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

Markus Plack Hannah Dröge Leif Van Holland Matthias B. Hullin
University of Bonn
Bonn, Germany
{mplack,droege,holland,hullin}@cs.uni-bonn.de

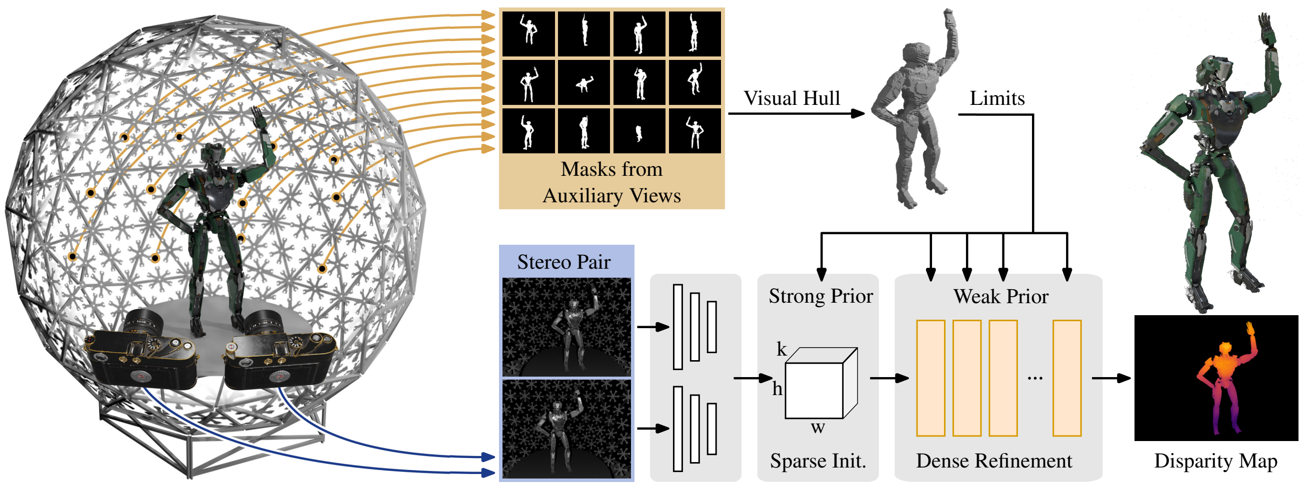

Figure 1. We propose a technique to induce a rough shape estimate from object masks (top) as prior information to a novel, sparse-dense stereo-matching network (bottom) for the application in capture stages (left) for accurate and memory-efficient disparity estimation (right).

Abstract

We present a stereo-matching method for depth estimation from high-resolution images using visual hulls as priors, and a memory-efficient technique for the correlation computation. Our method uses object masks extracted from supplementary views of the scene to guide the disparity estimation, effectively reducing the search space for matches. This approach is specifically tailored to stereo rigs in volumetric capture systems, where an accurate depth plays a key role in the downstream reconstruction task. To enable training and regression at high resolutions targeted by recent systems, our approach extends a sparse correlation computation into a hybrid sparse-dense scheme suitable for application in leading recurrent network architectures.

We evaluate the performance-efficiency trade-off of our method compared to state-of-the-art approaches and demonstrate the efficacy of the visual hull guidance. In addition, we propose a training scheme for a further reduction of memory requirements during optimization, facilitating training on high-resolution data.

1. Introduction

Stereo matching is a long-standing problem in the area of computer vision, driving core functionality in a wide range of applications, for example in the automotive industry, virtual and augmented reality systems, as well as in medical imaging, agriculture, remote sensing, and robotics domains. Recently, interest surged in telepresence and virtual production scenarios that use volumetric capturing systems [6, 11, 15, 33], which rely on fast and accurate depth estimates for downstream reconstruction tasks. The disparity regression problem is typically solved by initially computing the matching cost between a stereo image pair or a suitable feature representation thereof and searching for the best correspondences along the epipolar lines, resulting in a highly irregular cost landscape. Challenges include occlusion, view-dependent reflectivity, repetitive patterns, and insufficient calibration accuracy. With the rise of deep learning in the domain of computer vision, classical matching methods [3, 14, 31, 38] are surpassed by data-driven approaches [13, 17, 28, 51]. Recently, so-called all-pairs-correlation networks based on the optical flow network RAFT [42] have shown to perform remarkably well when applied in the stereo matching context [24]. Those methods compute a dense correlation volume for *all* possible matches and perform stereo regression in an iterative fashion akin to gradient descent methods. One distinct drawback of such approaches is that the size of the full correlation volume scales quadratically with the horizontal input resolution, limiting their applicability on high-resolution inputs. One solution to reduce the prohibitive memory requirement is to use sparse representations [47] that only store the k most relevant entries of the correlation volume, similar to k-nearest-neighbor (kNN) methods. While this still requires the computation of *all* correlation values, which does not reduce the computational costs, the memory demand only scales linearly with respect to the horizontal input resolution, but possibly discards valuable information.

In contrast, we propose a sparse-dense approach that allows us to consider all disparities, avoiding the limitations associated with missing values in sparse representations. We calculate disparities using a sparse method initially, followed by a refinement in a memory-efficient dense manner. As a crucial step to reduce the amount of sparse candidates, we propose to employ the visual hull [20] as a rough shape estimate that reduces the set of valid disparities to points inside the hull. The foreground segmentation masks required for this are available through the use of chroma-keying [35] or more sophisticated image-level segmentation approaches [11] in many capturing scenarios, and thus the visual hull can be computed easily. During the refinement step, we can further use the hull as a weak prior.

In summary, our contributions are as follows:

- We present a method to induce prior knowledge of visual hulls from auxiliary views into a recurrent stereo-matching network to reduce the initial disparity search space and as guidance for the iterative refinement.
- We demonstrate a sparse-dense correlation method that effectively reduces peak memory requirements while retaining the accuracy of all-pairs correlation methods through just-in-time computation for the updates.
- We propose an optimization scheme to realize high-resolution training of recurrent stereo network architectures and show how the visual hull-guided network can benefit from pre-training on conventional training data by making the input optional.

We share the model and training implementation of our **V**isual **H**ull **S**tereo (*VHS*) network and the custom kernels along with the data used for training and testing at `https://github.com/unlikelymaths/vhs`.

2. Related Work

Learning-based methods using correlation volumes to predict accurate disparity maps have shown great potential in stereo matching. We briefly review approaches for generating cost volumes and discuss previous work on further refinement of the disparities by iterative update methods before giving an overview of stereo vision approaches targeting efficiency aspects.

2.1. Matching Cost Volume

Recent developments in end-to-end learning approaches for cost volumes have successfully captured the similarity of pixel pairs across varied degrees of disparity in stereo matching [10, 17, 28, 52].

Correlation-Based Cost Volume In this context, Mayer *et al*. [28] introduced a method based on *correlation* for calculating cost volume, followed by subsequent work [23,43]. This approach measures the correlation between the features of two images within a 1D correlation layer applied horizontally along the disparity line.

Concatenation-Based Cost Volume *Concatenation*-based methods [1, 4, 22, 32], on the other hand, follow a different strategy. Kendall *et al*. [17] concatenated unary features with their corresponding features along the disparity line. They generated a 4D cost volume, subsequently processed through an encoder-decoder network with 3D convolutions across spatial dimensions and disparity. To further regularize the 4D cost volume, Chang *et al*. [5] discussed the implementation of a learned regularization using a stacked hourglass network. Addressing the lack of explicit similarity measures in previous concatenation-based approaches, Guo *et al*. [13] proposed integrating group-wise correlations into the 4D cost volume by dividing features into sub-groups and calculating correlations for each. To improve the performance even in regions with less texture, recent work [49] filters the concatenation volume with attention weights to suppress unnecessary information.

Cascading Cost Volume To overcome storage and runtime limitations, *cascading* cost volumes were created by building a cost volume pyramid and progressively refining depth estimation with a coarse-to-fine technique [10]. Other cascade formulations have been proposed for even higher resolutions [46] or address unbalanced disparity distributions [41].

2.2. Iterative Updates in Stereo Matching

Initially proposed for optical flow estimation, deep learning approaches have successfully employed traditional opti-

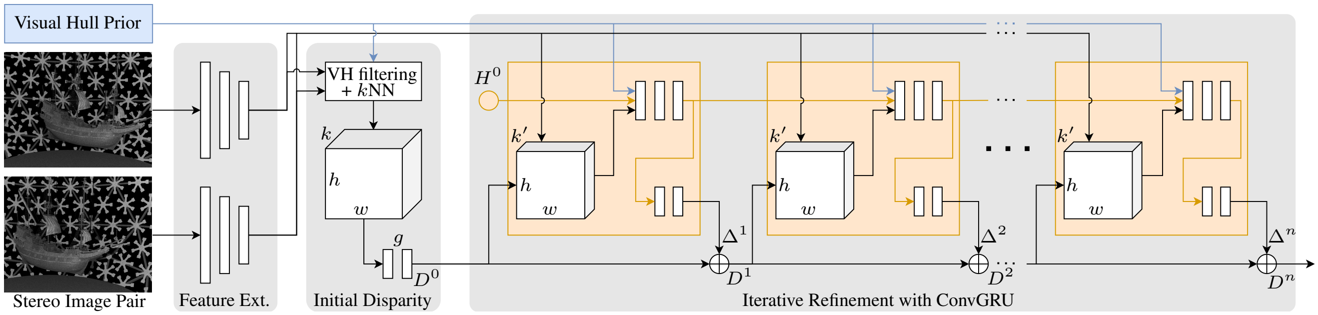

Figure 2. Overview of the three stages of our disparity estimation network VHS. Following the *Feature Extraction* we compute an *Initial Disparity* estimate D_0 from a sparse kNN cost volume restricted by the visual hull. Next, we perform an *Iterative Refinement* of the disparity guided by the visual hull prior using ConvGRU modules and dense local correlations with window size k'.

mization methods using learned updates to improve performance. These methods refine disparity maps through successive updates, as demonstrated by RAFT (Recurrent All-Pairs Field Transforms) [42]. RAFT consists of a feature encoding step, computation of correlation volumes containing the correlations between all pixel pairs, and a learned update operator that iteratively updates the optical flow estimation based on the correlation volumes. Based on this, Lipson *et al*. [24] introduced an adaptation of RAFT for stereo disparity estimation, called RAFT-Stereo, which recurrently updates the disparity map using local cost values.

Several works introduced modifications to this idea. IGEV-Stereo [50] introduces the geometry encoding volume to extend the all-pairs correlation volume and regress a better initial disparity. Instead of using the GRU to update the flow field, Wang *et al*. [45] repurposed it to predict the depth probability of each pixel. Zhao *et al*. [53] propose improvements in the iterative process to preserve detail in the hidden state by decoupling the disparity map from the hidden state and implementing a normalization strategy to handle large variations in disparities. EAI-Stereo [54] replaced the GRU with an error-aware iterative module.

Further work [19, 34, 36] uses global geometric priors to guide and iteratively refine depth estimations.

2.3. Efficiency

In a structured light setting [21, 27, 44], projected patterns are designed to uniquely identify the depth of objects at each position. Hence, the problem can be solved more efficiently for known light patterns, as demonstrated by *e.g.* Hyperdepth [8] using a random forest approach and the branching network in Gigadepth [39]. Note that this is different from our setting based on the work of Guo *et al*. [11] where multiple, potentially overlapping, patterns are projected into the scene.

Turning to wider stereo vision challenges, the bottleneck with cost volumes is their large search space, which requires considerable computation and storage to find the desired disparity. Khamis *et al*. [18] reduced the computational cost by refining the disparity from a low-resolution cost volume through multiple levels of resolution. Additionally, recent works [2, 48] stress real-time disparity estimation in stereo vision. While Shamsafar *et al*. [40] relies on lightweight architectures to optimize resources, Garrepalli *et al*. introduced DIFT [9] as a mobile architecture for optical flow that uses just-in-time computation of the correlation to reduce peak memory use and served as the inspiration for our correlation computation in the iterative updates. SCV-Net [26] builds a sparse correlation volume that resembles dilated convolutions controlled via a fixed sparsity value and without dependence on the inputs. Lastly, SCV-Stereo [47] is an alternative approach to sparse correlation volumes. Different from their method, we use kNN correlation for the initial disparity estimate instead of zero initialization and compute dense correlations on an ad hoc basis during the iterative stages.

3. Visual Hull Stereo

The overall structure of our method is based on RAFT-Stereo [24] and is shown in Fig. 2. It consists of three stages. First, the pair of input images is encoded into a feature representation using a pre-trained encoding network. These features are then used to compute an initial correlation cost volume. Together with prior information attained from a set of image masks of the scene, a sparse set of k disparities with the highest correlation values is selected from which an initial disparity value is estimated (Secs. 3.1 and 3.2). Following, the disparity is iteratively refined using a *Convolutional Gated Recurrent Unit* (ConvGRU)-based network and upsampling network [50], without the need to hold the full cost volume in memory at any time (Sec. 3.3) , which together with the sparse set of disparities we call the sparse-dense correlation network.

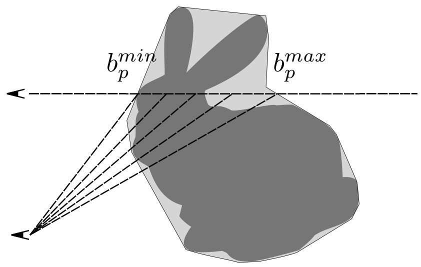

Figure 3. Estimation of the disparity boundaries (b_p^{min}, b_p^{max}), from two rectified views of an object's visual hull. The visual hull encloses the objects' surface, so the surface is guaranteed to lie within the disparity boundaries.

3.1. Sparse Correlation

Given a rectified stereo pair, we use a shared feature encoding network [50] to extract features at 25% of the original image size. This representation is used to compute an initial set of the k best matches. First, we define the correlation value $c_p(d) \in \mathbb{R}$ of disparity $d \in [0, w]$ at pixel $p \in \mathbb{N}^2$ as the inner product of the corresponding feature vectors $f_p, g_{p-(0,d)^T}$, from the left and right pictures of size $h \times w$, where $g_{p-(0,d)^T}$ represents the feature vector at the pixel in the right image offset by d:

$$c_p(d) = f_p \cdot g_{p-(0,d)^T} \tag{1}$$

Storing the full set of correlation values at high resolutions can be inefficient and resource-intensive, as the dense cost volume scales quadratically with the image width when the maximal disparity is properly adjusted. To decrease the memory requirements, we instead use a sparse correlation cost volume, which assigns to each pixel p a much smaller subset of correlation values c and corresponding disparity values d,

$$\mathcal{M}_p = \{(d, c_p(d)) \,|\, d \in \mathcal{D}_p^{k\text{NN}}\}, \tag{2}$$

where $\mathcal{D}_p^{k\text{NN}}$ represents the set of k best disparities for each pixel:

$$\mathcal{D}_p^{k\text{NN}} = \underset{\tilde{\mathcal{D}}_p \subset \mathcal{D}_p, |\tilde{\mathcal{D}}_p| = K}{\arg\max} \sum_{d \in \tilde{\mathcal{D}}_p} c_p(d) \tag{3}$$

Here, $\mathcal{D}_p$ is the set of all disparity candidates for pixel p.

3.2. Visual Hull Prior

This search for the best candidates can be further improved by inducing a prior based on image masks from the scene. The visual hull, as defined in [20], provides an efficient approximation of an object's shape derived from silhouettes captured by multiple cameras. In adherence to the representation proposed in [37], we compute the visual hull using a collection of masked input images, which is stored within an octree structure for compact storage and fast access. The octree is designed such that each leaf node indicates whether it is inside or outside the visual hull. Given this information, we calculate the hull boundaries by sampling rays projected into the scene from the reference view and evaluating these rays for transitions between outside and inside regions of objects. From these transitions, we create depth limits for each camera viewpoint and define disparity boundaries $b_p = (b_p^{min}, b_p^{max})$ based on pixel location p, as illustrated in Figure 3. The insight that the surfaces of the objects are confined within the interval $[b_p^{\min}, b_p^{\max}]$ can be leveraged to reduce computational requirements when computing the initial disparity map D^0.

We streamline the k-nearest-neighbor search, previously performed across an expansive set of disparity candidates $\mathcal{D}_p$ for pixel p as described in (3), by focusing only on disparities constrained within b_p:

$$\mathcal{D}_p^* = \{d \,|\, b_p^{min} \leq d \leq b_p^{max}\}, \qquad \mathcal{D}_p^* \subseteq \mathcal{D}_p \tag{4}$$

This approach allows for a faster computation of the restricted correlation cost volume $\mathcal{M}_p^*$ by skipping unnecessary evaluations of the correlation. Accordingly, we define our initial disparity map as follows:

$$D_p^0 = \sum_{l=1}^{K} d_l \cdot g(c_p(d))_l, \qquad (d, c_p(d)) \in \mathcal{M}_p^* \tag{5}$$

where g is an attention-based transformation network with a softmax function as the last layer.

3.3. Dense Iterative Disparity Refinement

We use a hierarchical ConvGRU network on three resolutions to iteratively refine the predicted disparities starting with the initial values D_p^0, similar to [50]: The network updates a hidden state H^i taking the current disparity values and contextual features extracted from the corresponding image data, and the correlated features around the current disparity estimate as input. The new state is used to predict an offset Δ_p^i from which the refined disparity values are computed as

$$D_p^{i+1} = D_p^i + \Delta_p^i. \tag{6}$$

Memory Efficient Correlation Instead of sampling correlation values from a pre-computed full cost volume, we propose to compute a local correlation volume ad hoc to reduce memory usage. This volume is bounded within a window W_p^i of size $2r + 1$, which is centered on the currently estimated disparity D_p^i,

$$W_p^i = [D_p^i - r, D_p^i + r], \tag{7}$$

where we fix $r = 4$, following [50]. We compute the correlations group-wise, as originally proposed by [13], by dividing the feature vectors into a set of subgroups. Please note that, for the initial disparity D_p^0, we strategically omitted the group-wise correlation calculation. This is due to the complexity of uniquely defining kNN for group-wise correlations, ensuring that our approach remains computationally efficient.

Visual Hull as Weak Prior As additional information, we supply the ConvGRU with a flag $v_p(d)$ that guides the network to predict a value within the visual hull,

$$v_p(d) = \begin{cases} 1 & \text{if } d \in D_p^*, \\ -1 & \text{otherwise} \end{cases} \tag{8}$$

for each disparity value d within the window W_p^i by concatenating this information with the other inputs. In that way, the limits b_p obtained from the visual hull operate as a weak prior guiding the disparity regression while retaining valuable correlation information for cases such as incorrect limits due to masking errors.

One distinct advantage of our visual hull guidance is that the disparity limits are an optional input to the whole pipeline. During the initial sparse correlation, we can fall back to sampling from all values below a pre-defined threshold in the same manner as established models, and during the dense updates, setting $v_p(d) = 0$ lets the network focus solely on correlation information without the additional constraint provided by the visual hull. This enables the application of our sparse correlation method even without masked measurements and pre-training of our method on existing datasets.

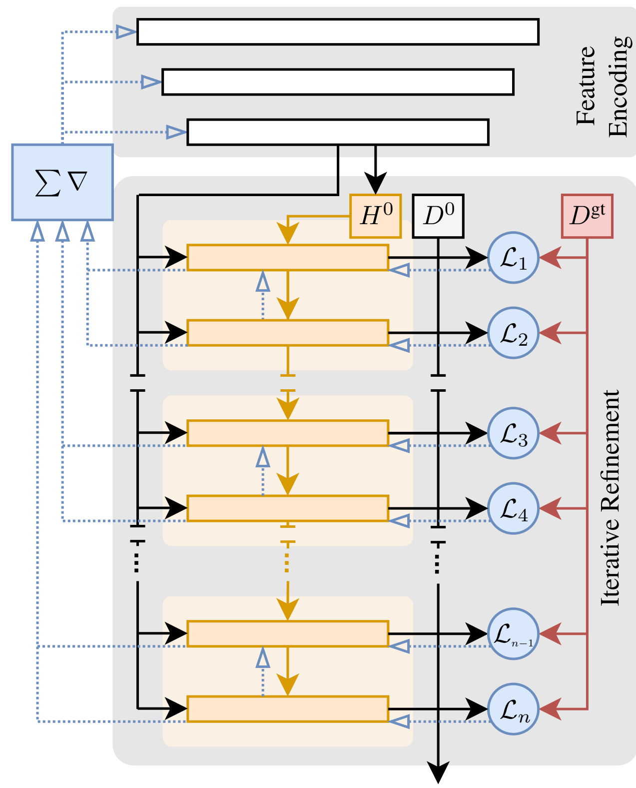

Figure 4. Memory efficient training scheme for $m = 2$ consecutive update steps. After the computation of the losses $\mathcal{L}_i$ and $\mathcal{L}_{i+1}$, we perform backpropagation to accumulate gradients of the update network parameters and detach the hidden state effectively freeing the computational graph. $\sum \nabla$ indicates an optional accumulation of gradients to avoid multiple backward passes through the feature extraction network.

4. Training Details

Given the particular nature of our method in terms of target application and required inputs, a boilerplate training procedure following the literature would be unproductive. Therefore, we present custom training details tailored to our use case, covering the preparation of custom data along with training strategies. We further introduce a memory-efficient approach, enabling training at even higher resolutions.

Common stereo datasets like SceneFlow [28] do not contain ground truth meshes or auxiliary views, which prevents the extraction of a meaningful visual hull. As an alternative, we render a custom dataset with Mitsuba 3 [16] and meshes from Objaverse-XL [7], which we call “FlyingObjaverse”, to train our network. The dataset generation loosely follows the approach of SceneFlow by placing objects on a virtual capture stage. For a detailed explanation explanation, please refer to the supplementary material.

4.1. Training Strategy

Having the visual hull guidance as an entirely optional component, allows our method to harness a more flexible training process and to predict the disparity map even without any pre-calculated masks. We use this flexibility in our experiments by pre-training a base model on SceneFlow and subsequently fine-tuning the network on our custom training data. The training is performed on SceneFlow *finalpass* for 20 epochs using AdamW [25] with a one-cycle learning rate schedule with a learning rate of 0.00015 and a batch size of 4. We use random crops of size 288×640, random y-jitter and occlusion as augmentation, and an L_1 loss following the weighting of RAFT-Stereo [24]. This model serves as our baseline for a benchmark evaluation on the SceneFlow test set. Subsequently, the network is fine-tuned on the FlyingObjaverse data for high-resolution stereo following the same settings, except for a magnified random cropping of 256×2048, batch size of 1 and with the additional visual hull inputs, which we randomly drop for $\frac{1}{8}$ of the samples. Note that we use RGB inputs for the benchmark comparison and grayscale for the simulation of IR images for all other experiments.

Method	#Params	$\text{EPE}_{\leq 192}$	EPE_{all}
CascadeStereo [10]	10.5M	0.67	3.30
CFNet [41]	23.0M	0.96	3.06
CoExNet [2]	3.5M	0.69	3.36
FADNet++ [48]	12.4M	0.88	3.55
GwcNet [13]	6.9M	0.76	3.52
IGEV-Stereo [50]	12.6M	0.48	3.01
MSNet2D [40]	2.3M	1.11	3.76
MSNet3D [40]	1.8M	0.79	3.44
PSMNet [5]	5.2M	1.02	3.69
VHS (ours)	12.7M	0.89	2.33

Table 1. Comparison on SceneFlow *finalpass* benchmark using the model implementations from [12]. All methods were trained on the SceneFlow dataset, without applying any additional priors. For details on additional experimental results on the Spring dataset [29], please refer to the supplementary material.

Memory Efficient Training During the training of most iterative methods, each refinement of the disparity consumes more VRAM since the full compute graph needs to be stored in memory. We propose to split the forward and backward computation in a manner that reduces the memory requirement while still retaining accurate gradient information, as shown in Fig. 4. For m consecutive refinement steps, we run the forward pass and compute the losses on the upscaled disparity predictions as usual. Then, we backpropagate the partial loss and erase the computational graph used to compute the gradients. To avoid multiple backward passes through the costly feature extraction network, we propose to optionally accumulate all gradients for the feature vectors first before performing a final backpropagation after all iterations are through. A detailed description of this procedure can be found in the supplementary material.

Technical Details Using CUDA, we build a visual hull octree from rendered masks from which the disparity limits are computed. Our network is implemented in PyTorch with custom CUDA kernels for the correlation computations, and we use warp-level shuffle operations to make the initial kNN correlation computation efficient. As such, the number of candidates is limited to 32, but we use 8 for all experiments following [47]. All our experiments were conducted on an NVIDIA GeForce RTX 4090, unless stated otherwise.

5. Experiments

We evaluate our method in terms of average end-point error (EPE) in pixels, proportion of errors ($>$ 4px in %) and the D1 outlier rate [30]. Runtime and video memory

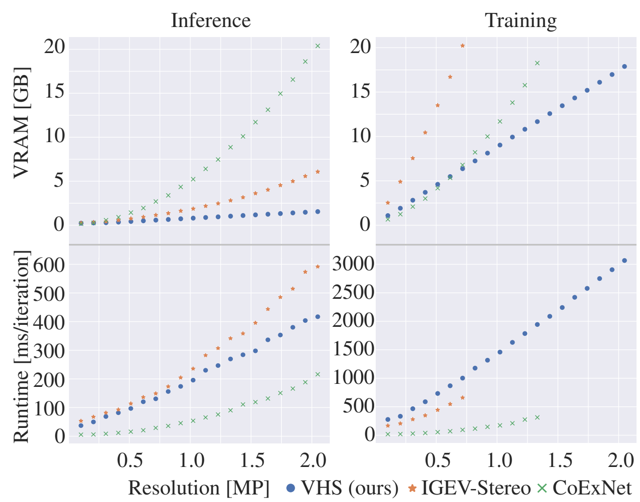

Figure 5. Memory and runtime statistics of our method compared to the best-performing (IGEV) and fastest (CoExNet) baseline methods. We fix the image height at 320 px and increase the width, adjusting the maximum disparity to $\frac{1}{4}$ of the latter.

measurements follow the literature and employ automatic mixed precision.

5.1. Benchmark Evaluation

We first validate the correctness of our sparse-dense correlation network compared to the state-of-the-art, with all methods being trained on SceneFlow. Tab. 1 shows that our method performs competitively in terms of EPE for disparities within the range that all methods can handle. Specifically, for pixels with true disparities less than or equal to 192 ($\text{EPE}_{\leq 192}$), our method matches with FADNet++ [48], with only three methods achieving better scores. Notably, when evaluated on all pixels (EPE_{all}), our method surpasses all baseline models as we do not have any upper limit to the possible disparity, even without any visual hull prior.

Also, our method requires less memory during both inference and training as shown in Fig. 5 and is as fast as IGEV-Stereo [50] during inference while having a minor runtime overhead during training.

5.2. Visual Hull Guidance

To further demonstrate our performance on high-resolution data with larger disparities using the additional visual hull input, we evaluate our method on the two test datasets after fine-tuning on the training dataset as described in Sec. 4. As shown in Tab. 2, our method outperforms all other methods on both the Poly Haven[1] and SMPL datasets across all metrics. Specifically, we achieve significantly lower EPE_{all} and EPE_{noc} which indicates higher overall accuracy, and a higher accuracy in non-occluded regions. We

[1] https://polyhaven.com/

Method	Poly Haven				SMPL			
	$\text{EPE}_{\text{all}}(\downarrow)$	$\text{EPE}_{\text{noc}}(\downarrow)$	$> 4\text{px}_{\text{all}}(\downarrow)$	$\text{D1}_{\text{all}}(\downarrow)$	$\text{EPE}_{\text{all}}(\downarrow)$	$\text{EPE}_{\text{noc}}(\downarrow)$	$> 4\text{px}_{\text{all}}(\downarrow)$	$\text{D1}_{\text{all}}(\downarrow)$
CascadeStereo [10]* @ ¼	16.97	14.37	31.1	6.77	8.31	6.51	13.8	2.97
CFNet [41]* @ ¼	14.50	11.98	31.4	7.80	13.28	12.48	9.8	3.74
CoExNet [2]* @ ½	9.78	8.57	25.9	7.21	2.98	2.38	8.6	1.56
FADNet++ [48]* @ ½	11.44	10.49	25.3	7.82	2.67	1.85	6.8	1.64
GwcNet [13]* @ ¼	19.97	17.04	35.8	9.60	11.27	10.34	14.9	3.86
IGEV-Stereo [50]* @ ½	5.22	4.10	16.6	3.94	1.68	1.27	6.2	0.83
MSNet2D [40]* @ ¼	10.08	8.69	44.2	5.67	5.24	4.44	28.9	2.38
MSNet3D [40]* @ ¼	14.41	11.95	32.3	7.65	9.78	8.36	12.3	3.40
PSMNet [5]* @ ¼	13.19	11.28	37.8	6.11	17.38	16.31	17.9	4.55
CoExNet$_{ft}$ [2]* @ ½	2.82	2.11	9.7	1.73	0.87	0.63	2.2	0.34
IGEV-Stereo$_{ft}$ [50]* @ ½	1.71	1.04	5.5	1.05	0.63	0.44	1.3	0.24
VHS$_{ft}$ (ours) @ ½	1.55	1.05	5.2	0.58	0.84	0.69	1.4	0.10
IGEV-Stereo$_{ft}$ [50]*	3.23	2.02	8.4	2.39	1.06	0.74	2.8	0.62
VHS$_{ft}$ (ours)	**0.98**	**0.55**	**3.2**	**0.40**	**0.54**	**0.41**	**0.9**	**0.10**

Table 2. Comparison on our data using the model implementations from [12]. Methods marked with * run with inputs aligned to set minimum disparity to zero, with resolution adjustments (half/quarter) indicated where applied. *ft* denotes methods fine-tuned on our FlyingObjaverse dataset.

Prior	EPE_{all}	EPE_{noc}	$> 4\text{px}_{\text{all}}$	D1_{all}
No	1.48	0.83	4.6	0.93
Initial	1.29	0.75	4.3	0.68
Update	1.04	0.57	3.3	0.46
Both	0.98	0.55	3.2	0.40

Table 3. Ablation of the visual hull guidance on the Poly Haven test set.

further highlight the robustness of our method by showing the lowest percentage of pixels with large disparity errors ($> 4\text{px}_{\text{all}}$, D1_{all}). We present qualitative results in Fig. 6. Note that the memory usage of previous methods scales with the size of the input in such a way that most models cannot perform inference at full resolution, even using server-grade GPUs. We ran the finetuning for the two best performing methods, CoExNet and IGEV-Stereo, at half and full resolution on an NVIDIA A100 GPU with 80GB memory, but CoExNet memory requirements still exceeded the available VRAM for the full resolution. Additionally, most methods cannot capture the large disparity values in our data as the correlation volumes are typically limited to 192 pixels. For this evaluation, we resort to running the models on $2\times$ or $4\times$ downsampled input images and reduce the offsets by aligning them using the known minimum ground-truth disparity of the foreground, selecting the best variant of both resolutions based on the smallest EPE.

To study the performance benefit of the visual hull, we perform an ablation study on the Poly Haven test set, as shown in Tab. 3. While applying visual hull guidance only for the initial disparity calculation ("Initial") already shows a minor improvement across all metrics compared to an uninformed run, the weak prior during the iterative updates ("Update") yields a major gain. Ultimately, we achieved the best results by employing visual hull guidance in both phases ("Both"). The improvement is particularly remarkable considering that the majority of the object points do not lie directly on the visual hull.

As the quality of the visual hull depends on the correctness of the masks, we additionally study the influence of incorrect matting on the performance of our method in Fig. 7. We find that our method is robust against binary dilation on the masks, while larger binary erosion reduces the accuracy. Intuitively, this makes sense as a correct visual hull always encloses the true surface, which is also the case for "inflated" visual hulls from dilated masks, while "deflated" hulls from eroded masks violate this assumption.

5.3. Training Scheme

To evaluate the impact of the memory-efficient training scheme on memory usage and runtime, we estimated these metrics for different numbers of connected updates before backpropagation in relation to the standard training procedure. We compared a setting with repeated backpropagation and weight update of the feature extractor ("Repeated") to a setting where the feature gradients are accumulated and the weights are only updated once all connected updates are computed ("Accumulated"). For the former, we measured a reduction in memory usage at the cost of increased runtime for a smaller number of connected updates m, as shown in Tab. 4. In comparison, the detached features offer a stable runtime even at as few as two connected layers, with an even further reduction in memory usage compared to full backpropagation. We did not observe any difference in error metrics between our proposed memory-efficient training

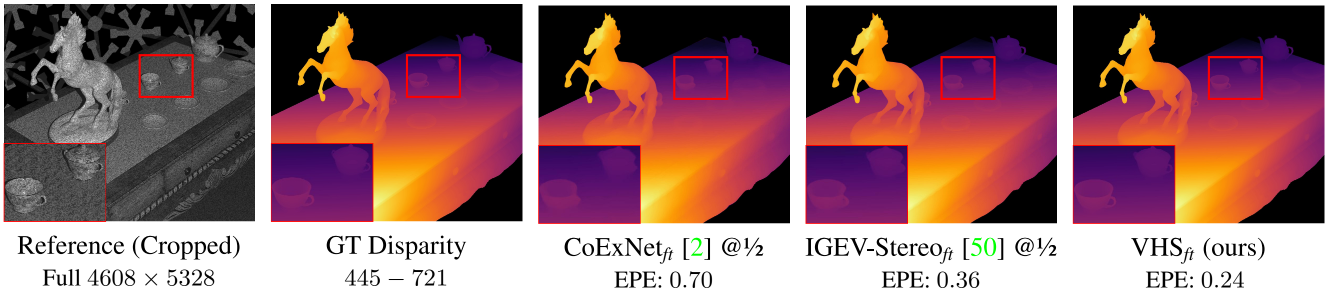

Figure 6. Qualitative results on samples from the Poly Haven. Note the faithful reconstruction of the tableware produced by our method. We show the range of disparity values below the GT disparity and the EPE below the methods.

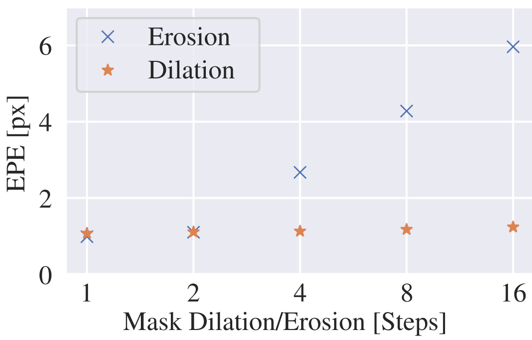

Figure 7. Correlation between mask accuracy and EPE, demonstrating robustness to binary dilations to the correct mask.

Variant	Repeated GB	Repeated ms	Accumulated GB	Accumulated ms
-	14.18	377	-	-
$m = 16$	8.71	441	8.49	586
$m = 8$	5.85	497	5.62	584
$m = 4$	4.42	611	4.19	583
$m = 2$	3.69	840	3.46	583

Table 4. Peak memory and average runtime per iteration comparing the standard training procedure (first row) with our proposed memory-efficient training running backpropagation and updating the weights of the feature extractor each time (left) and accumulating the feature gradients first (right) for different numbers of connected disparity updates. Updates for the ConvGRU are accumulated in both cases. Measured for a single stereo pair at 512×1024.

method, using $m = 4$ connected updates with detached features, and the standard training procedure.

Finally, we evaluate the impact of pre-training on SceneFlow in our training procedure. A network trained using only our FlyingObjaverse dataset yields an EPE of 1.33 on the Poly Haven test set, compared to 0.98 of a training on both SceneFlow and FlyingObjaverse, indicating a significant benefit of the hybrid approach.

5.4. Generality and Robustness

While the visual hull integration is limited to specific use cases where other views of the scene are given, our method of incorporating prior information is flexible and can accommodate other priors. As proof of concept, we show that replacing the visual hull prior with a sparse depth map prior decreases all error metrics compared to not using a prior. Those additional measurements were proposed by Poggi *et al.* [34] and we incorporate them by setting b_p^{min} and b_p^{max} to the measured disparity where given. The details and results are reported in the supplementary material.

Additionally, we evaluated our performance on data with varying noise levels, to estimate the robustness of our method, and test it on real-world data for a qualitative comparison. For more details and results, please refer to the supplementary material.

6. Conclusion

We have presented a technique to induce visual hull priors into recurrent stereo networks to improve matching performance. Combined with a novel sparse-dense correlation handling, our approach accurately regresses disparity for high-resolution images while retaining a favorable memory footprint and without an upper limit on the achievable disparity.

Acknowledgements

This work has been funded by the Ministry of Culture and Science North Rhine-Westphalia under grant number PB22-063A (InVirtuo 4.0: Experimental Research in Virtual Environments), and by the state of North Rhine-Westphalia as part of the Excellency Start-up Center.NRW (U-BO-GROW) under grant number 03ESCNW18B. Leif Van Holland acknowledges the support of the German Research Foundation (DFG) grant KL 1142/11-2 (DFG Research Unit FOR 2535 Anticipating Human Behavior).

References

[1] Shamsul Fakhar Abd Gani, Muhammad Fahmi Miskon, Rostam Affendi Hamzah, Mohd Saad Hamid, Ahmad Fauzan Kadmin, and Adi Irwan Herman. Refining disparity maps using deep learning and edge-aware smoothing filter. *Bulletin of Electrical Engineering and Informatics*, 13(3):1961–1969, 2024. 2

[2] Antyanta Bangunharcana, Jae Won Cho, Seokju Lee, In So Kweon, Kyung-Soo Kim, and Soohyun Kim. Correlate-and-excite: Real-time stereo matching via guided cost volume excitation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3542–3548. IEEE, 2021. 3, 6, 7, 8

[3] Stephen T. Barnard and William B. Thompson. Disparity analysis of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, (4):333–340, 1980. 1

[4] Rohan Chabra, Julian Straub, Christopher Sweeney, Richard Newcombe, and Henry Fuchs. Stereodrnet: Dilated residual stereonet. In *CVPR*, pages 11786–11795, 2019. 2

[5] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, pages 5410–5418, 2018. 2, 6, 7

[6] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM TOG*, 34(4):1–13, 2015. 1

[7] 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. 5

[8] Sean Ryan Fanello, Christoph Rhemann, Vladimir Tankovich, Adarsh Kowdle, Sergio Orts Escolano, David Kim, and Shahram Izadi. Hyperdepth: Learning depth from structured light without matching. In *CVPR*, pages 5441–5450, 2016. 3

[9] Risheek Garrepalli, Jisoo Jeong, Rajeswaran C. Ravindran, Jamie Menjay Lin, and Fatih Porikli. Dift: Dynamic iterative field transforms for memory efficient optical flow. In *CVPR*, pages 2219–2228, 2023. 3

[10] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *CVPR*, pages 2495–2504, 2020. 2, 6, 7

[11] 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, 2, 3

[12] Xianda Guo, Juntao Lu, Chenming Zhang, Yiqi Wang, Yiqun Duan, Tian Yang, Zheng Zhu, and Long Chen. Openstereo: A comprehensive benchmark for stereo matching and strong baseline. *arXiv preprint arXiv:2312.00343*, 2023. 6, 7

[13] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *CVPR*, pages 3273–3282, 2019. 1, 2, 4, 6, 7

[14] Rostam Affendi Hamzah and Haidi Ibrahim. Literature survey on stereo vision disparity map algorithms. *Journal of Sensors*, 2016, 2016. 1

[15] Jonathan Heagerty, Sida Li, Eric Lee, Shuvra Bhattacharyya, Sujal Bista, Barbara Brawn, Brandon Y. Feng, Susmija Jabbireddy, Joseph JaJa, Hernisa Kacorri, et al. Holocamera: Advanced volumetric capture for cinematic-quality vr applications. *IEEE TVCG*, 2024. 1

[16] 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. 5

[17] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *ICCV*, pages 66–75, 2017. 1, 2

[18] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In *ECCV*, pages 573–590, 2018. 3

[19] Numair Khan, Eric Penner, Douglas Lanman, and Lei Xiao. Temporally consistent online depth estimation using point-based fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9119–9129, 2023. 3

[20] Aldo Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(2):150–162, 1994. 2, 4

[21] Jacqueline LeMoigne and Allen Mark Waxman. Structured light patterns for robot mobility. *IEEE Journal on Robotics and Automation*, 4(5):541–548, 1988. 3

[22] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *CVPR*, pages 16263–16272, 2022. 2

[23] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In *CVPR*, pages 2811–2820, 2018. 2

[24] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *Int. Conf. on 3D Vision*, pages 218–227. IEEE, 2021. 2, 3, 5

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[26] Chuanhua Lu, Hideaki Uchiyama, Diego Thomas, Atsushi Shimada, and Rin-ichiro Taniguchi. Sparse cost volume for efficient stereo matching. *Remote sensing*, 10(11):1844, 2018. 3

[27] Manuel Martinez and Rainer Stiefelhagen. Kinect unleashed: Getting control over high resolution depth maps. In *MVA*, pages 247–250, 2013. 3

[28] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A

large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, pages 4040–4048, 2016. 1, 2, 5

[29] 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. 6

[30] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, pages 3061–3070, 2015. 6

[31] Karsten Mühlmann, Dennis Maier, Jürgen Hesser, and Reinhard Männer. Calculating dense disparity maps from color stereo images, an efficient implementation. *International Journal of Computer Vision*, 47:79–88, 2002. 1

[32] Guang-Yu Nie, Ming-Ming Cheng, Yun Liu, Zhengfa Liang, Deng-Ping Fan, Yue Liu, and Yongtian Wang. Multi-level context ultra-aggregation for stereo matching. In *CVPR*, pages 3283–3291, 2019. 2

[33] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L. Davidson, Sameh Khamis, Mingsong Dou, et al. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th annual symposium on user interface software and technology*, pages 741–754, 2016. 1

[34] 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. 3, 8

[35] Carolus Raditya, Muhammad Rizky, Sergio Mayranio, and Benfano Soewito. The effectivity of color for chroma-key techniques. *Procedia Computer Science*, 179:281–288, 2021. 2

[36] Alexander Rich, Noah Stier, Pradeep Sen, and Tobias Höllerer. 3dvnet: Multi-view depth prediction and volumetric refinement. In *2021 International Conference on 3D Vision (3DV)*, pages 700–709. IEEE, 2021. 3

[37] Hanno Scharr, Christoph Briese, Patrick Embgenbroich, Andreas Fischbach, Fabio Fiorani, and Mark Müller-Linow. Fast high resolution volume carving for 3d plant shoot reconstruction. *Frontiers in plant science*, 8:303692, 2017. 4

[38] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. In *CVPR*, volume 1, pages I–I. IEEE, 2003. 1

[39] Simon Schreiberhuber, Jean-Baptiste Weibel, Timothy Patten, and Markus Vincze. Gigadepth: Learning depth from structured light with branching neural networks. In *ECCV*, pages 214–229. Springer, 2022. 3

[40] Faranak Shamsafar, Samuel Woerz, Rafia Rahim, and Andreas Zell. Mobilestereonet: Towards lightweight deep networks for stereo matching. In *IEEE/CVF Winter Conf. on App. of Comput. Vision*, pages 2417–2426, 2022. 3, 6, 7

[41] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *CVPR*, pages 13906–13915, 2021. 2, 6, 7

[42] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419. Springer, 2020. 2, 3

[43] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *CVPR*, pages 195–204, 2019. 2

[44] Piet Vuylsteke and André Oosterlinck. Range image acquisition with a single binary-encoded light pattern. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(2):148–164, 1990. 3

[45] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. Itermvs: Iterative probability estimation for efficient multi-view stereo. In *CVPR*, pages 8606–8615, 2022. 3

[46] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *CVPR*, pages 14194–14203, 2021. 2

[47] Hengli Wang, Rui Fan, and Ming Liu. Scv-stereo: Learning stereo matching from a sparse cost volume. In *ICIP*, pages 3203–3207. IEEE, 2021. 2, 3, 6

[48] Qiang Wang, Shaohuai Shi, Shizhen Zheng, Kaiyong Zhao, and Xiaowen Chu. Fadnet++: Real-time and accurate disparity estimation with configurable networks. *arXiv preprint arXiv:2110.02582*, 2021. 3, 6, 7

[49] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *CVPR*, pages 12981–12990, 2022. 2

[50] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *CVPR*, pages 21919–21928, 2023. 3, 4, 6, 7, 8

[51] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *CVPR*, pages 1959–1968, 2020. 1

[52] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *CVPR*, pages 185–194, 2019. 2

[53] Haoliang Zhao, Huizhou Zhou, Yongjun Zhang, Jie Chen, Yitong Yang, and Yong Zhao. High-frequency stereo matching network. In *CVPR*, pages 1327–1336, 2023. 3

[54] Haoliang Zhao, Huizhou Zhou, Yongjun Zhang, Yong Zhao, Yitong Yang, and Ting Ouyang. Eai-stereo: Error aware iterative network for stereo matching. In *Proceedings of the Asian Conference on Computer Vision*, pages 315–332, 2022. 3