Probabilistic Labeling Cost for High-Accuracy Multi-View Reconstruction

Ilya Kostrikov, Esther Horbert, Bastian Leibe
Computer Vision Group
RWTH Aachen University
ilya.kostrikov@rwth-aachen.de {horbert , leibe}@vision.rwth-aachen.de

Abstract

In this paper, we propose a novel labeling cost for multi-view reconstruction. Existing approaches use data terms with specific weaknesses that are vulnerable to common challenges, such as low-textured regions or specularities. Our new probabilistic method implicitly discards outliers and can be shown to become more exact the closer we get to the true object surface. Our approach achieves top results among all published methods on the Middlebury DINO SPARSE *dataset and also delivers accurate results on several other datasets with widely varying challenges, for which it works in unchanged form.*

1. Introduction

We consider the classical computer vision problem of multi-view stereo, where an object's 3D shape is inferred from a set of calibrated 2D images. If the depth observations from each camera were perfect, multi-view reconstruction would be easy. In reality, the Lambertian assumption does not always hold, and noise and untextured regions make 3D reconstruction still difficult.

A large number of approaches have been proposed in recent years to address those issues [6, 7, 8, 12, 13, 19]. Volumetric methods have been particularly successful towards this goal [12, 13, 19]. Fueled by advances in convex optimization [14, 3], globally optimal formulations have been proposed for the multiview reconstruction problem [12, 13]. However, this line of research has so far mainly focused on the optimization methods. For highly accurate reconstruction results, the labeling cost (the data term in energy formulations) is just as important. As Fig. 1 clearly shows, even the best currently available approaches have major problems in low-textured image areas, leading to visible artifacts in the obtained reconstructions.

In this paper, we start from a formulation motivated by the volumetric approaches of [8, 12]. We present a detailed analysis of the reasons why those approaches have problems in specific challenging regions and we derive new insights into the multiview reconstruction problem from this analysis. Based on those insights, we propose a probabilistically well-founded formulation for the labeling cost that is more robust to outliers and that achieves improved reconstruction results (see Fig. 1). The key idea of our approach is to base the labeling cost on an independently selected subset of the available cameras for each voxel, such that the cameras can be trusted with high probability.

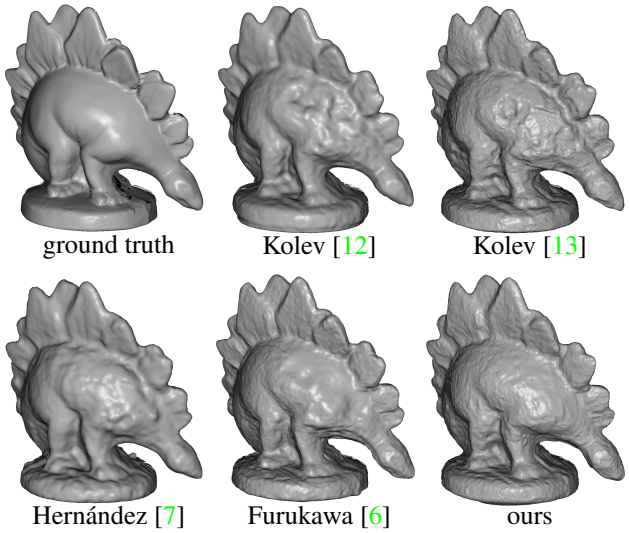


Figure 1. State-of-the-art results for the Middlebury DINO SPARSE dataset, compared to our approach. In contrast to other methods our approach is able to accurately reconstruct regions with high outlier rates, such as low-textured regions.

Our proposed method is elegant, fast, and simple to implement (also simple to port to the GPU). Moreover, it is general since we do not make any specific assumption regarding the shape of the object and works in unchanged form for a variety of datasets that offer different challenges, such as low-textured regions, specularities, concavities, and thin structures. We quantitatively evaluate our approach on the well known Middlebury benchmark [17] and achieve the best result among published methods for the DINO SPARSE dataset and high ranking results for three other datasets. Additionally, we also show qualitative results for difficult

datasets and compare our approach with other methods.

This paper is organized as follows: The next section discusses related work. Section 3 then describes the used energy minimization framework and discusses the specific problems arising in existing methods. We then present our new labeling costs in Section 4. Section C contains details about discretization and optimization of the energy functional. Finally, we show experimental results in Section D.

2. Related Work

The first approaches for multi-view stereo were carving techniques [18], which do not enforce any smoothness and result in quite noisy reconstructions. They were superseded by more elegant energy minimization techniques, which explicitly model smoothness constraints. First were active contours [2], where the shape evolves to photoconsistent locations. In the following years, the used surface representations and optimization techniques were steadily improved, starting with level sets [5], triangle meshes [7], followed by graph cuts [19, 15, 8], continuous global optimization [12], and in the last few years anisotropic measures [13, 16].

In this paper, we focus on volumetric reconstruction approaches [8, 12, 13, 19]. A problem arising in those approaches is that one cannot actually observe the inside of the object. The only observations possible are of the surface. Thus, when only using photoconsistency, which is only present on the surface, the global minimum of the energy functional is the empty solution. This problem was first solved with a ballooning term, which is a data-independent term that adds a constant outward force [19, 15]. This approach has several disadvantages. It is sensitive to oversmoothing and to finding the right balance between data term and ballooning term.

To address this problem, Boykov & Lempitsky [1] proposed a data-aware ballooning force using the surface orientation. Hernández *et al*. [8] proposed a data-aware volumetric term based on probabilistic visibility, where the photoconsistency score is propagated along the camera ray. Their approach treats all voxels as background if at least one camera votes for background, which leads to a systematic underestimation of the surface, as we show in Sec. 3.3. Kolev *et al*. [12] evaluated a number of approaches for setting labeling costs in their papers. They estimate surface normals and select only the cameras that are in front of the surface. As we show in Sec. 4.1, this does not guarantee that the considered part of the surface is visible to the selected cameras and misses important cameras actually seeing the considered part of the surface, while selecting outlier cameras.

While optimization methods were significantly improved over time, a robust labeling is still an open problem. Our approach uses the continuous global optimization framework by [12] and a novel labeling cost that was inspired by [8] but that avoids the problems of those approaches. Our approach selects the cameras with surface observations closest to the considered point to set the labeling costs. We treat the problem in a probabilistic way and obtain foreground and background probabilities that are neither systematically overestimated nor underestimated. Moreover, in our approach the surface estimation even becomes more accurate close to the real surface. In contrast to previous approaches, our approach implicitly selects the cameras seeing the relevant part of the surface and ignores outlier cameras with a high probability.

This idea of camera selection is also very different from the idea of image selection, as proposed by Hornung *et al*. [9]. Their approach selects a subset of *images* for reconstruction to avoid redundant information. In contrast, our method selects *cameras independently for each voxel*.

Table 1. Notation used in this paper

$\mathbf{x}$	3D point
C_i	observations taken from ith camera (depth and photo-consistency score)
$\mathbf{c}_i$	center of camera C_i
$\pi_i(\mathbf{x})$	projection of point $\mathbf{x}$ on ith image
$\sigma(\mathbf{x})$	labeling cost
$\rho(\mathbf{x})$	discontinuity cost
O	short for $O(\mathbf{x})$, observation that point $\mathbf{x}$ is invisible (and thus belongs to the object)
O_i	point $\mathbf{x}$ is visible from camera i
B	point $\mathbf{x}$ is *not* visible from any camera (and thus either background or inside the object)
B_i	point $\mathbf{x}$ is *not* visible from camera i
$S_i(\mathbf{x})$	consistency score for camera i

3. Multi-view Stereo as Energy Minimization

Kolev et al [12] introduced a framework that allows to partition 3D space into 'foreground' (or 'object') and 'background' regions. They propose a globally optimal continuous formulation of the problem that has several advantages compared to other approaches. In this section we will briefly introduce the framework, for more details see [12].

Let $V \in \mathbb{R}^3$ denote the volume that contains the object, $u : V \to \{0, 1\}$ is a characteristic function that implicitly represents the object. $\rho_{obj}(\mathbf{x})$ and $\rho_{bck}(\mathbf{x})$ are labeling costs for foreground/object and background, $\rho(\mathbf{x})$ is a discontinuity cost and $1/\lambda$ is the smoothness parameter. Multi-view reconstruction can then be formulated as the problem of minimizing the following energy:

$$E(u) = \underbrace{\int_V \rho(\mathbf{x})|\nabla u(\mathbf{x})|dx}_{\text{smoothing term}} + \lambda \underbrace{\int_V (\rho_{obj}(\mathbf{x}) - \rho_{bck}(\mathbf{x}))u(\mathbf{x})dx}_{\text{labeling cost}} \quad (1)$$

Let $u_{min}(x) = \arg\min_{u(x)} E(u)$. Then the surface can be extracted via thresholding $u_\nu(x) = \mathbb{1}\{u_{min}(x) \geq \nu\}$. The threshold ν can be set to 0.5 or determined as $\arg\min_\nu E(u_\nu)$.

The discontinuity cost $\rho(\mathbf{x})$ is defined very similar in most of the energy minimization approaches and has been shown to produce stable and reliable results. Defining robust labeling costs on the other hand is still an open problem. Some of the available approaches work well only under special conditions and for specific datasets. Therefore we concentrate on the labeling costs $\rho_{bck}(\mathbf{x})$ and $\rho_{obj}(\mathbf{x})$.

3.1. Depth Estimation

Given a set of images $I_1, I_2, ..., I_N$ with extracted background and calibrated camera parameters, the goal is to calculate the depth of each voxel from every camera. In this section we will briefly describe the process for camera j and voxel $\mathbf{x}$ that lies inside the convex hull.

We define a ray from camera center $\mathbf{c}_j$ to the point $\mathbf{x}$:

$$r_{j,\mathbf{x}}(t) = \mathbf{c}_j + \frac{\mathbf{x} - \mathbf{c}_j}{||\mathbf{x} - \mathbf{c}_j||}t, \tag{2}$$

where the parameter t is the position along the ray and corresponds to a certain distance from the camera. Let $t_{j,\mathbf{x}}$ be the position where the ray reaches the point $\mathbf{x}$: $t_{j,\mathbf{x}} = ||\mathbf{x} - \mathbf{c}_j||$, and thus $r_{j,\mathbf{x}}(t_{j,\mathbf{x}}) = \mathbf{x}$.

The location of the highest photo-consistency along the ray is computed using normalized cross-correlation between square patches around the projection of $\mathbf{x}$ in image j and the projections in the neighboring images:

$$NCC(\pi_i(\mathbf{x}), \pi_j(\mathbf{x})) = n_i \cdot n_j, \tag{3}$$

$$n_i = \frac{I_i(\pi_i(\mathbf{x})) - \overline{I_i(\pi_i(\mathbf{x}))}}{\|I_i(\pi_i(\mathbf{x})) - \overline{I_i(\pi_i(\mathbf{x}))}\|} \tag{4}$$

where $I_i(\pi_i(\mathbf{x}))$ is a vector that contains the intensities of the square patch around the projection of $\mathbf{x}$ on image i and $\overline{I_i(\pi_i(\mathbf{x}))}$ is a vector containing the mean intensity. In all our experiments, we used 7×7 square patches. Then we fuse all these NCC values by taking the weighted average:

$$S_j(\mathbf{x}) = \sum_{k=1}^{m} w_{j,i_k}(\mathbf{x}) NCC(\pi_{i_k}(\mathbf{x}), \pi_j(\mathbf{x})) \tag{5}$$

where, in order to obtain reliable depth estimates, we use only the neighboring cameras for computing NCC scores, as in [12]. Let $\alpha_{i,j}$ be the angle between the normalized viewing directions of cameras i and j. We use only the cameras with $\alpha_{i,j} \leqq \alpha_{max}$ (we set $\alpha_{max} = 45°$ in our experiments) and weight them with

$$w_{j,i}(\mathbf{x}) = \frac{\alpha_{max} - \alpha_{j,i}(\mathbf{x})}{\sum_{k=1}^{m} \alpha_{max} - \alpha_{j,k}(\mathbf{x})}. \tag{6}$$

Finally we choose the position with the maximum score S_j among the positions on the ray:

$$t_{j,\mathbf{x},max} = \arg\max_t S_j(r_{j,\mathbf{x}}(t)), \tag{7}$$

$$S_{j,\mathbf{x},max} = \max_t S_j(r_{j,\mathbf{x}}(t)). \tag{8}$$

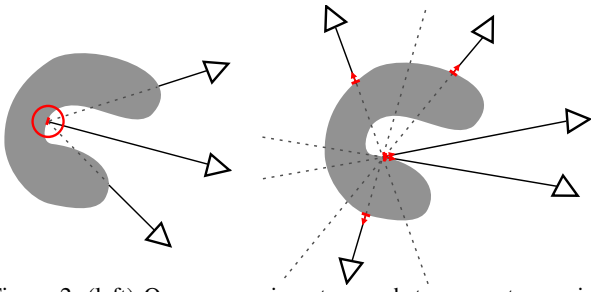


Figure 2. (left) One camera is not enough to compute a point's depth. (right) Observations by multiple cameras. For a given voxel each camera provides a depth observation for the corresponding ray. If the voxel is occluded from the position of the respective camera, the observed depth will be too small.

3.2. Discontinuity Cost

We use the discontinuity cost first proposed in [7]. Each camera gives a vote for point $\mathbf{x}$ only if $\mathbf{x}$ corresponds to the camera ray's intersection with the object surface:

$$VOTE_j(\mathbf{x}) = \begin{cases} S_j(\mathbf{x}) & \text{if } t_{j,\mathbf{x},max} = t_{j,\mathbf{x}} \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

and we accumulate the votes from all cameras:

$$\rho(\mathbf{x}) = e^{-\mu \sum_{j=1}^{N} VOTE_j(\mathbf{x})}. \tag{10}$$

where $\mu = 0.15$ as in [12]. This scheme is widely used in a number of approaches and has proven to be robust and to yield precise photo-consistency maps [12, 7, 8, 19].

3.3. Labeling Cost

For the labeling cost, we need to determine probabilities for foreground and for background for each voxel. We can then define the labeling cost as the following negative log probabilities:

$$\rho_{obj}(\mathbf{x}) = -\log P(O|C_1 \cdots C_N), \tag{11}$$

$$\rho_{bck}(\mathbf{x}) = -\log P(B|C_1 \cdots C_N). \tag{12}$$

where O corresponds to the event that point $\mathbf{x}$ belongs to the object and B to background. Since the true surface is unknown, we cannot compute (11) directly. However, we can introduce random variables $o_1, \cdots, o_N$ ($o_i \in \{O_i, B_i\}$). O_i means that camera i would label the voxel as object and B_i as background, respectively. Then we can decompose (11) using the law of total probability:

$$P(O|C_1 \cdots C_N) = \Sigma_{o_i \in \{O_i, B_i\}} P(O|o_1 \cdots o_N) P(o_1 \cdots o_N|C_1 \cdots C_N). \tag{13}$$

Hernández *et al.* [8] used the assumption that a point belongs to the foreground if and only if all cameras agree that it is foreground: $P(O|O_1 \cdots O_N) = 1$ (*c.f.* eq. (13)). All other combinations are attributed to background: $P(O|o_1 \cdots o_N) = 0$. However, if point $\mathbf{x}$ is visible

only from one camera then we cannot estimate $t_{j,\mathbf{x},max}$ in (8) because it requires its projections on other images to estimate the depth (*c.f*. Fig. 2 (left)). Thus, if a voxel is observed to be foreground by only one camera, this observation is very likely an outlier caused by noise (see also Fig. 2 (left)).

Therefore, if a point belongs to the background according to only one camera without support from a neighboring camera, this configuration is clearly wrong. In this case, we cannot be sure if this camera is an outlier or if a neighboring camera is the outlier, but in [8], all these impossible configurations are assumed to belong to the background. Since $P(O|O_1 \cdots O_{k-1} B_k O_{k+1} \cdots O_N) > 0$, a considerable amount of probability mass is shifted from $P(O|C_1 \cdots C_N)$ to $P(B|C_1 \cdots C_N)$. This probabilistic model thus systematically underestimates the foreground probability, leading to shrinkage. In order to correct for this problem, it would be necessary to sort out all impossible configurations and compare only the true foreground and background probabilities. However, since (13) contains 2^N terms, this is infeasible in practice.

4. Our Method

Instead of applying the law of total probability to all cameras in (13), we can first select some subset of cameras $\{i_1, \ldots, i_k\} \in \mathcal{P}(\{1, \ldots, N\})$ and then apply it to this subset. We will show that by selecting a subset of cameras in a specific way we can achieve some useful properties.

Not all cameras can provide useful information for a given voxel. Each camera can only observe the surface of the object. The volume in front of the surface is assumed to be empty, thus background. The volume immediately behind the surface corresponds to object. However, we can only be sure for a narrow band behind the surface. Behind this narrow band the rest of the volume is *unknown*, as it is not visible to this camera. Voxels near the surface will be classified as invisible (*i.e*. air or inside the object) by many cameras, namely the cameras that do not see the surface from the right angle (*c.f*. Fig. 2(right)). However, these surface voxels will be classified as foreground by the cameras that actually see this part of the surface. These observations are the only ones actually containing information.

Suppose that we selected a subset of cameras such that they all have similar observations, *i.e*. they agree about the labeling of the respective voxel. This means that we have only two events ($O_1 \cdots O_N$ and $B_1 \cdots B_N$) in the total probability formula. We can derive:

$$\begin{aligned} P(B|C_{i_1} \cdots C_{i_k}) = & \underbrace{P(B|B_{i_1} \cdots B_{i_k})}_{=1} P(B_{i_1} \cdots B_{i_k}|C_{i_1} \cdots C_{i_k}) \\ & + \underbrace{P(B|O_{i_1} \cdots O_{i_k})}_{=0} P(O_{i_1} \cdots O_{i_k}|C_{i_1} \cdots C_{i_k}) \quad (14) \\ = & P(B_{i_1} \cdots B_{i_k}|C_{i_1} \cdots C_{i_k}) \quad (15) \end{aligned}$$

and similarly

$$P(O|C_{i_1} \cdots C_{i_k}) = P(O_{i_1} \cdots O_{i_k}|C_{i_1} \cdots C_{i_k}) \quad (16)$$

Variables O_i are not independent in this case, since the visibility depends on several cameras, therefore we cannot factorize them directly. Assuming equal probabilities for $P(O_{i_1} \cdots O_{i_k}) = P(B_{i_1} \cdots B_{i_k})$ and applying Bayes' theorem several times we can get (for the complete derivation and additional results see our extended version[1]):

$$\rho_{bck}(\mathbf{x}) - \rho_{obj}(\mathbf{x}) = -\log \prod_{j=1}^{k} \frac{P(B_{i_1} \cdots B_{i_k}|C_{i_j})}{P(O_{i_1} \cdots O_{i_k}|C_{i_j})} \quad (17)$$

$$= \sum_{j=1}^{k} \rho_{bck}^{i_j}(\mathbf{x}) - \sum_{j=1}^{k} \rho_{obj}^{i_j}(\mathbf{x}) \quad (18)$$

In contrast to other approaches, this formulation does not systematically underestimate either of the probabilities.

We now need to reliably select a subset of cameras. As mentioned earlier, a widely used assumption is that a point is invisible if and only if it is invisible from all cameras. However, in multiple view stereo, if point $\mathbf{x}$ is visible only from one camera then we cannot estimate the depth because it requires its projections on other images (*c.f*. Fig. 2(left)). Our first criterion for selecting cameras is thus that we need *at least two cameras with similar observations.*

Another important insight is that *only those observations are reliable which place the surface near this voxel*. For each voxel, each camera provides an estimate how far this point is away from the surface and if it is in front of or behind the surface. The further the surface observation is from the voxel, the less reliable the observation is, because the camera does not actually see this voxel. This means that the most reliable observations are the ones that have surface observations nearest the current voxel. Moreover, on a small scale there will be no self occlusions and therefore all cameras should have similar observations. These insights lead us to the following selection of cameras (see also Fig. 3). Let N be the set of cameras and $N_d(\mathbf{x})$ the cameras with surface observations at most a distance of d away from $\mathbf{x}$:

$$N_d(\mathbf{x}) = \{\, j \in \{1, \ldots, N\} \mid |t_{j,\mathbf{x},max} - t_{j,\mathbf{x}}| \leq d \,\} \quad (19)$$

$$d_{min}(\mathbf{x}) = \min_d \text{ s. t. } |N_d(\mathbf{x})| \geq k \quad (20)$$

where d_{min} is the minimum distance to select k cameras.

When we estimate labeling probabilities for the voxels that lie on the surface, the proposition that the cameras selected with (19) with some small k have similar observations, is always correct in case there are no outliers. Our

[1] http://www.vision.rwth-aachen.de/projects/cvpr14

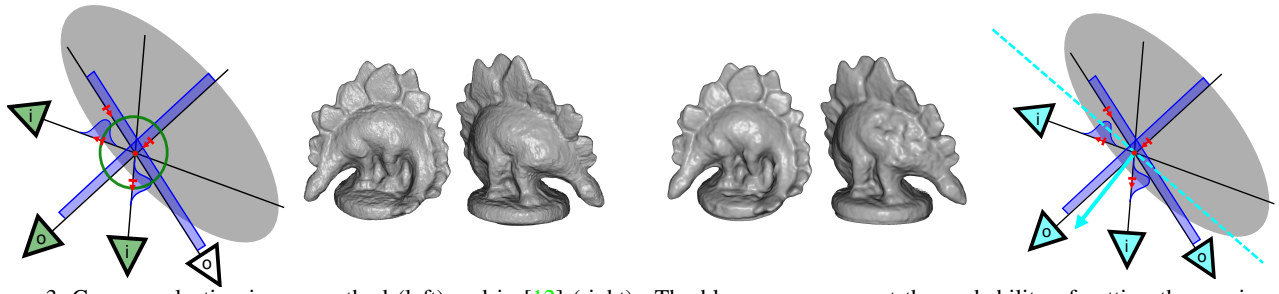


Figure 3. Camera selection in our method (left) and in [12] (right). The blue areas represent the probability of getting the maximum score at a position along a ray. Red arrows represent the surface observations. Uniform distributions indicate outliers (o), where the surface observation is at a random position. Outliers mainly occur in untextured regions and the score will be approximately the same for all patches along the ray. The position of the maximum score is determined by noise, thus randomly. Inlier cameras (i) have a surface observation very near the true surface. (left) Our method selects the cameras with the three nearest surface observations (green circle, green cameras), selecting only one of two outliers in the example. It can be seen that the probability of choosing an outlier becomes smaller for voxels near the true surface.(right) Kolev *et al*. [12] select all front-facing cameras, according to an estimated surface normal (cyan arrow). This means they select all outliers.

criterion is however also robust to outliers since outliers produce randomly distributed predictions that are thus very rarely located close to the surface. Hence we can assume that our criterion is correct in sufficiently many cases in order to provide enough data for the convex optimization.

For voxels $\mathbf{x}$ that are far from the real surface this assumption can be violated depending on the properties of the surface, the camera placements and because of the presence of outliers. However, as stated above, we need to estimate accurate probabilities only for the voxels that are near the actual surface. Therefore we get *accurate labeling near the object surface* even though we *do not know where this surface is located* during the estimation of these probabilities.

The parameter k can be regarded as the minimal number of cameras that observe a point on the surface. This parameter can be selected depending on the number of cameras that are used for reconstruction. We performed all our experiments with $k = 2$ and $k = 3$, with better results achieved by setting $k = 3$. Our formulation is stable even when there are only 2 cameras with similar observation with $k = 3$.

Monte Carlo Experiment. In order to give a better intuition about the assumptions made above, we performed a simple experiment in 2D. Consider an object as in Fig. 4, surrounded by N uniformly spaced cameras (blue circles) at distance R from the object center and pointing towards its center. We examine a part of the object, from the center of the object to the border of the convex hull (red line in Fig. 4), and subdivide it into M uniformly distributed points. For each point on this line, we collect a depth measurement from each camera. In the scope of this experiment this means we sample a distribution, as described below.

We assume that each depth measurement measures the true surface with probability $(1 - \eta)$ or is caused by a noise process (*e.g.*, due to weak texture) with probability η. In the former case, the measurement corresponds to the true surface depth, plus a Gaussian noise component:

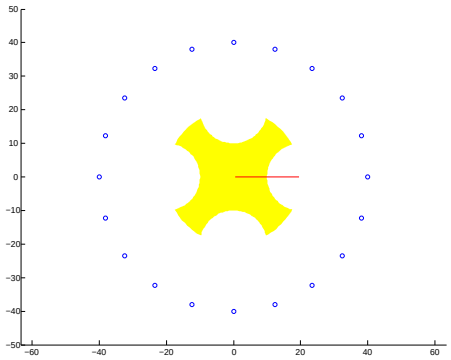


Figure 4. Object used for Monte Carlo experiment: the yellow area represents the 2D object and the blue circles indicate the cameras looking at the object. The points on the red line are examined for the experiment. The frequency of finding at least two inliers out of three cameras is displayed in Fig. 5.

$d_m \sim d_{true} + N(0, \sigma)$. In the latter case, it will be uniformly distributed on the part of the view ray that lies inside the convex hull. We consider all measurements that provide correct labeling of the point (either background or foreground) as inliers. It is important to note that even if a measurement is caused by noise, its outcome may still be an inlier.

Now we sample from the distributions to obtain a depth 'measurement' for each camera. We first sample correct or incorrect cameras with probability $1 - \eta$ or η respectively. Then, depending on the type of the camera, we sample the surface contact point. Finally, we select the three surface contact points that are closest to the point of evaluation. If the majority of those three points are inliers, we consider the camera selection as correct, else as incorrect. We repeat the experiment 1,000 times and calculate the frequency of correct camera selections for each examined point, as shown in Fig. 5.

As can be seen in Fig. 5, in our approach the probability of finding at least two correct observations rises near the

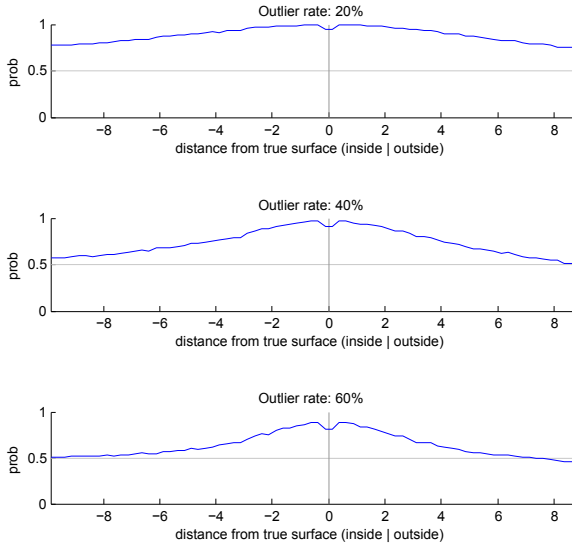


Figure 5. Frequency of finding at least two inliers out of three cameras at different distances from the surface and for different outlier ratios. The probability of finding at least two correct observations rises near the true object surface. The small decrease around the object surface caused by the Gaussian observation noise.

true object surface. The small decrease around the object surface arises from the fact that correct observations of the surface are afflicted by a small noise component regarding the position.

Finally, we need to define labeling costs for a single camera. For $\rho^{ij}_{bck}(\mathbf{x})$ and $\rho^{ij}_{obj}(\mathbf{x})$ we use the costs defined in [12]. They approximately correspond to the following Bernoulli distribution:

$$\rho^{ij}_{bck}(\mathbf{x}) = -\log(\mu^{\mathbb{1}\{t_{j,\mathbf{x},max} > t_{j,\mathbf{x}}\}}(1-\mu)^{(1-\mathbb{1}\{t_{j,\mathbf{x},max} > t_{j,\mathbf{x}}\})}) \quad (21)$$

$$\rho^{ij}_{obj}(\mathbf{x}) = -\log(\mu^{\mathbb{1}\{t_{j,\mathbf{x},max} < t_{j,\mathbf{x}}\}}(1-\mu)^{(1-\mathbb{1}\{t_{j,\mathbf{x},max} < t_{j,\mathbf{x}}\})}) \quad (22)$$

where

$$\mu = 0.25 + \frac{f(C_{i,\mathbf{x},max})}{4}, \text{ and} \quad (23)$$

$$f(s) = 1 - e^{-\frac{\tan(\frac{\pi}{4}(s-1))^2}{\sigma^2}} \quad (24)$$

is a smooth function that maps from $[-1, 1]$ to $[0, 1]$. In all our experiments σ was set to 0.5. In fact, many other smooth functions with similar properties can be used. In general we need to map -1 to 1 and 1 to 0 and the values in between accordingly. For details see [19].

The term 0.25 corresponds to the idea that even high consistency scores may correspond to outliers. The maximum value of $\mu = 0.5$ means that for low consistency scores, we simply cannot make a decision and we have equal probability for foreground and background.

As mentioned above, we determine costs only for the points that lie inside the convex hull. For all other points we set low costs for background and high for object, for example $\rho_{obj}(x) = 1$ and $\rho_{bck}(x) = 0$.

4.1. Comparison with similar approaches

Kolev *et al.* [12, 10] represent labeling costs as an average of costs from a subset of cameras. This subset is chosen using the estimated surface normal, using only the front facing cameras (*c.f.* Fig. 3):

$$N(\mathbf{x}) = \{i \in \{1, \dots, N\} | \angle(V_i, N_\mathbf{x}) \leq \gamma_{max}, i \in Vis(\mathbf{x})\} \quad (25)$$

where $Vis(\mathbf{x})$ denotes the set of visible cameras. Their approach thus requires the estimation of $N_\mathbf{x} = \frac{\nabla d(\mathbf{x})}{\|\nabla d(\mathbf{x})\|}$ where $d(\mathbf{x})$ is a signed distance function to the surface and of the visibility set $Vis(\mathbf{x})$. It is necessary to estimate $N_\mathbf{x}$ also for points, that are not actually on the surface, which leads to a disadvantageous selection of cameras. Moreover they do not include any outlier handling. Kolev *et al.* perform several estimations of the 3D reconstruction and use a surface reconstructed in the previous iteration to estimate $N_\mathbf{x}$ and $Vis(\mathbf{x})$ for the next iteration. Therefore, errors from previous iterations propagate to the following ones. This leads, among others, to problems reconstructing thin structures. It is possible to use this approach without precomputing results for sparse resolutions, but this would make the approach significantly slower. In Section D we will show that this model produces errors in several difficult regions.

5. Discretization and Optimization

We computed the solution of (29) as in [10] using the Primal-Dual method. We used a multi-resolution scheme to increase the speed of computations. However, in contrast to [12, 10], no errors or inaccuracies are propagated to the higher levels, because we only limit the search space, thus the actual estimation is not affected. In all our experiments the multi-resolution scheme affected only the runtime, but not the quality of the results. We do not precompute depth images, but compute the depth for each voxel, thereby avoiding problems with grid size and achieving subvoxel precision.

6. Experimental Results

We evaluated our approach on the well-known Middlebury datasets DINO and TEMPLE[2] [17] and on HEAD[3] [4]. Each of these datasets exhibits different features: DINO contains a smooth poorly textured object, TEMPLE is better textured but contains many small and sharp details. The HEAD dataset violates the Lambertian assumption due to reflections, moreover it contains very thin structures that are challenging to reconstruct.

Middlebury: Figures 1 and 6 show our results for DINO SPARSE. It can be seen that [12, 13, 7, 6] have problems in

[2] http://vision.middlebury.edu/mview/

[3] http://vision.in.tum.de/data/datasets/3dreconstruction

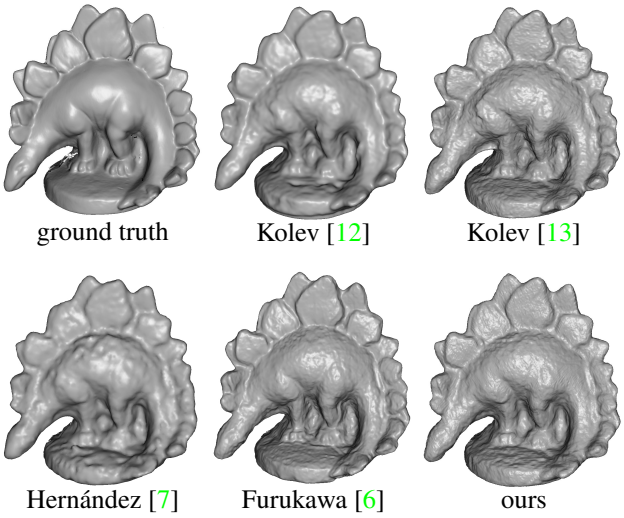


Figure 6. Results for Middlebury DINO SPARSE (*c.f.* Fig. 1).

low textured areas such as the shoulders of the dino, where many outliers occur. Our method reconstructs these parts accurately, because our approach properly estimates probabilities even under a significant portion of outliers. Fig. 7 shows the Middlebury evaluation at the time of publication. The numbers confirm the visual impression. We currently have the top score among published methods for DINO SPARSE and are on rank 4 for DINO RING. Table 2 compares our results for TEMPLE SPARSE and TEMPLE RING to the top scores according to the different measures and to methods similar to ours [12, 13]. Our method reaches ranks between 11 and 16 depending on the measure. The reason for the lower performance is that total variation regularization tends to penalize non-smooth areas.

It can be seen that many approaches which perform well on DINO perform not so well on TEMPLE and vice versa. The only approach with top results on all datasets is [6], who use a complicated pipeline, whereas our approach is simple and essentially parameter free. Our approach got significantly better results than Kolev *et al.*, although the approach in [12] uses the normal orientation to calculate depths and [13] uses a more sophisticated anisotropic optimization. Our results are better than all other approaches that use total variation for regularization.

Head: Fig. 9 shows our results for the HEAD dataset compared to [12] and [4]. [12] produces poor results in many regions and was not able to recover the thin structure. [4, 11] proposed a more complex framework to deal with these problems. Our method achieves slightly better results than [4, 11], although using the simpler framework from [12], except for our novel labeling cost. This clearly shows that the improvement is due to our new labeling method.

Runtime: For DINO we used a maximal resolution of 256^3 and for TEMPLE and HEAD 396^3. On the CPU (i7 3770, 4 cores) computation time was less than 2 hours and for the GPU version at most 20 minutes on a GTX 680.

Sort By	Temple Full 312 views Acc [mm]	Temple Full 312 views Comp [%]	Temple Ring 47 views Acc [mm]	Temple Ring 47 views Comp [%]	Temple Sparse 16 views Acc [mm]	Temple Sparse 16 views Comp [%]	Dino Full 363 views Acc [mm]	Dino Full 363 views Comp [%]	Dino Ring 48 views Acc [mm]	Dino Ring 48 views Comp [%]	Dino Sparse 16 views Acc [mm]	Dino Sparse 16 views Comp [%]
Kostrikov			0.57	99.1	0.79	95.8			0.35	99.6	0.37	99.3
Furukawa 2	0.54	99.3	0.55	99.1	0.62	99.2	0.32	99.9	0.33	99.6	0.42	99.2
Zaharescu			0.55	99.2	0.78	95.8			0.42	98.6	0.45	99.2
Furukawa 3	0.49	99.6	0.47	99.6	0.63	99.3	0.33	99.8	0.28	99.8	0.37	99.2
Tsinghua_BBNC											0.3	99.1
Liu2					0.65	96.9					0.51	98.7
Kolev3			0.7	98.3	0.97	92.7			0.42	99.5	0.48	98.6
Schroers	0.57	99.1	0.64	96.4	2.12	62.9	0.33	99.7	0.33	99.7	0.54	98.6
Hernandez	0.36	99.7	0.52	99.5	0.75	95.3	0.49	99.6	0.45	97.9	0.6	98.5
Hongxing	0.83	95.7	0.79	96.3	0.97	93.9	0.62	96.3	0.5	99.1	0.52	98.4
Liu					0.96	89.6					0.59	98.3
Kolev2			0.72	97.8	1.04	91.8			0.43	99.4	0.53	98.3

Figure 7. Middlebury Multi-view Stereo Evaluation [17]. The best result is marked in red. For DINO SPARSE we have the top result in terms of completeness. In terms of accuracy our method ties with Furukawa3 [6]. [21] are listed with a better accuracy, but this method was not published yet.

	TEMPLE RING		TEMPLE SPARSE	
	Acc	Comp	Acc	Comp
Vu [20]	0.45	99.8		
Bradley	0.57	98.1	0.48	93.7
Furukawa3 [6]	0.47	99.6	0.63	99.3
Kostrikov (ours)	0.57	99.1	0.79	95.8
Kolev3 [13]	0.7	98.3	0.97	92.7
Kolev2 [12]	0.72	97.8	1.04	91.8

Table 2. Middlebury Multi-view Stereo results [17] for the TEMPLE dataset compared to top ranking and similar approaches. At the time of publication our approach was ranked between 11th and 17th place depending on the dataset and measure. The top ranking method for TEMPLE RING by Vu *et al.* [20] produces significantly worse results on the DINO dataset than our approach. Compared to Kolevs methods [12, 13] our method performs significantly better. [8] did not submit to the benchmark.

Limitations and future work: It is worth mentioning that the used optimization framework operates on a discrete grid. This fact enforces some restriction on the accuracy of the reconstructed objects. Our method could be improved by using more elaborate ways to compute depth, higher resolution or using anisotropic optimization from [13].

7. Conclusion

We presented a novel labeling cost for multi-view reconstruction, that is probabilistically derived and becomes more accurate the closer to the true object surface. Our method is robust to outliers, simple to implement, general and it is among the top performing methods in the popular Middlebury benchmark.

Acknowledgements This research has been funded by the DFG project "Situiertes Sehen zur Erfassung von Form und Affordanzen von Objekten" (LE 2708/1-1) under the D-A-CH lead agency programme.

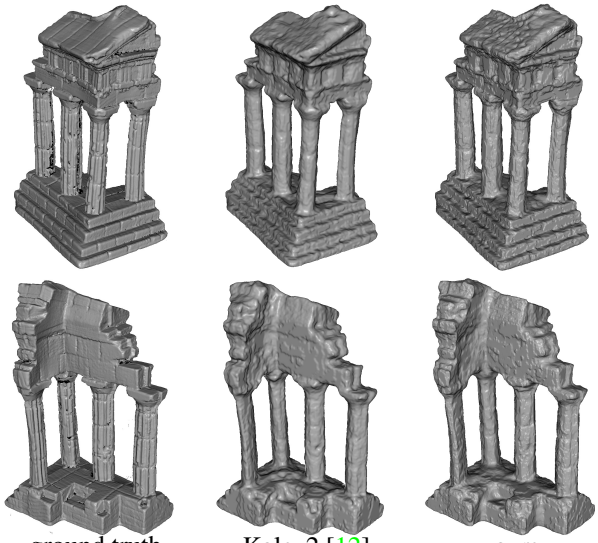


Figure 8. Results for Middlebury TEMPLE RING.

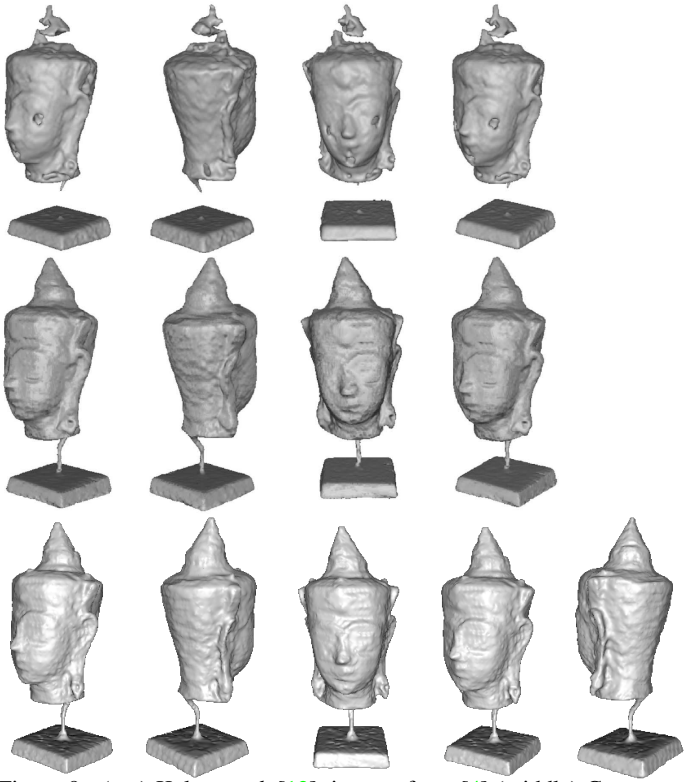


Figure 9. (top) Kolev *et al*. [12], images from [4] (middle) Cremers *et al*., images from [4]. [4] does not present results for DINO, but our reimplementation shows that this approach has problems in the dino dataset, especially in concave regions. (bottom) our proposed approach performs significantly better than [12] and slightly better than [4], although we use the simpler optimization framework from [12]. This clearly shows that the improvements reached by our method are due to our novel labeling term.

References

[1] Y. Boykov and V. Lempitsky. From Photohulls to Photoflux Optimizations. In *BMVC*, 2006. 2

[2] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic Active Contours. In *ICCV*, 1995. 2

[3] A. Chambolle. An Algorithm for Total Variation Minimization and Applications. *Journal of Mathematical Imaging and Vision*, 20:89–97, 2004. 1

[4] D. Cremers and K. Kolev. Multiview Stereo and Silhouette Consistency via Convex Functionals over Convex Domains. *PAMI*, 33:1161–1174, 2011. 6, 7, 8, 10, 11

[5] O. Faugeras and R. Keriven. Variational principles, surface evolution, PDEs, level set methods, and the stereo problem. *Transactions on Image Processing*, 7:336–344, 1998. 2

[6] Y. Furukawa and J. Ponce. Accurate, Dense, and Robust Multiview Stereopsis. *PAMI*, 32:1362–1376, 2010. 1, 6, 7

[7] C. Hernández and F. Schmitt. Silhouette and stereo fusion for 3D object modeling. *CVIU*, 96:367–392, 2004. 1, 2, 3, 6, 7

[8] C. Hernández, G. Vogiatzis, and R. Cipolla. Probabilistic visibility for multi-view stereo. In *CVPR*, 2007. 1, 2, 3, 4, 7

[9] A. Hornung, B. Zeng, and L. Kobbelt. Image Selection For Improved Multi-View Stereo. In *CVPR*, 2008. 2

[10] K. Kolev. *Convexity in Image-Based 3D Surface Reconstruction*. PhD thesis, Department of Computer Science, Technical University of Munich, Germany, January 2012. 6

[11] K. Kolev and D. Cremers. Integration of Multiview Stereo and Silhouettes via Convex Functionals on Convex Domains. In *ECCV*, 2008. 7

[12] K. Kolev, M. Klodt, T. Brox, and D. Cremers. Continuous Global Optimization in Multiview 3D Reconstruction. *IJCV*, 84:80–96, 2009. 1, 2, 3, 5, 6, 7, 8, 9

[13] K. Kolev, T. Pock, and D. Cremers. Anisotropic Minimal Surfaces Integrating Photoconsistency and Normal Information for Multiview Stereo. In *ECCV*, 2010. 1, 2, 6, 7

[14] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *ICCV*, 2001. 1

[15] V. Lempitsky, Y. Boykov, , and D. Ivanov. Oriented Visibility for Multiview Reconstruction. In *ECCV*, 2006. 2

[16] C. Schroers, H. Zimmer, L. Valgaerts, A. Bruhn, O. Demetz, and J. Weickert. Anisotropic range image integration. In *DAGM*, 2012. 2

[17] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. In *CVPR*, 2006. 1, 6, 7

[18] S. Seitz and C. Dyer. Photorealistic Scene Reconstruction by Voxel Coloring. In *CVPR*, 1997. 2

[19] G. Vogiatzis, P. H. S. Torr, and R. Cipolla. Multi-view stereo via volumetric graph-cuts. In *CVPR*, 2005. 1, 2, 3, 6

[20] H. H. Vu, R. Keriven, P. Labatut, and J.-P. Pons. Multi-view stereo via volumetric graph-cuts. In *CVPR*, 2009. 7

[21] Z. Xue, X. Cao, Y. Liu, Q. Dai, and N. Zhang. Continuous depth-map estimation with explicit occlusions for multi-view stereo. In *submitted to CVIU*, 2011. 7

Appendix

This appendix contains four sections. In Section A we give details about the camera selection process. Section B contains detailed derivations of the formulas in the paper and Section C gives more details about the discretization and optimization used in the framework. In Section D we show results for more datasets.

A. Effect of the Number of Cameras

Fig. 10 shows our results for DINO SPARSE using $k = 1, 2, 3, 4$. The results given for 2 cameras look slightly better in some concave areas, but the reconstruction for $k = 3$ cameras is less noisy. The reconstruction using $k = 4$ cameras has problems in concave areas, where there are fewer cameras available that are able to see these points. All our results submitted to the Middlebury evaluation were created using $k = 3$.

B. Derivations

Complete derivation of the formula in Section 4:

$$
\begin{aligned}
&\rho_{bck}(\mathbf{x}) - \rho_{obj}(\mathbf{x}) \\
&\quad = -\log \frac{P(B|C_{i_1} \cdots C_{i_k})}{P(O|C_{i_1} \cdots C_{i_k})} \\
&\quad = -\log \frac{P(B_{i_1} \cdots B_{i_k}|C_{i_1} \cdots C_{i_k})}{P(O_{i_1} \cdots O_{i_k}|C_{i_1} \cdots C_{i_k})} \\
&\quad = -\log \frac{P(C_{i_1} \cdots C_{i_k}|B_{i_1} \cdots B_{i_k})}{P(C_{i_1} \cdots C_{i_k}|O_{i_1} \cdots O_{i_k})} \\
&\quad = -\log \prod_{j=1}^{k} \frac{P(C_{i_j}|B_{i_1} \cdots B_{i_k})}{P(C_{i_j}|O_{i_1} \cdots O_{i_k})} \\
&\quad = -\log \prod_{j=1}^{k} \frac{P(B_{i_1} \cdots B_{i_k}|C_{i_j})}{P(O_{i_1} \cdots O_{i_k}|C_{i_j})} \\
&\quad = \sum_{j=1}^{k} \rho(B_{i_1} \cdots B_{i_k}|C_{i_j}) - \sum_{j=1}^{k} \rho(O_{i_1} \cdots O_{i_k}|C_{i_j}) \\
&\quad = \sum_{j=1}^{k} \rho_{bck}^{i_j}(\mathbf{x}) - \sum_{j=1}^{k} \rho_{obj}^{i_j}(\mathbf{x})
\end{aligned}
$$

C. Discretization and Optimization

In our experiments we used the following discretization. We used a uniform 3-dimensional grid $N_x \times N_y \times N_z$ with the maximal resolution N_{max} and define the grid step as

$$
h = \max\left(\frac{|\mathbf{x}_{max} - \mathbf{x}_{min}|}{N_{max}}, \frac{|y_{max} - y_{min}|}{N_{max}}, \frac{|z_{max} - z_{min}|}{N_{max}}\right) \quad (26)
$$

and resolutions for all dimensions:

$$
N_x = \frac{|x_{max} - x_{min}|}{h}, N_y = \frac{|y_{max} - y_{min}|}{h}, N_z = \frac{|z_{max} - z_{min}|}{h}. \quad (27)
$$

Function values at the points of the grid are defined as

$$
g_{i,j,k} = g(x_{min} + hi, y_{min} + hj, z_{min} + hk). \quad (28)
$$

The minimum of

$$
E(u) = \underbrace{\int_V \rho(\mathbf{x})|\nabla u(\mathbf{x})|dx}_{\text{smoothing term}} + \lambda \underbrace{\int_V (\rho_{obj}(\mathbf{x}) - \rho_{bck}(\mathbf{x}))u(\mathbf{x})dx}_{\text{labeling cost}} \quad (29)
$$

can then be found using the Primal-Dual method. We start with some initialization $u^{(0)}$ and then for $t = 0, 1, \ldots$ we iterate

$$
\xi_{i,j,k}^{(t+1)} = \Pi_K(\xi_{i,j,k}^{(t)} + \eta \nabla \bar{u}_{i,j,k}^{(t)}) \quad (30)
$$

$$
u_{i,j,k}^{(t+1)} = \Pi_{[0,1]}(u_{i,j,k}^{(t)} + \theta(div(\xi_{i,j,k}^{(t+1)}) - b_{i,j,k})) \quad (31)
$$

$$
\bar{u}_{i,j,k}^{(t+1)} = 2u_{i,j,k}^{(t+1)} - u_{i,j,k}^{(t)} \quad (32)
$$

until convergence. $b(x) = \lambda(\rho_{obj}(x) - \rho_{bck}(x))$, Π_X denotes the projection into the set X and $K = \{\xi \in \mathbb{R}^3 | \|\xi\| \leq \rho_{i,j,k}\}$. In all our experiments $\eta = \theta = 0.1$. We used

$$
\left| \frac{E(u^{(t+1)}) - E(u^{(t)})}{E(u^{(t)})} \right| < \epsilon \quad (33)
$$

as a convergence criteria with $\epsilon = 10^{-9}$.

In order for the algorithm to converge, $\nabla u_{i,j,k}$ has to be defined as forward difference:

$$
\nabla u_{i,j,k} = \begin{pmatrix} u_{i+1,j,k} - u_{i,j,k} \\ u_{i,j+1,k} - u_{i,j,k} \\ u_{i,j,k+1} - u_{i,j,k} \end{pmatrix}, \quad (34)
$$

and divergence as backward difference:

$$
\begin{aligned}
&div(\xi_{i,j,k}) = \\
&(\xi_{i,j,k}^{(0)} - \xi_{i-1,j,k}^{(0)}) + (\xi_{i,j,k}^{(1)} - \xi_{i,j-1,k}^{(1)}) + (\xi_{i,j,k}^{(2)} - \xi_{i,j,k-1}^{(2)}). \quad (35)
\end{aligned}
$$

Also when we search for the depth we discretized the camera ray in the following way:

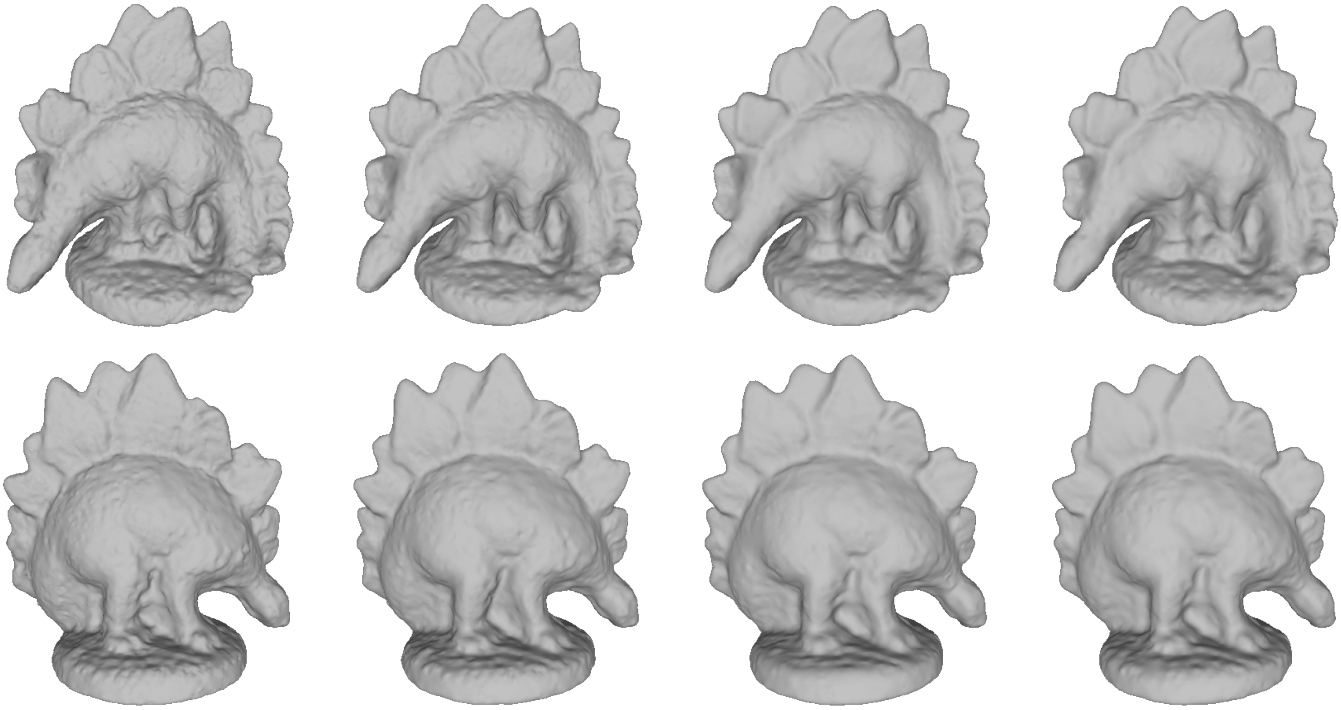
$$
t_{j\mathbf{x},max} = arg \max_{t \in \{t_{j,\mathbf{x}} + ih - \frac{h}{2} | i \in \mathbb{N}\}} S_j(r_{j,\mathbf{x}}(t)). \quad (36)
$$

This choice of the ray discretization is due to the fact that we search for the surface not at the grid points but between them.

Furthermore, since we approximate $\nabla u(x)$ with forward differences it is important to make the discontinuity cost consistent with discretization:

$$
VOTE_j(x) = \begin{cases} S_j(x) & \text{if } r_{j,\mathbf{x}}(t_{j,\mathbf{x},max}) \in \\ & [x, x+h] \times [y, y+h] \times [z, z+h] \\ 0 & \text{otherwise.} \end{cases} \quad (37)
$$

We use a multi-resolution scheme to increase the speed of computations. However, in contrast to [12], no errors



(a) $k = 1$, *i.e.* ≥ 1 camera (b) $k = 2$, *i.e.* ≥ 2 cameras (c) $k = 3$, *i.e.* ≥ 3 cameras (d) $k = 4$, *i.e.* ≥ 4 cameras

Figure 10. Results for DINO SPARSE for different number of cameras.

or inaccuracies are propagated to the higher levels, because we only limit the search space, thus the actual estimation is not affected. In all our experiments the multi-resolution scheme affected only the runtime, but not the quality of the results. We do not precompute depth images, but compute the depth for each voxel, thereby avoiding problems with grid size and achieving sub-voxel precision.

D. Additional Results

In the paper we discussed results for the HEAD dataset [4] because it provides some interesting special cases for multi-view reconstruction. We also evaluated our approach on other datasets from that repository and include the results in Fig. 11.

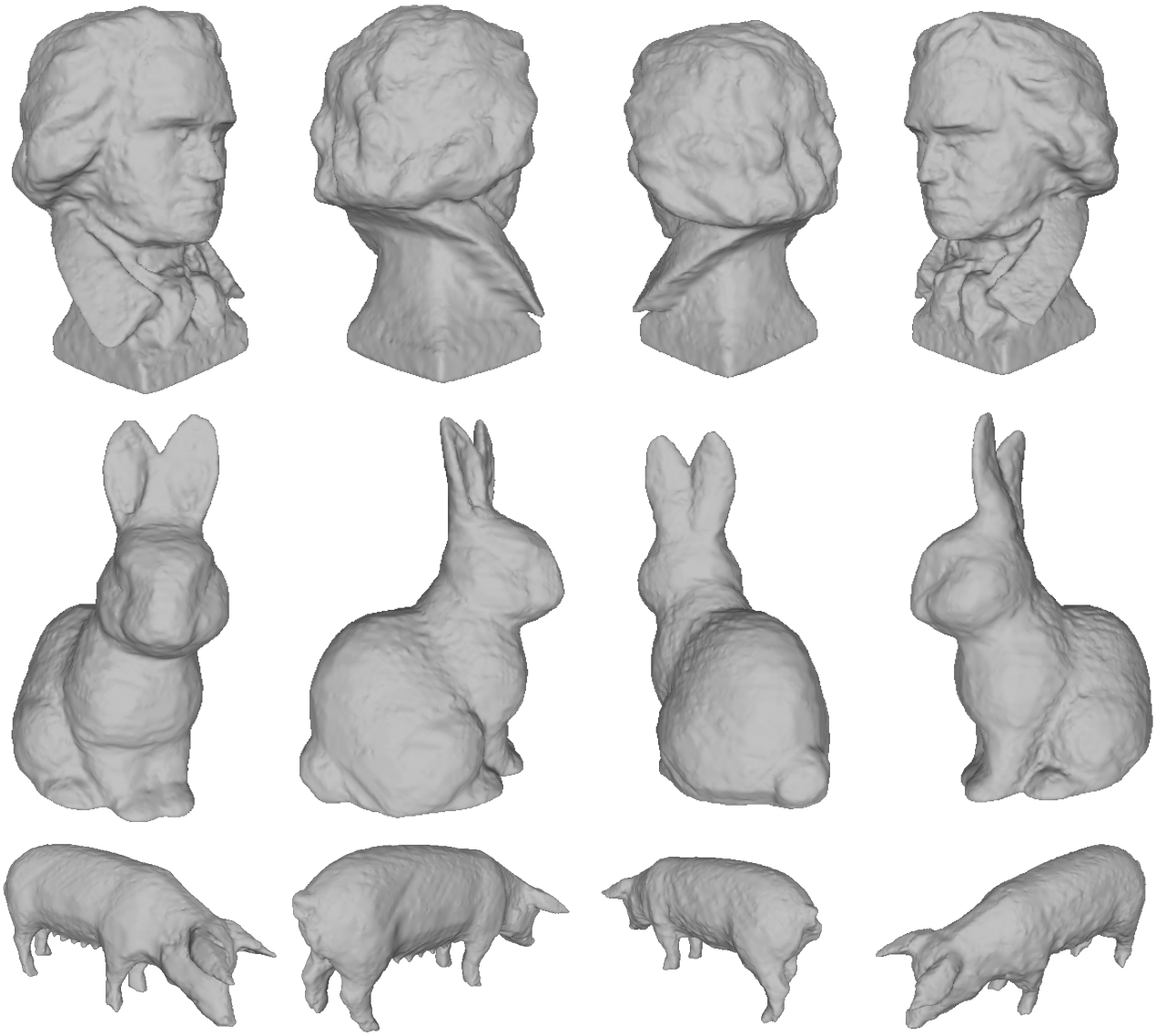


Figure 11. Results for the datasets BEETHOVEN, BUNNY, PIG [4].