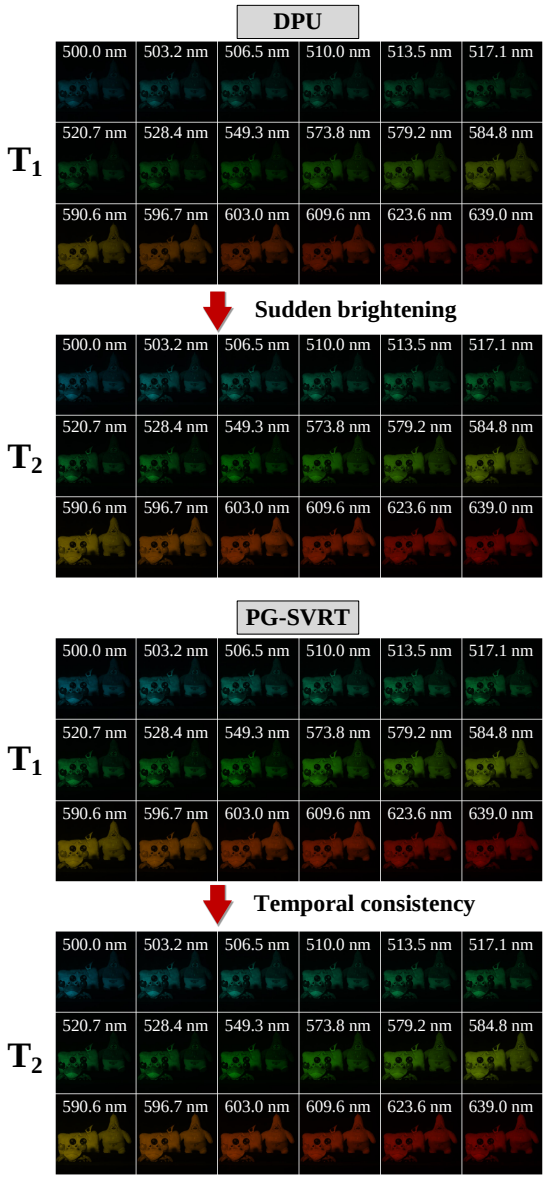


Figure S5. Reconstruction results of adjacent frames from scene 3 using PG-SVRT and DPU. DPU exhibits sudden brightening, whereas PG-SVRT preserves the signal's temporal consistency.

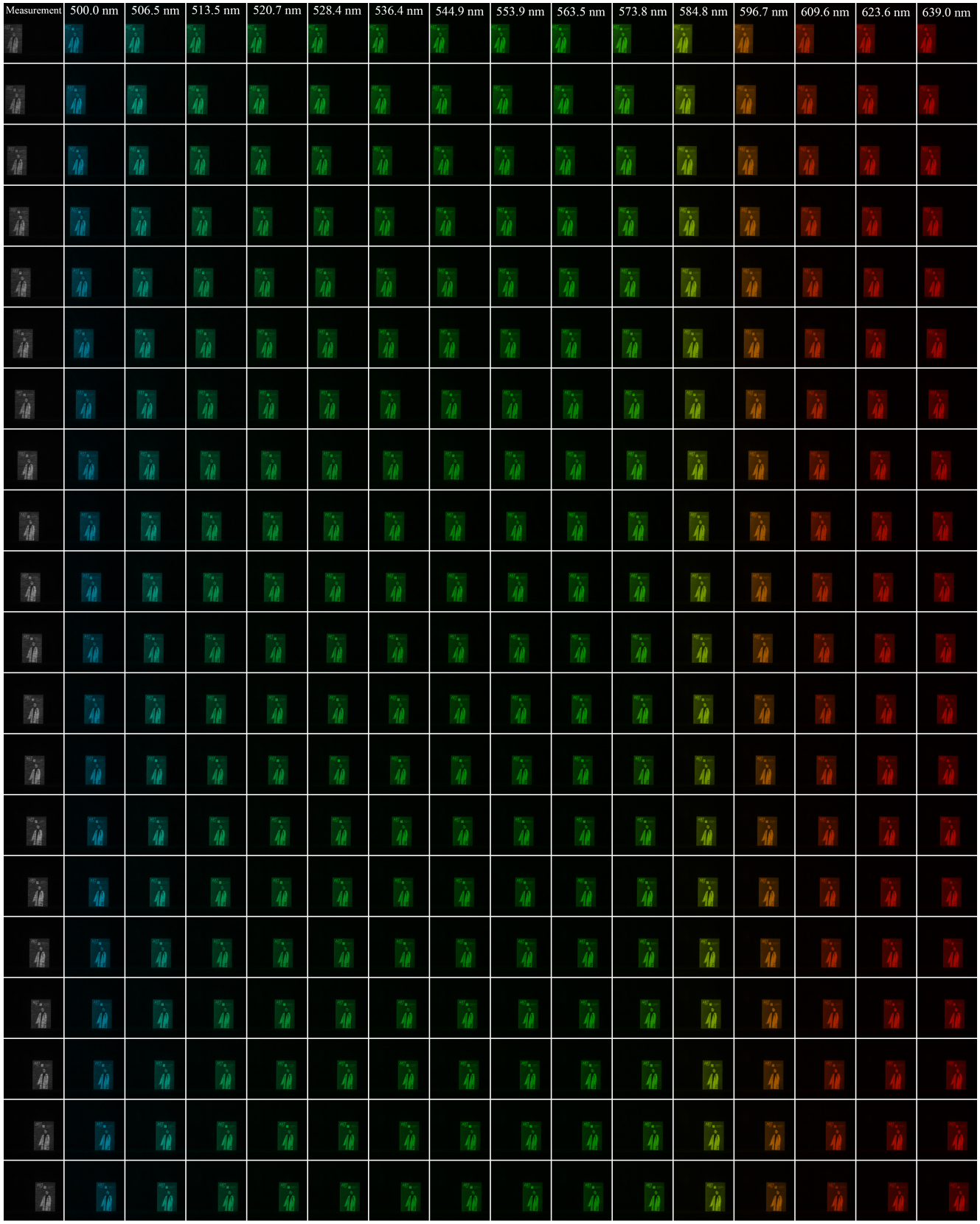


Figure S6. Reconstruction results of PG-SVRT on the gray book translation scene.

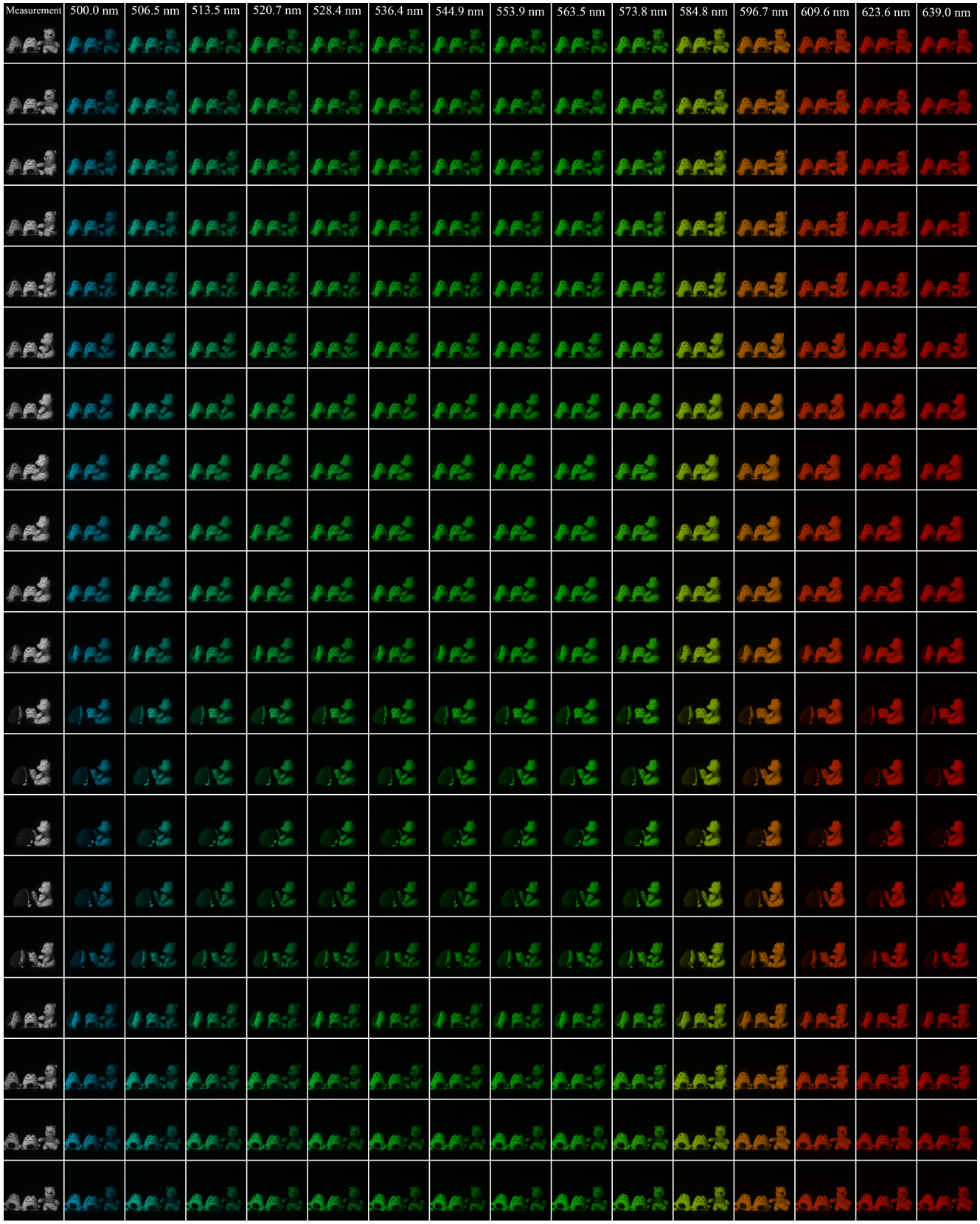


Figure S7. Reconstruction results of PG-SVRT on the toy rotation and translation scene.

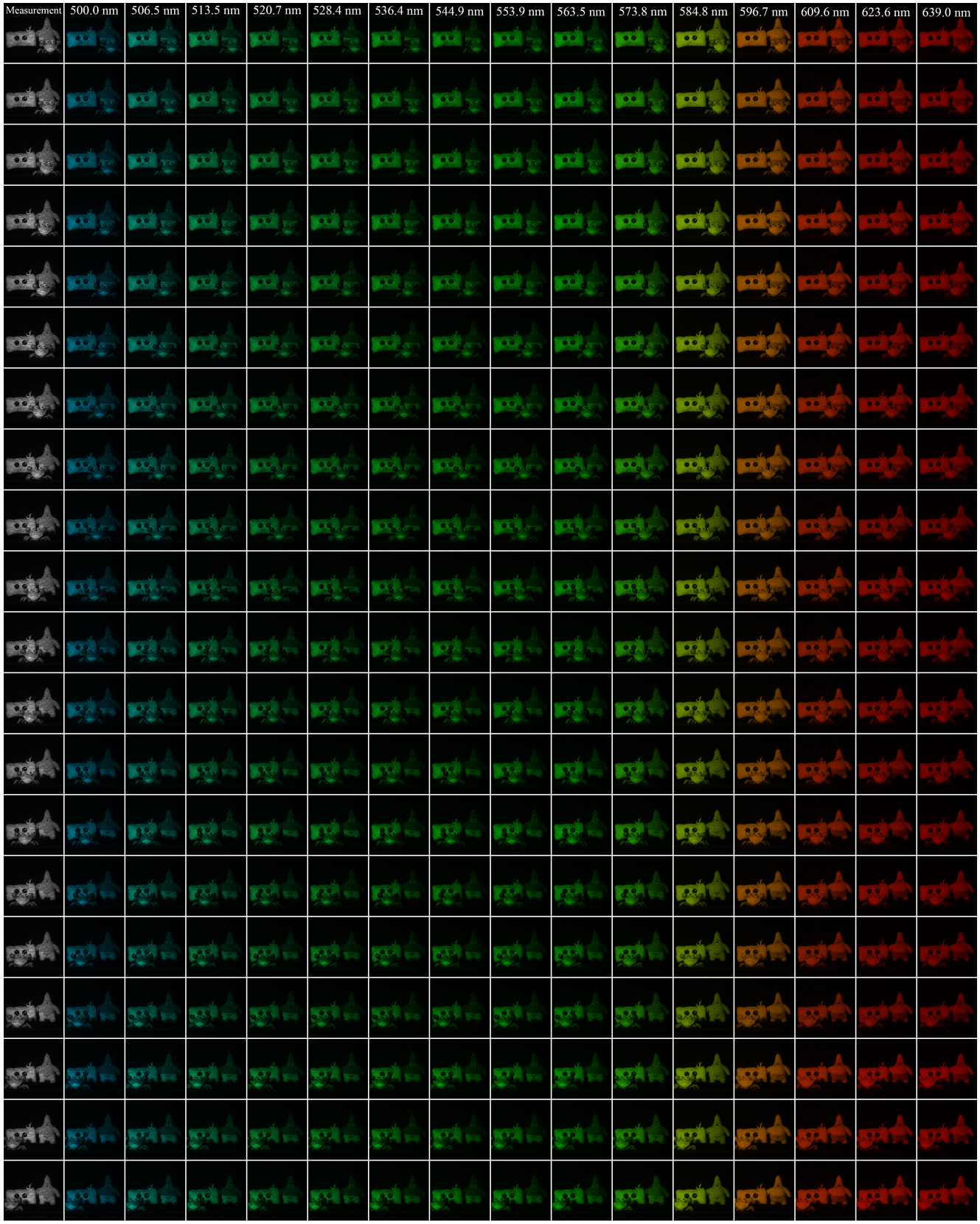


Figure S8. Reconstruction results of PG-SVRT on the mechanical crab scene with high-degree-of-freedom motion.

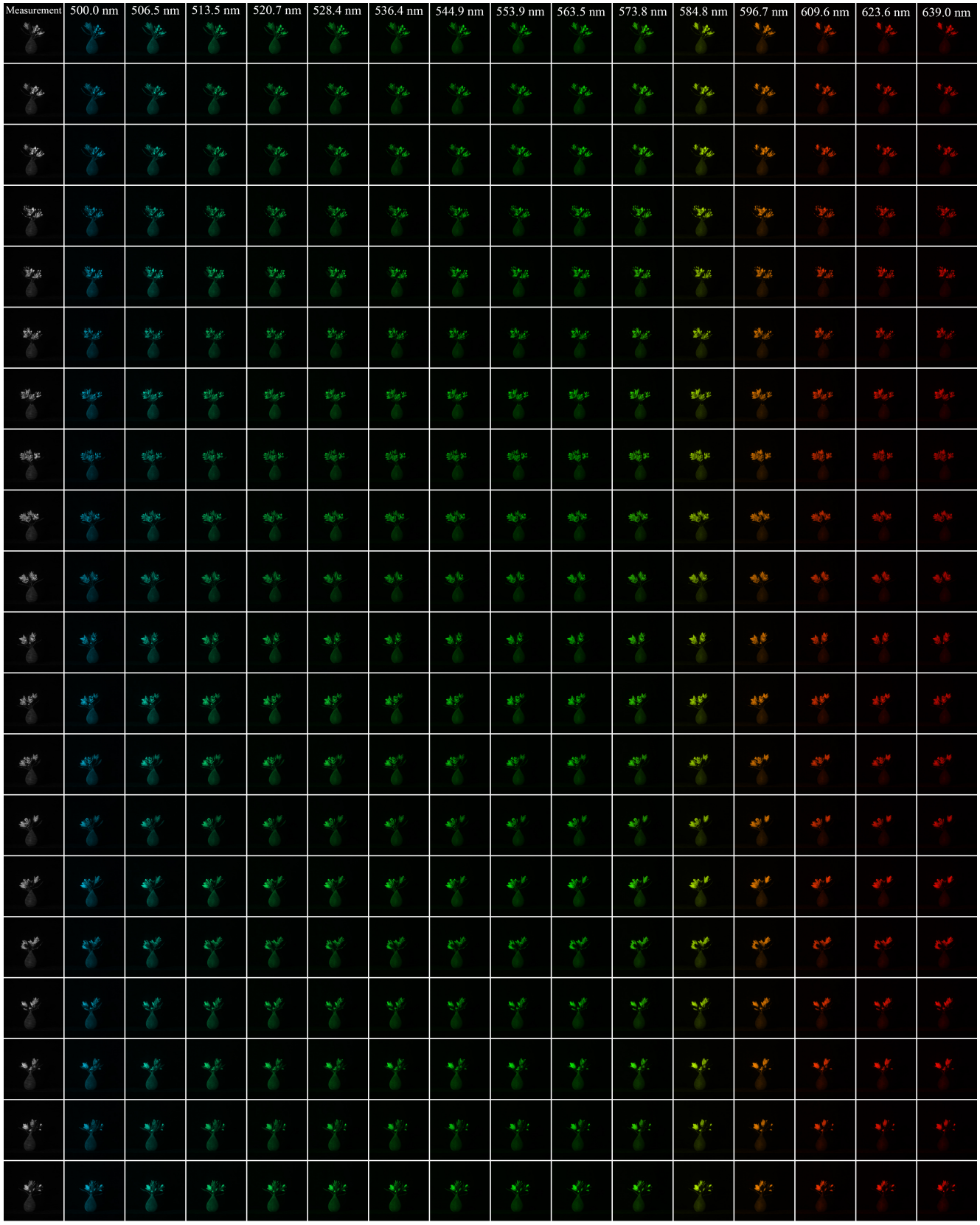


Figure S9. Reconstruction results of PG-SVRT on the LEGO flower rotation scene.

Figure S10. Reconstruction results of PG-SVRT on the toy-built complex street scene.