VHS: High-Resolution Iterative Stereo Matching with Visual Hull Priors

∼ *Supplementary Material* ∼

Markus Plack Hannah Dröge Leif Van Holland Matthias B. Hullin
University of Bonn
Bonn, Germany

Our supplementary material includes more detailed explanations and experimental results on our proposed method. We will discuss the following:

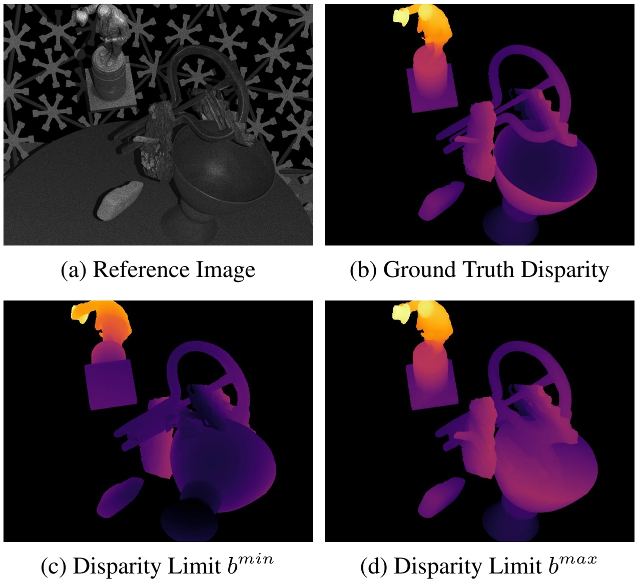
A. Preparation of our custom dataset "FlyingObjaverse"

B. Memory efficient training

C. Evaluation on priors beyond the visual hull

D. Evaluation on the high-resolution Spring dataset [6]

E. Robustness to noise and real-world data



(a) Reference Image

(b) Ground Truth Disparity

(c) Disparity Limit b^{min}

(d) Disparity Limit b^{max}

Figure 1. Sample from the FlyingObjaverse training dataset. Notice how the true disparity is close to the upper disparity limit except for the basin in the bottom right, which cannot be recovered from the visual hull.

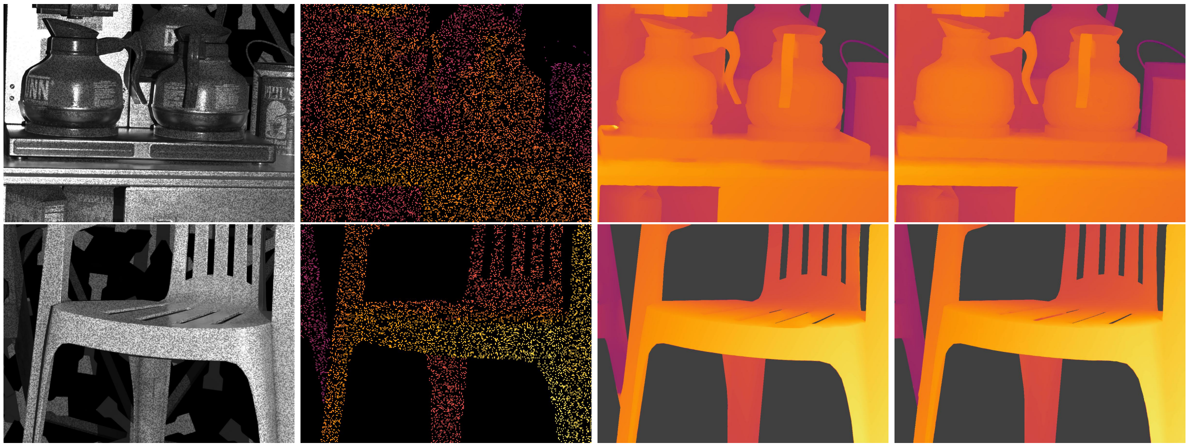
A. Dataset Preparation

We render a custom dataset using Mitsuba 3 [4] and meshes from Objaverse-XL [2], and place objects on a virtual capture stage. Each scene contains a randomly transformed arrangement of $1 - 10$ objects, as shown in Fig. 1, with an infrared camera stereo setup using active illumination with projected patterns similar to [3] and a total of 68 cameras for the masks, all captured at a resolution of 4608×5328. We render 2 stereo pairs for 500 scenes. For testing, we follow the same rendering pipeline but select meshes from different sources to avoid contamination of the training dataset. To test performance on difficult lighting effects, we curated scenes with objects that include challenging reflectance properties and fine details using high-quality meshes from Poly Haven[1] and built eight scenes, each viewed from four different angles. As a second test set, we used SMPL [5] human models with texture from SMPLitex [1] to evaluate performance on human subjects. We create 100 scenes by combining random poses from the animations with random textures and render 2 stereo pairs for each scene.

B. Memory Efficient Training

To update the weights of the recurrent network using backpropagation, the recursive computation has to be unrolled, which can get prohibitively expensive in terms of memory requirements. Instead, we group the unrolled graph into blocks of m evaluations of the ConvGRU and assume that no gradient flow exists between these blocks. We can run the forward pass for m steps and at the end of the block compute the gradients for both the ConvGRU weights θ_{GRU} and the feature encoding weights θ_{enc}, because the iterative refinement outputs a new disparity estimate D^i after each step, which allows us to compute losses $\mathcal{L}_i$ between these intermediate predictions and the ground truth and accumulate them as $\mathcal{L}_{\text{block}}$. This in turn yields gradients $\frac{\partial \mathcal{L}_{\text{block}}}{\partial \theta_{\text{GRU}}}$ and $\frac{\partial \mathcal{L}_{\text{block}}}{\partial \theta_{\text{enc}}}$ that we accumulate into buffers $\mathcal{G}_{\text{GRU}}$ and $\mathcal{G}_{\text{enc}}$ respectively. The computational graph that was recorded to perform the backpropagation can now be discarded, as the gradients in the next block do not depend on the previous

[1] https://polyhaven.com/



(a) Reference Image (b) Depth Priors (c) Baseline Disparity (d) Predicted Disparity

Figure 2. Stereo matching using sparse depth as prior.

operations, and the process is repeated until all blocks are processed. Now the accumulated gradients can be used to compute the parameter update based on gradient descent. Crucially, the eager computation decouples the memory requirements from the number of refinement steps, allowing to increase the input resolution without sacrificing the quality of the refinement. Algorithm 1 shows the pseudocode of the memory-efficient training procedure. This is what we call the "Repeated" strategy since the gradient computation in line 13 always backpropagates through the whole feature extraction network. Alternatively, in the "Accumulated" strategy, we accumulate the gradient with respect to the feature maps in line 13 and backpropagate this gradient to the feature extraction network weights directly before the update in line 19. Additionally, we shift the block boundaries in each training step of the optimization to prevent a bias in the gradients.

Algorithm 1 Memory Efficient Training

```
1:  Initialize $k \leftarrow n$, $\mathcal{G}_{\text{enc}} \leftarrow 0$, $\mathcal{G}_{\text{GRU}} \leftarrow 0$, $\mathcal{L}_{\text{block}} \leftarrow 0$
2:  for $k = 1, ..., \lfloor \frac{n}{m} \rfloor$ do
3:      // Compute forward pass inside block
4:      for $l = 0, ..., m-1$ do
5:          // Update Disparity Estimate
6:          $D^{km+l} \leftarrow D^{km+l-1} + \Delta^{km+l-1}$
7:          // Accumulate loss
8:          $\mathcal{L}_{\text{block}} \leftarrow \mathcal{L}_{\text{block}} + \mathcal{L}(D^{km+l}, D^{GT})$
9:      end for
10:     // Accumulate gradients
11:     for $l = 0, ..., m-1$ do
12:         $\mathcal{G}_{\text{GRU}} \leftarrow \mathcal{G}_{\text{GRU}} + \frac{\partial \mathcal{L}_{\text{block}}}{\partial \theta_{\text{GRU}}}$
13:         $\mathcal{G}_{\text{enc}} \leftarrow \mathcal{G}_{\text{enc}} + \frac{\partial \mathcal{L}_{\text{block}}}{\partial \theta_{\text{enc}}}$
14:     end for
15:     // Discard computational graph
16:     $\mathcal{L}_{\text{block}} \leftarrow 0$
17: end for
18: $\theta_{\text{GRU}} \leftarrow \texttt{update}(\theta_{\text{GRU}}, \mathcal{G}_{\text{GRU}})$
19: $\theta_{\text{enc}} \leftarrow \texttt{update}(\theta_{\text{enc}}, \mathcal{G}_{\text{enc}})$
```

C. Priors Beyond Visual Hull

To further demonstrate the flexibility of our approach, we replaced the Visual Hull prior with a sparse depth map prior, similar to the method proposed by Poggi *et al.* [7], in an additional experiment. In their work, sparse depth measurements are used as prior information for multi-view 3D reconstruction. For our experiment, we simulated this prior by generating a sparse depth map where 20% of pixels contained depth information and applied additional Gaussian noise with a standard deviation of 1 pixel. We pass the known values $\hat{d}_p$ into our method by setting the lower/upper disparity limit as $b_p = (b_p^{\text{min}}, b_p^{\text{max}}) = (\hat{d}_p - 0.5, \hat{d}_p + 0.5)$. The training of our pipeline followed a similar setup to the experiments in Tab. 2 of the original paper. Fig. 2 shows the qualitative results of the predicted disparity from samples of the Poly Haven test set, alongside their estimated depth priors and the predicted baseline disparity for comparison. The quantitative results are reported in Tab. 1, which demonstrates the positive effect of incorporating the depth prior on reducing the resulting error values.

Prior	EPE_{all}	EPE_{noc}	$> 4\text{px}_{\text{all}}$	D1_{all}
No Prior	1.48	0.83	4.6	0.93
Sparse Depth Map	0.84	0.61	1.8	0.24

Table 1. Comparison of our method on the Poly Haven test set, without prior and with sparse noisy depth measurements as prior.

D. Experiments on Spring Dataset

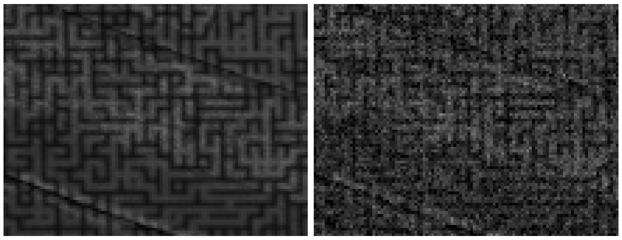
We performed additional experiments on the publicly available Spring dataset [6], which contains high-resolution images. In these experiments, we trained our model for 100,000 iterations using a combined dataset composed of 50% Sceneflow and 50% Spring data.

We upsampled the Spring dataset to $4K$ resolution using bilinear interpolation and trained the model on cropped image patches of size 256 × 1600 from the Spring dataset and 288 × 640 from the Sceneflow dataset. During training, common stereo augmentation techniques were applied, including scaling, flipping, y-jitter, color adjustments, and occlusion inpainting, to improve robustness.

For evaluation, we processed the images at $4K$ resolution, applying downsampling before submitting the results to the benchmark. The evaluation results are shown in Tab. 2 and also available on the official benchmark website[2] where our method ("VHS") ranked third at the time of submission.

E. Robustness

To evaluate our method's robustness, we performed additional experiments on data with increasing noise levels. The experimental setup follows a similar approach to that of the experiments referenced in Tab. 2 of the original paper, with the addition of Gaussian noise at varying levels during the evaluation phase. Please see Fig. 3 for zoomed-in examples of evaluation images with and without noise. The influence of noise on the resulting error values is visualized in Fig. 4, for our method and in comparison for IGEV-Stereo [8] evaluated on half resolution. Please note that because IGEV-Stereo was evaluated at half resolution, the results were also less influenced by the noise applied at full resolution, due to the reduction of noise during the downsampling process.



(a) Without Noise

(b) $\sigma = 0.06$

Figure 3. Zoomed-in reference images: one without noise and the other with added Gaussian noise at a standard deviation of $\sigma = 0.06$.

To demonstrate the robustness of our method regarding real-world data, we evaluated our method on data captured in a dome setup where we predicted the visual hull based on 7 camera views. The resulting visual hull and the predicted disparities of our method compared to IGEV-Stereo are shown in Fig. 5.

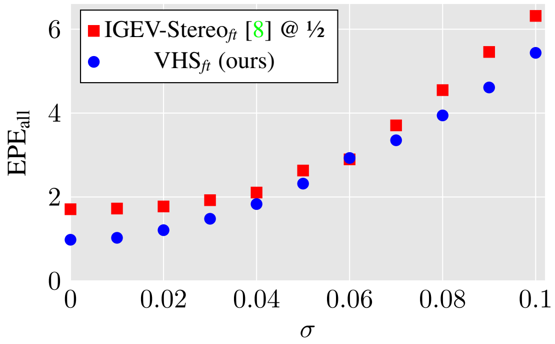


Figure 4. Influence of the noise level (Gaussian noise with standard deviation σ) in the evaluation data on the performance (EPE$_{\text{all}}$).

References

[1] Dan Casas and Marc Comino Trinidad. Smplitex: A generative model and dataset for 3d human texture estimation from single image. *arXiv preprint arXiv:2309.01855*, 2023. 1

[2] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 1

[3] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM TOG*, 38(6):1–19, 2019. 1

[4] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. Mitsuba 3 renderer. 2022. https://mitsuba-renderer.org. 1

[5] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 34(6):1–16, 2015. 1

[6] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 3, 4

[7] Matteo Poggi, Andrea Conti, and Stefano Mattoccia. Multi-view guided multi-view stereo. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8391–8398. IEEE, 2022. 2

[8] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *CVPR*, pages 21919–21928, 2023. 3, 4

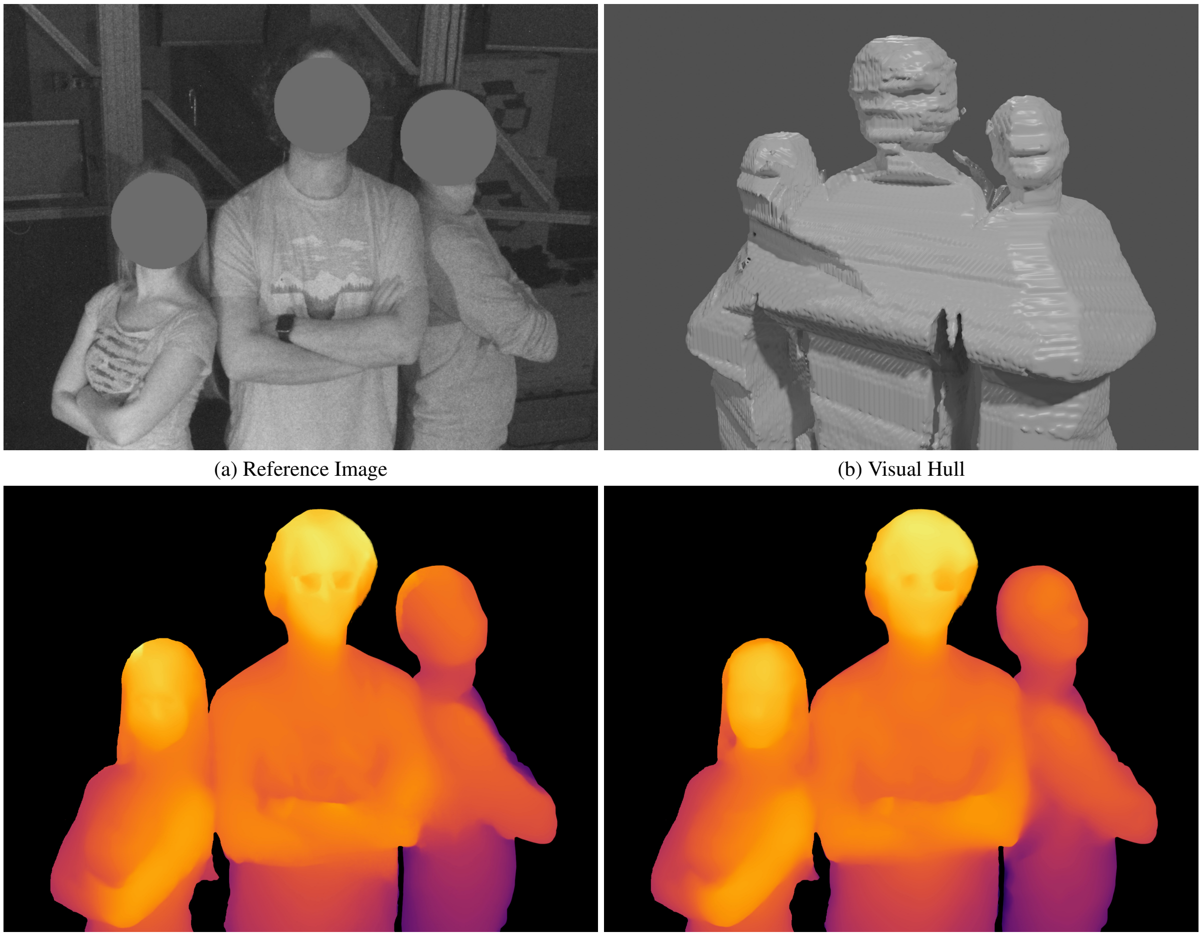
[2] https://spring-benchmark.org/stereo

1px Total	1px Low-detail	1px High-detail	1px Matched	1px Unmatched	1px Not sky	1px Sky	1px S0-10	1px S10-40	1px S40+
13.726	13.368	35.317	11.646	55.022	12.618	30.565	12.106	13.249	18.223

Abs Total	Abs Low-detail	Abs High-detail	Abs Matched	Abs Unmatched	Abs Not sky	Abs Sky	Abs S0-10	Abs S10-40	Abs S40+
4.235	4.125	10.833	2.770	33.318	3.301	18.426	6.896	3.183	2.327

D1 Total	D1 Low-detail	D1 High-detail	D1 Matched	D1 Unmatched	D1 Not sky	D1 Sky	D1 S0-10	D1 S10-40	D1 S40+
5.940	5.863	10.582	4.483	34.874	5.369	14.620	6.341	6.245	4.271

Table 2. Evaluation on the Spring Benchmark [6] measured over 1-pixel error (1px), absolute error (Abs), and disparity error (D1). The metrics are reported across various categorized areas. For further details on the categorization, refer to [6].



(a) Reference Image

(b) Visual Hull

(c) Predicted Disparity by IGEV-Stereo$_{ft}$ [8] @ ½

(d) Predicted Disparity by VHS$_{ft}$ (ours)

Figure 5. Qualitative results from a dome capture setup.