V-FUSE: Volumetric Depth Map Fusion with Long-Range Constraints Supplementary Material

Nathaniel Burgdorfer   Philippos Mordohai
Stevens Institute of Technology

In this document, we provide additional implementation details (Section S.1), equations for learnable geometric constraints (Section S.2), details on confidence computation and evaluation (Section S.3), ablation studies evaluating architecture contributions (Section S.4), further quantitative evaluations (Section S.5), and additional qualitative comparisons and examples (Section S.6).

S.1. Implementation Details

During training and inference, we set the number of input views for fusion to $N = 5$ and the number of depth planes to $M = 8$ for input depth maps from all methods. For the DTU [1] dataset, we use the maximum output resolution of each method as the input for V-FUSE. Specifically, we use 400×296 for MVSNet, 1600×1184 for UCSNet, and 1600×1152 for NP-CVP-MVSNet and GBi-Net. For training, we scale the input by a factor of 0.5, with the exception of MVSNet, for which we train at the full resolution. For Tanks & Temples [5], we use input resolutions of 1920×1056 for UCSNet and 1920×1024 for both NP-CVP-MVSNet and GBi-Net. As the minimum and maximum allowable search window radii, we use $\psi_{min} = 0.005$ and $\psi_{max} = 0.50$. The terms of the loss are weighed by $\lambda_d = 0.5$, $\lambda_c = 20.0$, and $\lambda_r = 0.5$.

S.2. Hyper-parameters

We define two parameters used in the formulations of the geometric constraints. For support, we use σ_p to determine the sharpness of the support response boundary.

$$\sigma_p = \gamma_\sigma \frac{(B_p^{max} - B_p^{min})}{M(b_{max} - b_{min})} \tag{1}$$

Here, γ_σ is a learned hyper-parameter, B_p^{min} and B_p^{max} are the minimum and maximum depth bounds per pixel, M is the number of depth hypotheses, and b_{min} and b_{max} are the overall minimum and maximum depth bounds that are given as input for the current reference view. Lower values of γ_σ correspond to a tighter support window. Since this is a function of the per-pixel depth bounds, support adapts to the confidence at each pixel.

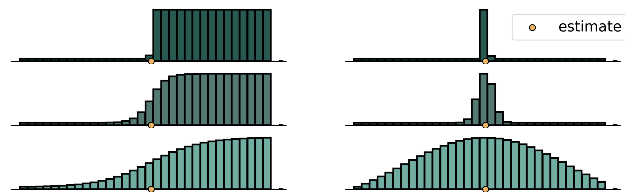


Figure S.1. Visualization of the effects of modifying (a), the multiplier used in the sigmoid and (b), the sigma used in the Gaussian. The *response boundary* is the region around the current depth estimate where high response values transition to low response values. Decreasing the multiplier for occlusions (as well as free-space violations) (a; top→bottom) causes the sigmoid response boundary to soften around the current estimated depth. Increasing the sigma for support (b; top→bottom) causes the support response boundary to soften around the current estimated depth.

Method	AUC↓	
	DTU	Tanks & Temples
MVSNet [9]	6.12	-
MVSNet + V-FUSE	**4.08**	-
UCSNet [2]	20.29	0.394
UCSNet + V-FUSE	**4.96**	**0.208**
NP-CVP-MVSNet [8]	8.28	0.212
NP-CVP-MVSNet + V-FUSE	**4.35**	**0.154**
GBi-Net [6]	3.69	0.464
GBi-Net + V-FUSE	**2.48**	**0.448**

Table S.1. Quantitative comparison of the average AUC (lower is better) between V-FUSE and each baseline method on the DTU [1] evaluation set and the Tanks & Temples [5] training set.

For occlusions and free-space violations, we use λ_p to determine the sharpness of the sigmoid response boundary.

$$\lambda_p = \gamma_\lambda \frac{M(b_{max} - b_{min})}{(B_p^{max} - B_p^{min})} \tag{2}$$

Here, γ_λ is a learned hyper-parameter. Lower values of γ_λ correspond to a softer sigmoid response boundary. Occlusions and free-space violations also adapt to the confidence at each pixel. See Figure S.1 for a visualization of the response boundary.

(c) Image (d) Input Confidence (c) Fused Confidence

Figure S.2. Comparison between the input confidence maps with the fused confidence maps on scenes from the DTU [1] benchmark using GBi-Net [6] (top), and UCSNet [2] (bottom) as input. (Darker pixel value corresponds to lower confidence.)

Method	DTU
	MAE↓
GBi-Net [6]	5.845
+ V-FUSE [brute-force]	5.344
+ V-FUSE [sup]	4.813
+ V-FUSE [fsv+occ]	4.477
+ V-FUSE	**4.196**

Table S.2. Ablation study evaluating the contribution from different aspects of the network architecture. [brute-force] indicates the network is trained using the entire search space (using 128 depth planes) without the SWE sub-network. [sup] and [fsv+occ] indicate that the network only utilizes the support channel, or only utilizes the free-space violation and occlusion channels, respectively.

S.3. Confidence Estimates

To evaluate the quality of confidence estimates, we report the *area under the curve (AUC)*. The AUC is the area under the ROC curve, which maps the error rate for the depth map as a function of density based on the sorted confidence values [4]. We first sort the estimated depths according to confidence, and form the sparsification curve of MAE vs. depth map density by progressively dropping the least confident depths [4] to obtain depth maps of lower density, and presumably lower error. Small area under the sparsification curve indicates that confidence has been estimated well and can be used to rank depth estimates accurately.

The fused confidence maps are directly computed from the estimated search window radius. After we normalize the per-pixel window radii from 0 to 1, the confidence value at each pixel is $C_p^f = 1 - R_p$. Intuitively, a larger estimated radius for a given pixel should indicate lower confidence in the final depth estimate. This relationship is also reflected and enforced in our loss function. To compare the quality of the confidence maps, we report the AUC for all methods in Table S.1. The output confidence values after fusion prove to be more reliable estimates of confidence. Qualitative confidence map results can be seen in Figure S.2.

Method	V-FUSE [brute-force]	V-FUSE [swe]
Memory(GB)	37.734	**4.439**
Parameters	**289,587**	297,429
Inference Time(s)	22.95	**2.51**

Table S.3. Ablation study between the brute-force approach and the SWE sub-network. We use 192 depth planes for the brute-force approach, with every pixel having the same depth bounds given as input by the dataset. For the SWE sub-network, we use 8 depth planes with each depth bound estimated per-pixel. We can observe significant memory and run-time improvements, with minimal additional model parameters.

Method	DTU
	AUC↓
V-FUSE [pv]	7.66
V-FUSE [pv+swe]	6.06
V-FUSE [swe]	**2.48**

Table S.4. Ablation study on the confidence map computation. We evaluate using only the output probability volume, as well as the incorporated and stand-alone radius outputs from the SWE sub-network (*pv: probability volume; swe: search window estimate*).

S.4. Ablation Studies

We provide ablation studies evaluating the contributions of several aspects of the network architecture. In Table S.2, we evaluate the MAE isolating the SWE sub-network, as well as the different constraints. We show the contributions of using the brute-force search space approach, as well as the contributions of using only support and only occlusions and free-space violations. In Table S.3, we show the memory, run-time, and parameter difference between the brute-force approach and the SWE sub-network.

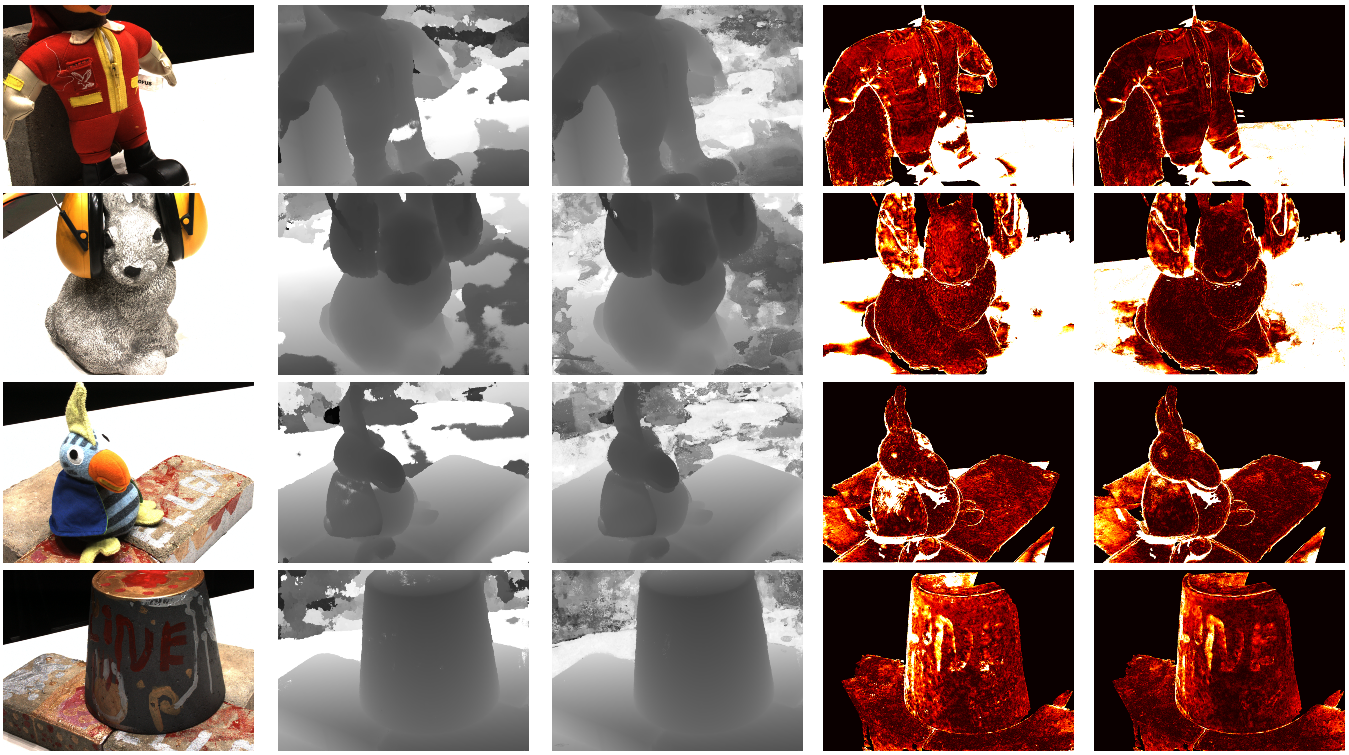
Most *state-of-the-art* Deep MVS architectures directly compute confidence estimates from the output probability volume of the network, using a small window around the estimated depth voxel. We explore using this method, incorporating the radius estimate from the SWE sub-network, as well as using only the radius to compute confidence. Specifically, following previous work in Deep MVS, we compute confidence maps from the output probability volume of the network by summing the probability values for the four sur-

Method	Tanks & Temples		
	Precision ↑	Recall ↑	F-Score ↑
UCSNet [2]			
+ Gipuma [3]	46.66	**70.34**	54.83
UCSNet + V-FUSE	**47.08**	68.64	**55.03**
GBi-Net [6]			
+ Gipuma [3]	**54.49**	71.25	**61.42**
+ V-FUSE	50.16	**73.08**	59.08

Table S.5. Precision, Recall, and F-Score on the intermediate set of Tanks & Temples [5].

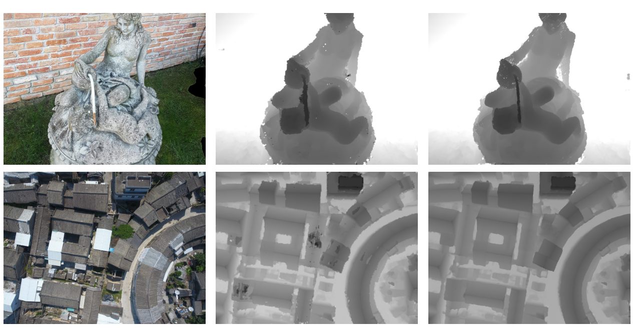
Method	BlendedMVS [MAE↓]							
	Mean	scan106	scan107	scan108	scan109	scan110	scan111	scan112
GBi-Net [6]	0.319	1.661	0.006	0.027	0.017	0.025	0.031	0.462
+ V-FUSE	**0.288**	**1.539**	**0.001**	**0.015**	**0.011**	**0.010**	**0.024**	**0.417**

Table S.6. Quantitative comparison of the 2D depth map errors on the validation set of BlendedMVS [10] benchmark. The best results are marked in bold. Without any fine-tuning, V-FUSE improves the inputs of GBi-Net [6] on all scenes in the validation set.



(a) Image (b) NP-CVP-MVSNet Depth [8] (c) V-FUSE Depth (d) NP-CVP-MVSNet Error [8] (e) V-FUSE Error

Figure S.3. Qualitative comparison on the DTU [1] dataset between NP-CVP-MVSNet [8] and V-FUSE. We compare the input and fused depth and error maps. Error maps are colored using a heat map (larger errors correspond to brighter colors).



(c) Image (d) Input Depth (c) Fused Depth

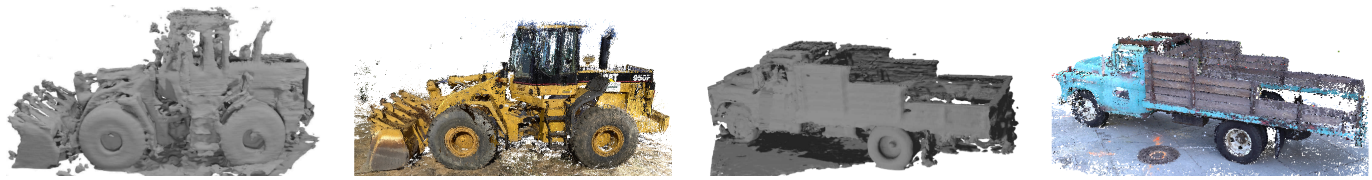
Figure S.4. Qualitative examples comparing the input depth maps with the fused output depth maps from the BlendedMVS [10] dataset using GBi-Net [6] as input.

rounding voxels corresponding to the index of the selected depth. We then test adding the inverse of the radius value from the SWE sub-network as a weighting to this confidence. Finally, we test using only the inverse radius value to compute our confidence. We evaluate these different methods of producing confidence maps in Table S.4. In all instances, the confidence maps produced by V-FUSE are better indications of depth estimate quality, and can be used to more effectively rank depth estimates.

S.5. Additional Quantitative Evaluations

We provide additional results on the **BlendedMVS** [10] dataset. We report the MAE on the validation set between the GBi-Net [6] input depth maps and the V-FUSE output fused depth maps. We used the pre-trained GBi-Net model trained on the BlendedMVS training set and tested V-FUSE using the model trained on DTU without any fine-tuning. The quantitative results are presented in Table S.6. We show significant improvements in all scenes. Qualitative results can be viewed in Figure S.4.

We also provide the *Precision* and *Recall*, alongside the



(a) NeuralFusion [7] (b) V-FUSE (c) NeuralFusion [7] (d) V-FUSE

Figure S.5. Qualitative examples comparing the reconstruction results of NeuralFusion [7] and V-FUSE on scenes from Tanks & Temples [5]. The visuals of the 3D models produced by NeuralFusion [7] are sampled from the paper.

F-Score on the **Tanks & Temples** [5] intermediate test set, retrieved from the benchmark leaderboard. The *Precision* score is the percentage of points in the reconstructed point cloud that have a Chamfer distance to the closest point in the ground-truth point cloud below some threshold, τ. The *Recall* score is the percentage of points in the ground truth-point cloud that have a Chamfer distance to the closest point in the reconstructed point cloud below the same threshold, τ. The *F-Score* is then the harmonic mean of these two scores. Quantitative results can be found in Table S.5. We improve the Precision and F-Score using UCSNet as input, and improve the Recall using GBi-Net as input.

S.6. Additional Qualitative Results

We show additional qualitative results for all baselines. Figure S.3 shows a comparison of depth and error maps between NP-CVP-MVSNet [8] and V-FUSE on several scenes from the DTU [1] evaluation set. In Figure S.5, we provide a comparison between the final 3D reconstruction results of NeuralFusion [7] and V-FUSE. We only provide a qualitative comparison, as NeuralFusion does not provide any quantitative results on any of the datasets used in our experiments. Figure S.6 shows a comparison of depth and confidence maps between GBi-Net [6] and V-FUSE on several scenes from the DTU evaluation set. In Figure S.7, we can observe the depth maps comparison of scenes from the Tanks & Temples [5] dataset between GBi-Net [6] and V-FUSE. We provide depth map comparisons from several scenes of the validation set from BlendedMVS [10] using GBi-Net [6] as input in Figure S.8. We also show final reconstructions from the DTU and Tanks & Temples datasets in Figure S.9 and Figure S.10, respectively.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-Scale Data for Multiple-View Stereopsis. *IJCV*, 2016.

[2] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep Stereo Using Adaptive Thin Volume Representation With Uncertainty Awareness. In *CVPR*, 2020.

[3] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively Parallel Multiview Stereopsis by Surface Normal Diffusion. In *ICCV*, 2015.

[4] Xiaoyan Hu and Philippos Mordohai. A Quantitative Evaluation of Confidence Measures for Stereo Vision. *IEEE TPAMI*, 34(11):2121–2133, 2012.

[5] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction. *ACM TOG*, 36(4), 2017.

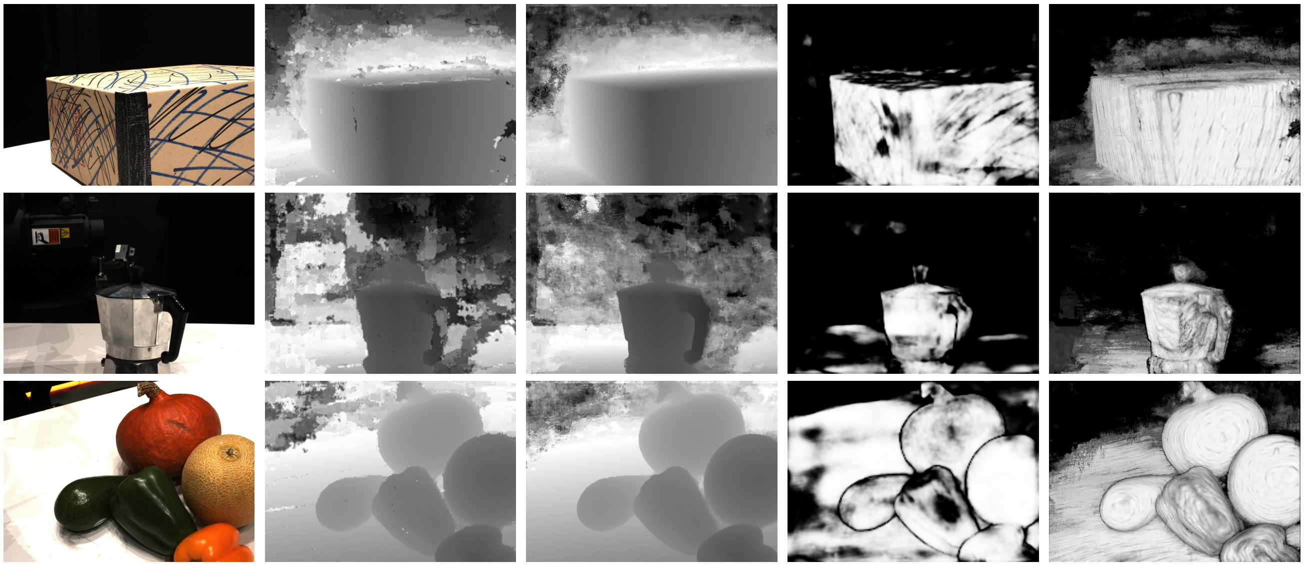
[6] Zhenxing Mi, Chang Di, and Dan Xu. Generalized Binary Search Network for Highly-Efficient Multi-View Stereo. In *CVPR*, pages 12991–13000, 2022.

[7] Silvan Weder, Johannes L. Schonberger, Marc Pollefeys, and Martin R. Oswald. NeuralFusion: Online Depth Fusion in Latent Space. In *CVPR*, pages 3162–3172, 2021.

[8] Jiayu Yang, Jose M. Alvarez, and Miaomiao Liu. Non-Parametric Depth Distribution Modelling Based Depth Inference for Multi-View Stereo. In *CVPR*, pages 8626–8634, 2022.

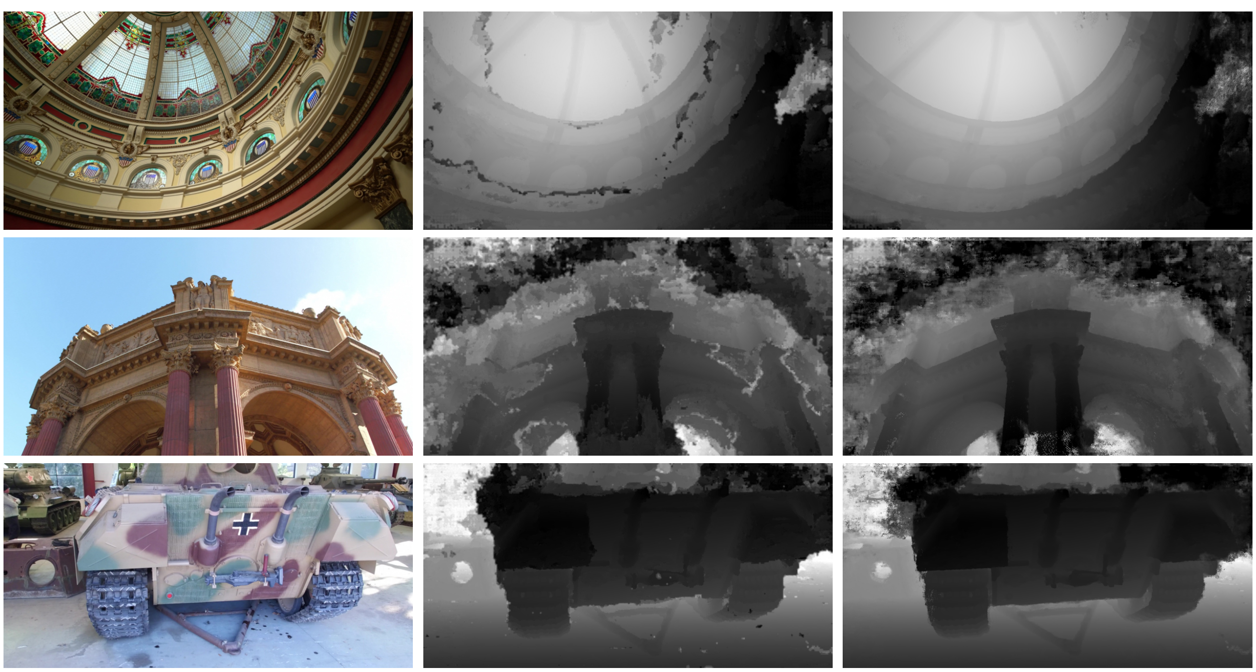
[9] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth Inference for Unstructured Multi-view Stereo. In *ECCV*, 2018.

[10] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020.



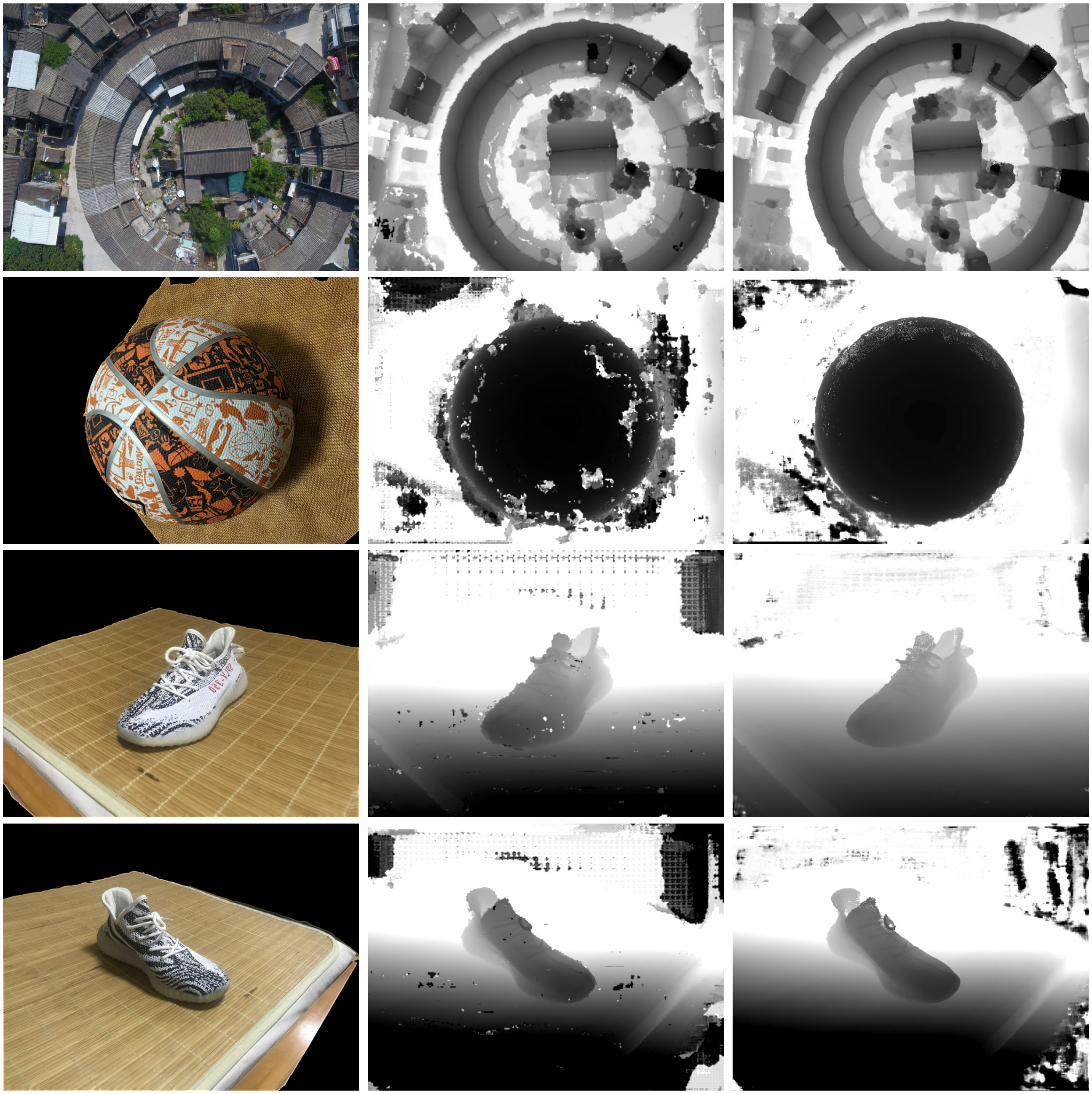
(a) Image (b) GBi-Net Depth [6] (c) V-FUSE Depth (d) GBi-Net Confidence [6] (e) V-FUSE Confidence

Figure S.6. Qualitative comparison on the DTU [1] dataset between GBi-Net [6] and V-FUSE. We compare the input and fused depth and confidence maps. The improvements in surface boundary definition in the fused depth maps are also present in the fused confidence maps. We can observe much more detailed confidence maps, with more continuous changes in confidence values as opposed to very abrupt changes in the input confidence maps.



(a) Image (b) GBi-Net [6] Depth (c) V-FUSE Depth

Figure S.7. Qualitative depth map comparison on the Tanks & Temples [5] dataset between GBi-Net [6] and V-FUSE.



(c) Image | (d) GBi-Net [6] Depth | (c) V-FUSE Depth

Figure S.8. Qualitative examples comparing the input depth maps with the fused output depth maps from the BlendedMVS [10] dataset using GBi-Net [6] as input.

Figure S.9. Output point clouds of V-FUSE from the DTU [1] dataset using NP-CVP-MVSNet [8] as input.

Figure S.10. Output point clouds of V-FUSE on the intermediate set from the Tanks & Temples [5] dataset using UCSNet [2] as input.