ϕ-SfT: Shape-from-Template with a Physics-Based Deformation Model —Supplementary Material—

Navami Kairanda[1,2] Edith Tretschk[1] Mohamed Elgharib[1] Christian Theobalt[1] Vladislav Golyanik[1]
[1]Max Planck Institute for Informatics, SIC [2]Saarland University, SIC

This supplementary material provides more details on the experiments, including a detailed evaluation on the synthetic dataset (Sec. 1), per-scene results of the ablation study (Sec. 2), preliminary results on material editing and details on the adaptive optimisation scheme (Sec. 3), and qualitative results on all real sequences (Sec. 4).

1. Evaluation on the Synthetic Dataset

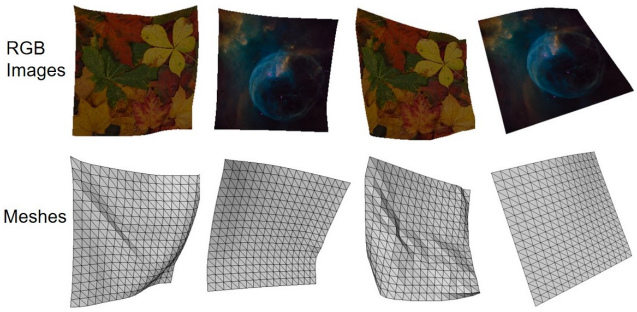

Figure 1. We generate a new "ϕ-SfT synthetic dataset" of four sequences with reference ground-truth meshes to facilitate quantitative evaluation of monocular 3D surface reconstruction methods.

The ϕ-SfT Synthetic Dataset is a new dataset of four monocular RGB sequences of naturalistically deforming surfaces with different textures generated using physics-based simulation [1]. This is the same simulator used as part of our reconstruction pipeline and hence can lead to a small bias favouring our method. A flat square cloth of dimensions $1\times1m$ is provided in the form of a textured mesh to the simulator at the beginning of the simulation. The deformations at subsequent time points are caused by the varying gravity and wind forces acting on the cloth. Moreover, we vary elastic material properties of the cloth across the sequences, following Wang *et al.* [7]. Each sequence contains 50 frames, and the mesh contains 289 regularly-sampled vertices. Finally, the simulated cloth states are rendered as virtual images using PyTorch3D tools [4]. The rendered images serve as inputs to the evaluated methods, and the obtained meshes are 3D ground truth. Fig. 1 shows an overview of the generated synthetic sequences.

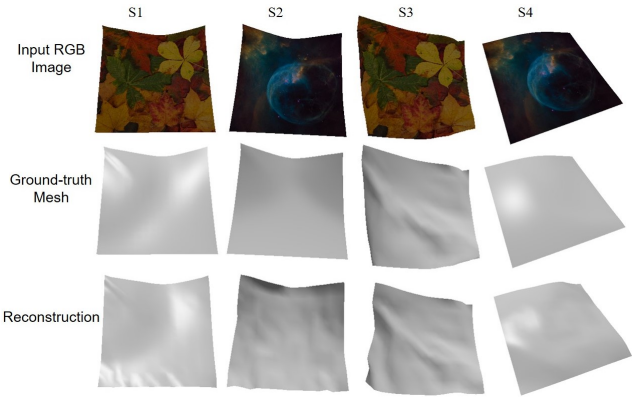

Figure 2. We show qualitative results on all *synthetic* sequences. For the given RGB image, the ground-truth mesh and reconstructed mesh are visualised in the input camera view. ϕ-SfT reconstructs physically plausible and accurate surfaces.

Evaluation Metrics The availability of vertex correspondences in the synthetic dataset allows for more faithful alignment as well as better metrics, *i.e.,* per-frame Procrustes over per-frame ICP, and e_{3D} and e_n over $\tilde{Ch}_{G,M}$. In particular, since vertex correspondences are known across all surface states in the synthetic dataset, we align reconstructions of all methods to the ground truth in a rigid-body fashion using *per-frame Procrustes*. As in previous methods [2, 3, 5, 6], we use the 3D error assuming known correspondences to express the reconstruction accuracy on the new dataset:

$$e_{3D} = \frac{1}{|T|}\sum_{t=1}^{T}\frac{\|\mathbf{G}_t - \mathbf{V}_t\|_F}{\|\mathbf{G}_t\|_F}, \tag{1}$$

where $\mathbf{G}_t$ and $\mathbf{V}_t$ are the vertices of the ground-truth and reconstructed mesh, respectively, and $\|\cdot\|_F$ denotes the Frobenius norm. To better capture the error in local deformations, we additionally compute the average per-vertex angular error in degrees:

$$e_n = \frac{180^\circ}{\pi|T||N|}\sum_{t=1}^{T}\sum_{i=1}^{N}\cos^{-1}(\check{\mathbf{g}}_t^i \cdot \check{\mathbf{x}}_t^i), \tag{2}$$

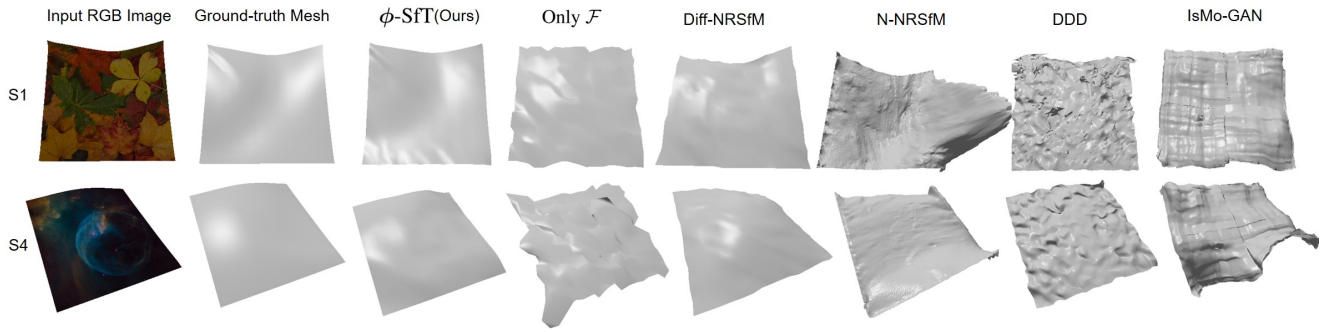

Figure 3. Qualitative comparisons of several tested methods [3, 5, 6, 8], including ϕ-SfT, for an arbitrary frame of the *synthetic* S1 and S4 sequences. Our results are significantly more accurate and, unlike the other methods, are physically plausible.

Seq.	IsMo-GAN		N-NRSfM		DDD		Diff-NRSfM		Only $\mathcal{F}$		ϕ-SfT	
	e_{3D}	e_n	e_{3D}	e_n	e_{3D}	e_n	e_{3D}	e_n	e_{3D}	e_n	e_{3D}	e_n
S1	0.066	34.27	0.167	34.34	0.043	33.86	0.053	**11.30**	0.054	12.73	**0.042**	11.86
S2	0.077	45.11	fail	fail	0.036	25.20	0.055	11.35	0.069	14.92	**0.023**	**10.62**
S3	0.096	36.72	0.113	26.36	0.066	42.16	0.077	17.59	0.059	15.83	**0.033**	**9.12**
S4	0.078	41.16	0.077	24.36	0.023	19.86	0.063	5.69	0.043	9.33	**0.005**	**2.61**
Avg.	0.079	39.32	0.119*	28.35*	0.042	30.27	0.062	11.48	0.056	13.20	**0.026**	**8.55**

Table 1. Evaluation of compared methods and ϕ-SfT on our synthetic dataset using e_{3D} and e_n after rigid alignment with per-frame Procrustes. N-NRSfM [6] failed on S2 (the missing error is not considered when computing the averages, which are marked with "*"). Our full method gives the most accurate results.

where $\check{\mathbf{g}}_t^i \in \mathbb{R}^3$ and $\check{\mathbf{x}}_t^i \in \mathbb{R}^3$ are the unit normals at the ith vertex in frame t of the ground-truth and reconstructed mesh, respectively.

Comparison/Results Similar to ϕ-SfT's real dataset evaluation, we compare our synthetic dataset results to SfT methods, namely Yu *et al.*'s Direct, Dense, Deformable (DDD) [8], and Shimada *et al.*'s IsMo-GAN [5], as well as Sidhu *et al.*'s Neural NRSfM (N-NRSfM) [6] and Parashar *et al.*'s Diff-NRSfM [3]. Since ground-truth meshes are available for the synthetic dataset, we provide ground-truth 2D point correspondences as input to N-NRSfM and Diff-NRSfM. We provide DDD with the required hierarchy of coarse-to-fine templates. Additionally, we show the result of a baseline (Only $\mathcal{F}$) where the only optimisation parameters are the correctives $\{\mathcal{F}_t\}_t$.

In Fig. 2, we show qualitative results on all synthetic sequences. ϕ-SfT reconstructs physically plausible and accurate surfaces. Fig. 3 shows that ϕ-SfT outperforms related methods qualitatively, similar to its performance on the real dataset. This demonstrates that SfT and NRSfM, both relying on simple geometric prior assumptions, struggle to estimate physically plausible surfaces. We also note that Diff-NRSfM performs better on our synthetic dataset than our real dataset, as the deformations here are global and smooth (see Fig. 3, fifth column). We refer to Tab. 1 for mean vertex error, e_{3D}, and mean angular normal error in degrees, e_n, on our synthetic data with per-frame Procrustes using ground-truth correspondences. We outperform others on all synthetic sequences, except for Diff-NRSfM using e_n on S1. As shown in Tab. 1, our method has the lowest average e_{3D}, suggesting it better reconstructs global deformations, and the lowest average e_n, suggesting it also better captures local folds.

Sequence	w/o $\mathcal{F}$	w/o adaptive	w/o E_s	Full
S1	4.61	1.90	0.89	**0.79**
S2	3.28	5.78	**2.75**	4.16
S3	3.55	7.17	**3.54**	4.21
S4	36.13	13.37	8.91	**7.60**
S5	25.55	8.64	7.54	**6.15**
S6	**3.74**	6.84	7.26	6.20
S7	10.31	4.66	**4.73**	6.34
S8	4.17	3.34	2.71	**2.52**
S9	4.87	4.19	2.65	**2.36**
Average	10.69	6.21	5.33	**4.48**

Table 2. We report the Chamfer distance $\tilde{Ch}_{G,M}$ (multiplied by 10^4 for readability) when ablating various design choices of our method: the external forces (w/o $\mathcal{F}$), the adaptive optimisation scheme (w/o adaptive), and the silhouette energy term (w/o E_s). We use sequences from the new ϕ-SfT *real* dataset.

2. Ablation Study

We ablate the following design choices of our method: 1) Operation without corrective forces $\mathcal{F}$, 2) Influence of the adaptive training by considering all frames from the start, and 3) Disabling the silhouette term E_s. The per-scene breakdown of the ablation experiments is in Tab. 2. The correctives forces lead, on average, to the largest improvement despite three scenes showing worse performance. The adaptive scheme improves results on all scenes. E_s helps when the structure deforms and changes its shape in the reprojection significantly, as is the case for most sequences. However, S3, for instance, has fewer global deformations and more significant local folds, in which case E_s does not

help (Fig. 5). Note that E_s is susceptible to errors in the input segmentation whereas the photometric energy E_p is (empirically) robust to them.

3. Miscellaneous

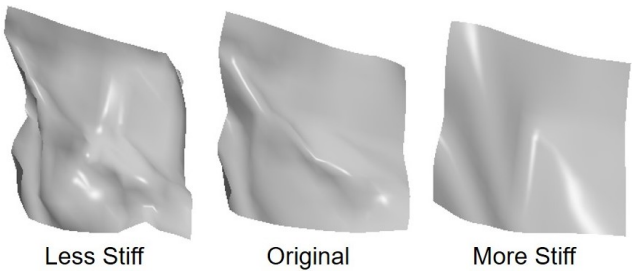

Figure 4. We scale the bending parameters of our result of synthetic scene S3 by factor 10 for less/more stiffness (we show the last frame). The deformations introduced in coarse shape and local folds suggest that intuitive editing is possible.

Material Editing We conduct a preliminary experiment to demonstrate that the optimised physical parameters are meaningful and inferred well enough to enable intuitive editing. Given optimised physical parameters ϕ^*, we aim to generate new deformations at test time by modifying the inferred material parameters. To this end, we run the physics simulation after scaling stretching stiffness $\mathcal{S}$ while re-using the other physical parameters $\{d^*, \mathcal{B}^*, w^*, \mathcal{F}^*\}$. See Fig. 4 for the result.

Adaptive Optimisation Parameters By default, we start with an initial active temporal window of $t_a = 5$. However, we set $t_a = 3$ on the real sequence S3 because it has large global deformations in early frames. In case the optimisation cannot reach the threshold b for a given temporal window, we set a maximum number of iterations i_{max}, after which we grow the window regardless. We set $i_{max} = 5$ by default. However, we use a higher value of $i_{max} = 10$ for sequences with difficult folds (real S2, S3) or large global deformations (synthetic S1, S2, S3).

4. Qualitative Results

We show qualitative reconstruction results for arbitrary frames on all real sequences in Fig. 5. ϕ-SfT reconstructs challenging surface deformations well by capturing both coarse shape and local folds. Especially S3 and S4 in Fig. 5 show that our physics-based approach provides a reasonable prior for self-occluded surface parts. Fig. 6 contains depth maps reconstructed by our approach.

References

[1] Junbang Liang, Ming Lin, and Vladlen Koltun. Differentiable cloth simulation for inverse problems. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. 1

[2] Dat Tien Ngo, Sanghyuk Park, Anne Jorstad, Alberto Crivellaro, Chang D. Yoo, and Pascal Fua. Dense image registration and deformable surface reconstruction in presence of occlusions and minimal texture. In *International Conference on Computer Vision (ICCV)*, 2015. 1

[3] Shaifali Parashar, Mathieu Salzmann, and Pascal Fua. Local non-rigid structure-from-motion from diffeomorphic mappings. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[4] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 1

[5] Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Didier Stricker. IsMo-GAN: Adversarial learning for monocular non-rigid 3d reconstruction. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. 1, 2

[6] Vikramjit Sidhu, Edgar Tretschk, Vladislav Golyanik, Antonio Agudo, and Christian Theobalt. Neural dense non-rigid structure from motion with latent space constraints. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[7] Huamin Wang, Ravi Ramamoorthi, and James F. O'Brien. Data-driven elastic models for cloth: Modeling and measurement. *ACM Transactions on Graphics*, pages 71:1–11, 2011. 1

[8] Rui Yu, Chris Russell, Neill D. F. Campbell, and Lourdes Agapito. Direct, dense, and deformable: Template-based non-rigid 3d reconstruction from rgb video. In *International Conference on Computer Vision (ICCV)*, 2015. 2

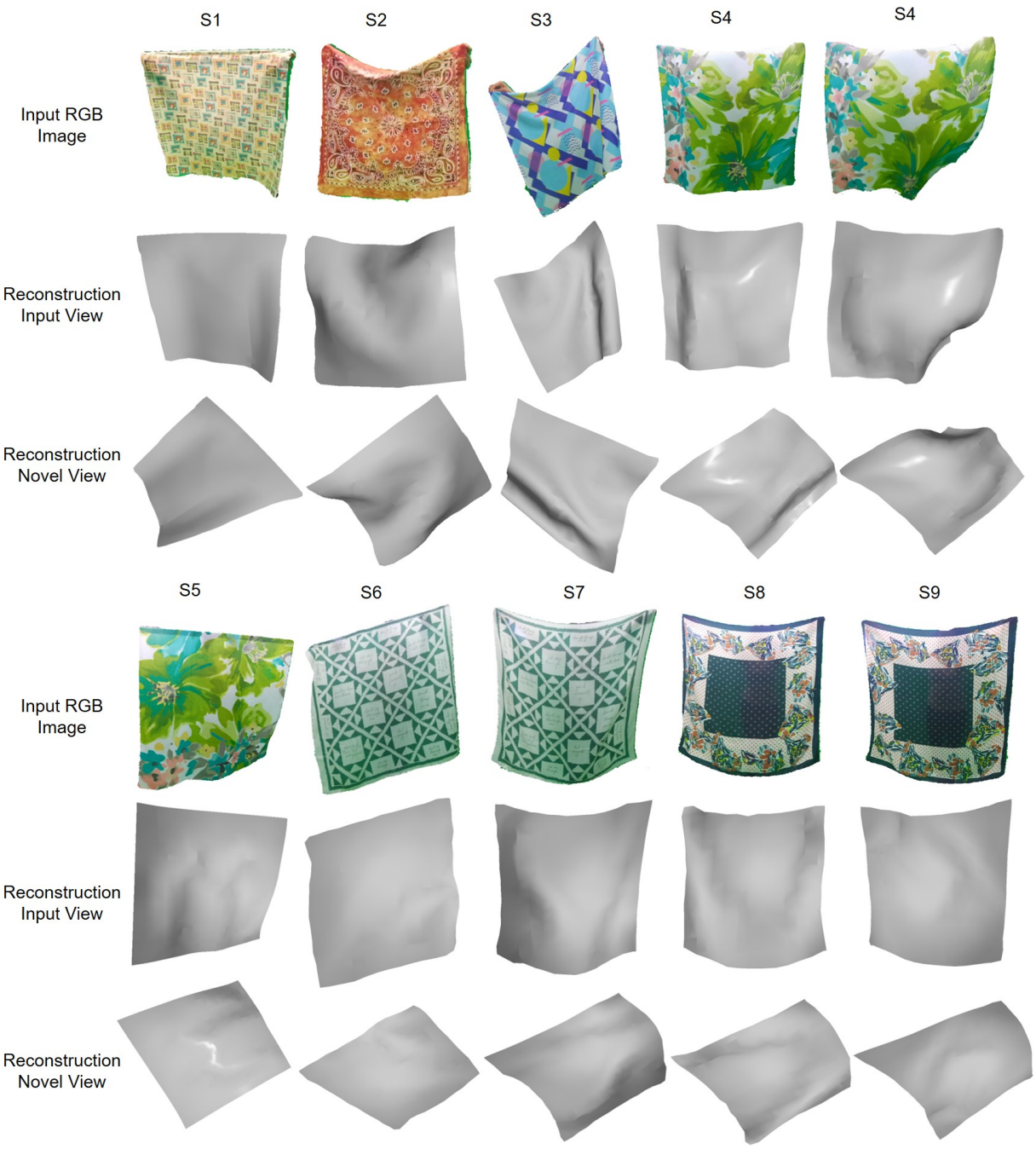

Figure 5. We show qualitative results on all real sequences. For the given RGB image, the reconstructed mesh is visualised in the input camera view as well as novel camera view. ϕ-SfT accurately reconstructs the coarse shape and local folds.

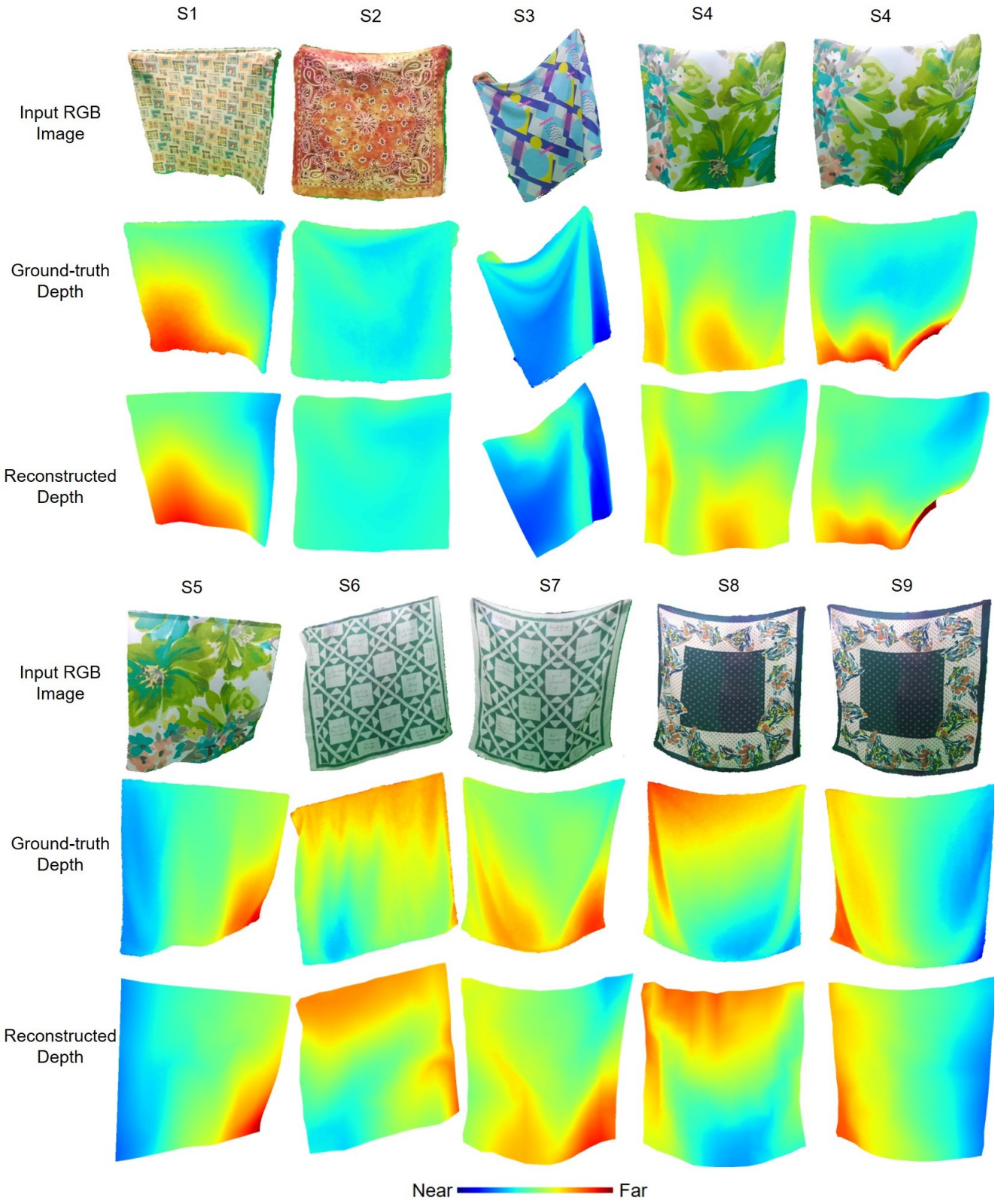

Figure 6. We show qualitative results as colour-coded depth maps on all real sequences. For the given RGB image, the ground-truth depth map exhibits similar features as our reconstructed depth. Both the coarse shape and local folds are well captured.