On Using Coplanar Shadowgrams for Visual Hull Reconstruction

Shuntaro Yamazaki
shuntaro@ni.aist.go.jp

Srinivasa Narasimhan
srinivas@cs.cmu.edu

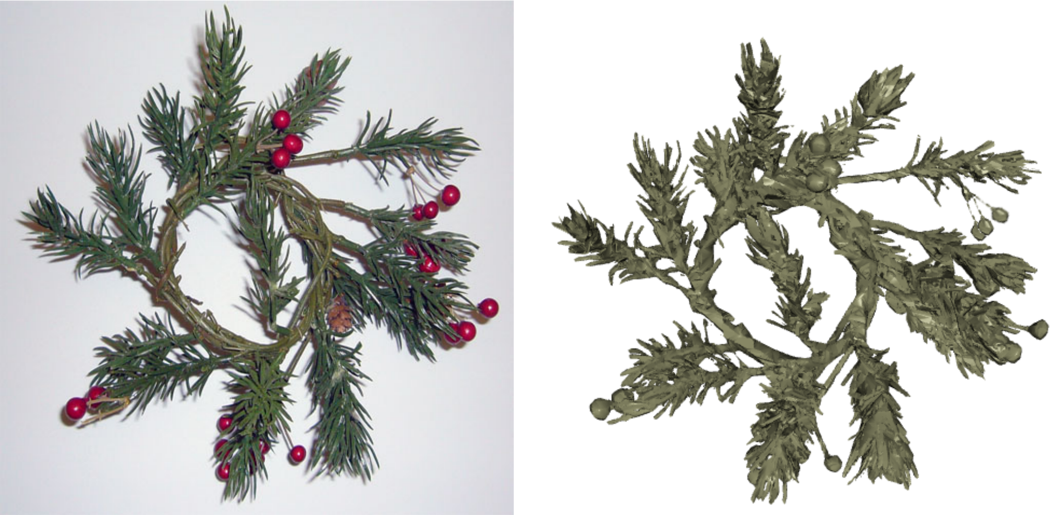
Simon Baker
sbaker@microsoft.com

Takeo Kanade
tk@cs.cmu.edu

CMU-RI-TR-07-29

Abstract

Acquiring 3D models of intricate objects (like tree branches, bicycles and insects) is a hard problem due to severe self-occlusions, repeated thin structures and surface discontinuities. In theory, a shape-from-silhouettes (SFS) approach can overcome these difficulties and use many views to reconstruct visual hulls that are close to the actual shapes. In practice, however, SFS is highly sensitive to errors in silhouette contours and the calibration of the imaging system, and therefore not suitable for obtaining reliable shapes with a large number of views. We present a practical approach to SFS using a novel technique called coplanar shadowgram imaging, that allows us to use dozens to even hundreds of views for visual hull reconstruction. Here, a point light source is moved around an object and the shadows (silhouettes) cast onto a single background plane are observed. We characterize this imaging system in terms of image projection, reconstruction ambiguity, epipolar geometry, and shape and source recovery. The coplanarity of the shadowgrams yields novel geometric properties that are not possible in traditional multi-view camera-based imaging systems. These properties allow us to derive a robust and automatic algorithm to recover the visual hull of an object and the 3D positions of light source simultaneously, regardless of the complexity of the object. We demonstrate the acquisition of several intricate shapes with severe occlusions and thin structures, using 50 to 120 views.



(a) An intricate object (b) Our reconstruction

Figure 1: Obtaining 3D models of intricate shapes such as in (a) is hard due to severe occlusions and correspondence ambiguities. (b) By moving a point source in front of the object, we capture a large number of shadows cast on a single fixed planar screen (122 views for this object). Applying our techniques to such *coplanar shadowgrams* results in accurate recovery of intricate shapes.

1 Introduction

Acquiring 3D shapes of objects that have numerous occlusions, discontinuities and repeated thin structures is challenging for vision algorithms. For instance, the wreath object shown in Figure 1(a) contains over 300 branch-lets each 1-3mm in diameter and 20-25mm in length. Covering the entire surface area of such objects requires a large number (dozens or even a hundred) of views. Thus, finding correspondences between views as parts of the object get occluded and "dis-occluded" becomes virtually impossible, often resulting in erroneous and incomplete 3D models.

If we only use the silhouettes of an object obtained from different views, it is possible to avoid the issues of correspondence and occlusion in the object, and reconstruct its *visual hull* [1]. The top row of Figure 2 illustrates the visual hulls estimated using our technique from different numbers of silhouettes. While the visual hull computed using a few (5 or 10) silhouettes is too coarse, the visual hull estimated from a large number of views (50) is an excellent model of the original shape.

In practice, however, SFS algorithms are highly sensitive to errors in the geometric

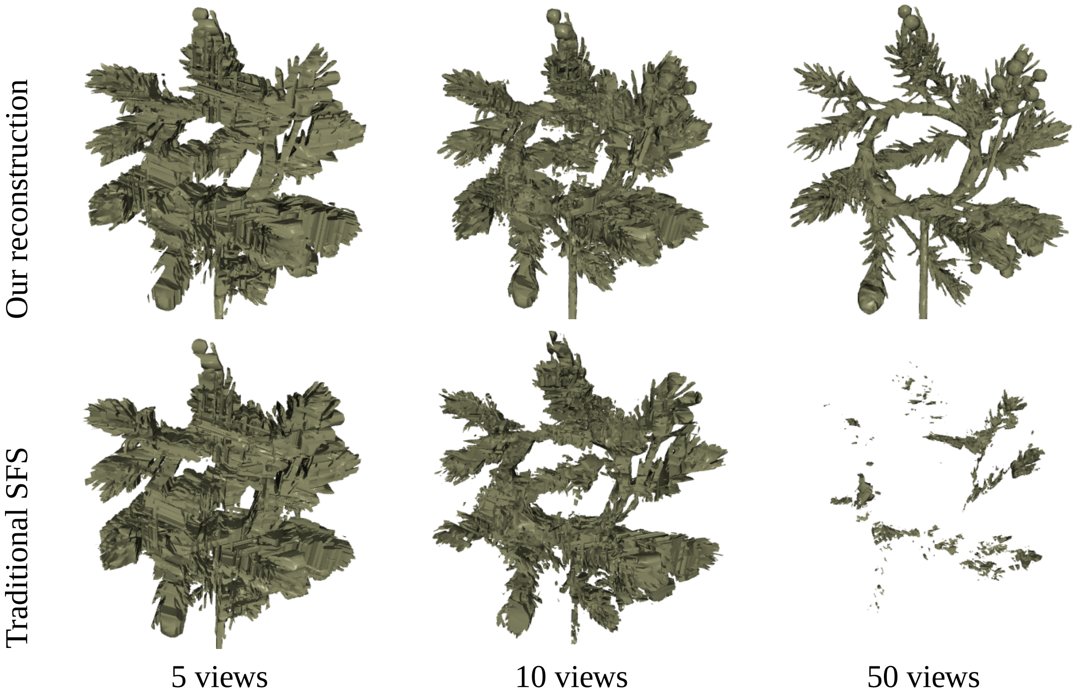
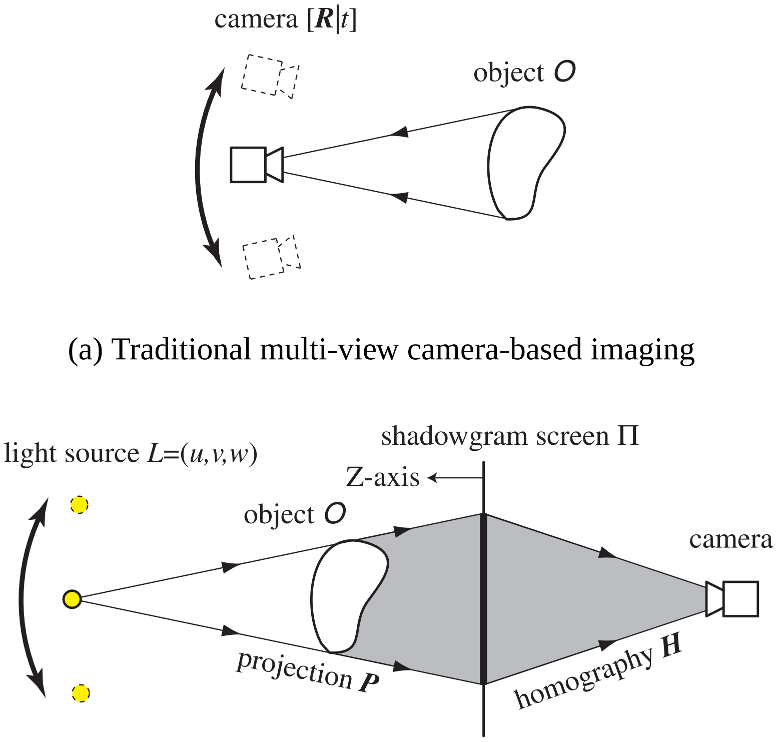


Figure 2: Sensitivity of SFS reconstruction. (Top) The visual hulls reconstructed using the light source positions estimated by our method. As the number of silhouettes increases, the visual hull gets closer to the actual shape. (Bottom) The reconstructions obtained from slightly erroneous source positions. As the number of views increases, the error worsens significantly.

parameters of the imaging system (camera calibration) [19]. This sensitivity worsens as the number of views increases, resulting in poor quality models. The bottom row in Figure 2 shows the visual hulls of the wreath object obtained using a naïve SFS algorithm. This drawback must be addressed in order to acquire intricate shapes reliably.

In traditional SFS, a camera observes the object, and the silhouette is extracted from obtained images by matting [20]. Multiple viewpoints are captured by moving either the camera or the object (see Figure 3(a)). For each view, the relative pose between the object and the camera is described by six parameters (3D translation and 3D rotation). Savarese *et al.* [15] proposed a system that avoids silhouette matting. When an object is illuminated by a single point light source, the shadow cast onto a background plane (also known as a shadowgram [18]) is sharp and can be directly used as its silhouette. Silhouettes from multiple views are obtained by rotating the object. In terms of multi-view geometry, this


camera [$\boldsymbol{R}|t$]
object O
light source $L=(u,v,w)$
shadowgram screen Π
Z-axis
object O
camera
projection $\boldsymbol{P}$
homography $\boldsymbol{H}$

(a) Traditional multi-view camera-based imaging

(b) Coplanar shadowgram imaging

Figure 3: (a) The object of interest is observed directly by a projective camera. The silhouette of the object is extracted from the captured image. Multiple views are obtained by moving the camera or the object. (b) A point source illuminates the object and its shadow cast on a planar rear-projection screen represents the silhouette of the object. Coplanar shadowgrams from multiple viewpoints are obtained by translating the light source. Note that the relative transformation between the object and the screen remains fixed across different views. This is the key difference between the systems in (a) and (b).

is equivalent to traditional SFS, requiring six parameters per view.

In this paper, we present a novel approach to SFS called *coplanar shadowgram imaging*. We use a setup similar in spirit to that proposed by Savarese *et al.* [15] The key difference here is that the point source is moved, while the object, the camera and the background screen all remain stationary. The central focus of this work is the acquisition of visual hulls for intricate and opaque objects from a large number of coplanar shadowgrams. Our main contributions are described below.

Multi-view geometry of coplanar shadowgram imaging:

Figure 3 shows the difference between the traditional camera-based and coplanar shadowgram imaging systems. Observe that the relative transformation between the object and screen remains fixed across different views. The image projection model is described by only three parameters per view (3D translation of the source) instead of six in the traditional system. Our geometry is similar in spirit to the parallax geometry [16, 4] where the homography between image planes is known to be an identity, which allows us to derive novel geometric properties that are not possible in the traditional multi-view camera-based imaging system. For instance, we show that epipolar geometry can be uniquely estimated from only the shadowgrams, without requiring any correspondences, and independent of the object's shape.

Recovery of light source positions:

When the shape of the object is unknown, the locations of all the point sources can be recovered from coplanar shadowgrams, only up to a four parameter linear transformation. We show how this transformation relates to the well-known *Generalized Perspective Bas-Relief* (GPBR) ambiguity [11] that is derived for a single viewpoint system. We break this ambiguity by simultaneously capturing the shadowgrams of two spheres.

Robust reconstruction of visual hull:

Even a small amount of blurring in the shadow contours may result in erroneous estimates of source positions that in turn can lead to erroneous visual hulls. We propose a two-step optimization of the light source positions that can robustly reconstruct the visual hulls of intricate shapes. First, the error in light source positions is corrected by enforcing the reconstructed epipolar geometry. This step achieves significant improvement over the initial shape. Second, we minimize the mismatch between the *convex polygons* of the acquired shadowgrams and those obtained by reprojecting the estimated visual hull. The source positions obtained serve as a good initial guess to the final step that minimizes the mismatch between the actual shadowgrams. In practice, the convex polygon step also leads to faster convergence.

For the analogous camera-based imaging, a number of algorithms have been proposed to make SFS robust to errors in camera position and orientation. These techniques optimize camera parameters by exploiting either epipolar tangency [19, 3, 21] or silhouette consistency [23, 10], or assume orthographic projection [8]. However, they all require non-trivial parameter initializations and the knowledge of silhouette feature correspondences (known as frontier points [9]). This restricts the types of objects that one can reconstruct using these methods; silhouettes of simple objects such as spheres do not have enough features and intricate objects like branches have too many, making it hard to find correspondences automatically. As a result, previous approaches have succeeded in only acquiring the 3D shape of *reasonably complex* shapes like people and statues that can be modeled using a small number of views.

In contrast, our algorithm is effective for a large number of views (dozens to a hundred), does not require any feature correspondences and does not place any restriction on the shapes of the objects. The minimization of silhouette mismatch is also easier requiring optimization of source translation (3 DOF per view), instead of the harder (and sometimes ambiguous [9]) joint estimation of camera rotation and translation (6 DOF per view) in the traditional system. As a result, we achieve good quality reconstructions of real objects such as wreaths, wiry balls and palm trees, that show numerous occlusions, discontinuities and thin structures. In addition, we have also evaluated our techniques quantitatively using simulations with objects such as corals, branches, bicycles whose ground truth shapes are known beforehand.

Despite significant progress in optical scanning hardware [5, 12] and multi-view geometry [9, 17], reconstruction of intricate shapes remains an open problem. We believe this work is an initial step in the right direction. In the future, we will extend our techniques to include multiple screens covering $360^\circ \times 360^\circ$ views of the objects, and combine our techniques with stereo and photometric stereo, to obtain reconstructions that are smoother than visual hulls, including concavities.

2 Coplanar Shadowgrams

We define shadowgrams as the shadows cast on a background plane by an object that occludes a point source. If the object is opaque, the shadowgram accurately represents the silhouette of the object. Henceforth, we shall use shadowgrams and silhouettes interchangeably. Coplanar shadowgram imaging is the process of acquiring several shadowgrams on *a single plane* by moving the light source. Our setup shown in Figure 4 includes a 6M pixel Canon EOS-20D digital camera, a 250 watt 4mm incandescent bulb, and a 4ft × 4ft translucent rear-projection screen.

Figure 3(b) illustrates the viewing and illumination geometry of coplanar shadowgram imaging. Without loss of generality, let the shadowgram plane Π be located at $Z = 0$ in the world coordinate system **W**. The $Z-$direction is defined so that it is aligned with the optical axis of the camera. Suppose a point light source is at $L = (u, v, w)^T$ and **L** be a

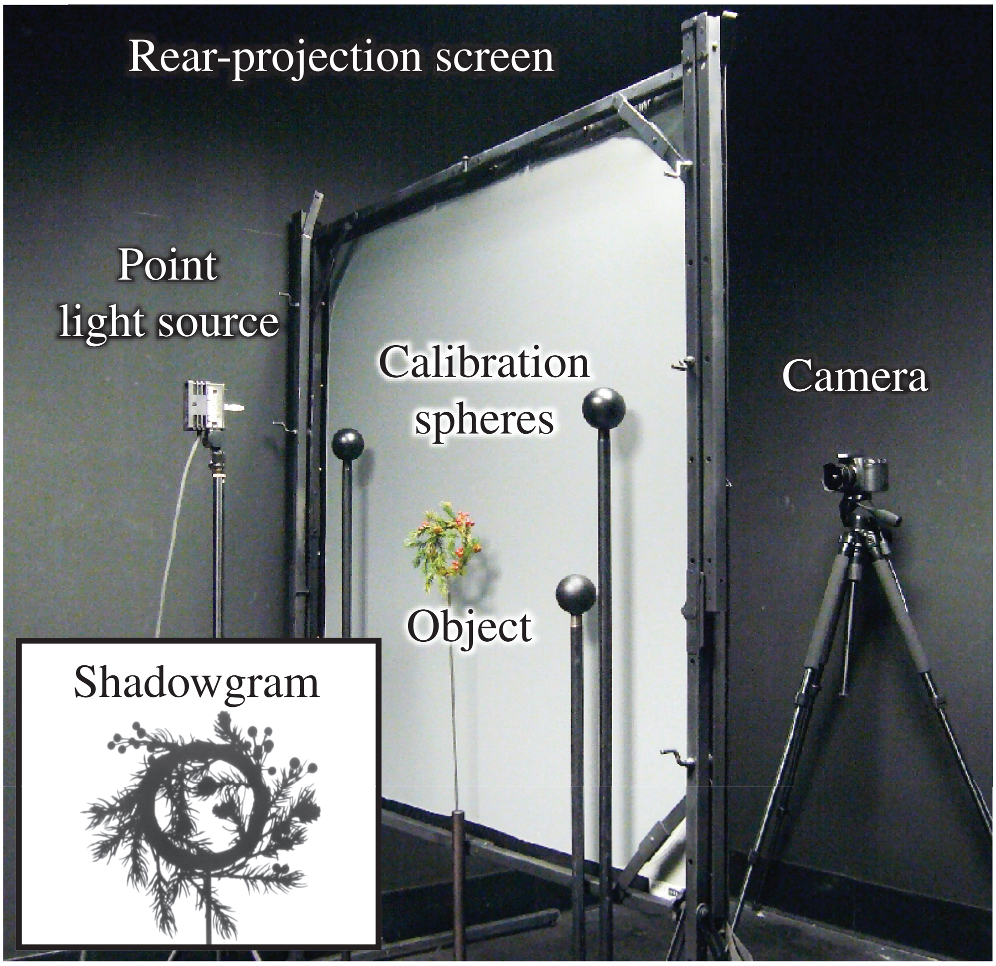


Figure 4: The setup used to capture coplanar shadowgrams includes a digital camera, a single point light source, and a rear-projection screen. The object is placed close to the screen to cover a large field of view. Two or more spheres are used to estimate the initial light source positions. (Inset) An example shadowgram obtained using the setup.

translated coordinate system whose origin is at L. Then, the resulting shadowgram S is obtained by applying a source dependent projective transformation $\boldsymbol{P}(L)$ to the object O as:

$$S = \boldsymbol{P}(L)\,O \tag{1}$$

where the projective transformation $\boldsymbol{P}(L)$ from 3D space to the 2D screen is:

$$\boldsymbol{P}(L) = \underbrace{\begin{pmatrix} 1 & 0 & 0 & u \\ 0 & 1 & 0 & v \\ 0 & 0 & 0 & 1 \end{pmatrix}}_{\mathbf{L} \text{ to } \mathbf{W} \text{ on } \Pi} \underbrace{\begin{pmatrix} -w\mathbf{I_3} & \mathbf{0}_3 \\ (0,0,1) & 0 \end{pmatrix}}_{\text{projection to } \Pi} \underbrace{\begin{pmatrix} \mathbf{I_3} & -L \\ \mathbf{0}_3^T & 1 \end{pmatrix}}_{\mathbf{W} \text{ to } \mathbf{L}} \tag{2}$$

$$= \begin{pmatrix} -w & 0 & u & 0 \\ 0 & -w & v & 0 \\ 0 & 0 & 1 & -w \end{pmatrix}. \tag{3}$$

$\mathbf{I}_3$ is a 3×3 identity matrix and $\mathbf{0}_3 = (0,0,0)^T$.

In Equation (1), S represents the set of 2D points (in homogeneous coordinates) within the shadowgram on the plane Π, and O represents the 3D points on the object surface. The image I captured by the camera is related to the shadowgram S on the plane Π by a 2D homography: $I = \boldsymbol{H}S$. This homography $\boldsymbol{H}$ is independent of the light source position and can be estimated separately using any computer vision algorithm (such as the four-point method [9]). In the following, we assume that the shadowgram S has been estimated using $S = \boldsymbol{H}^{-1} I$.

Now let a set of shadowgrams $\{S_k\}$ be captured by moving the source to n different locations $\{L_k\}$ $(k = 1, \cdots, n)$. Then, the visual hull V of the object is obtained by the intersection:

$$V = \bigcap_k \boldsymbol{P}(L_k)^{-1} S_k \tag{4}$$

Thus, given the 3D locations L_k of the light sources, the visual hull of the object can be estimated using Equation (3) and Equation (4). Table 1 summarizes and contrasts the geometric parameters that appear in the traditional multi-view camera-based and coplanar shadowgram imaging systems.

3 Source Recovery using two Spheres

When the shape of the object is unknown, it is not possible to uniquely recover the 3D source positions using only the coplanar shadowgrams. In the technical report [22], we discuss the nature of this ambiguity and show that the visual hull and the source positions can be computed up to a 4 parameter linear transformation. This transformation is similar in spirit to the 4 parameter *Generalized Perspective Bas-Relief* (GPBR) transformation [11] with one difference: in the context of coplanar shadowgrams, the GPBR

Table 1: Comparison between the geometric parameters of silhouette projection. For n views, the traditional multi-view system is described by $5 + 6n$ parameters. In comparison, the coplanar imaging system requires only $8 + 3n$ parameters.

	View independent	View dependent
Projective cameras	1 (focal length) 1 (aspect ratio) 1 (skew) 2 (image center)	3 (rotation) 3 (translation)
Coplanar shadowgrams	8 (homography $\boldsymbol{H}$)	3 (translation L)

transformation is separately defined with respect to the local coordinate frame defined at each source location, whereas our transformation is defined with respect to a global coordinate frame defined on the screen. We also derive a relationship between the two transformations.

3.1 Geometric Solution to 3D Source Recovery

We now present a simple calibration technique to break this ambiguity. The 3D location $L = (u, v, w)^T$ of a light source is directly estimated by capturing shadowgrams of two additional spheres that are placed adjacent to the object of interest.

Two (out of three) coordinates $L' = (u, v)^T$ of the light source can be estimated by analyzing the shadowgrams of two spheres. Figure 5 illustrates the coplanar elliptical shadowgrams cast by the two spheres. [1] The ellipses are localized using a constrained least squares approach [6]. The intersection of the major axes $\overline{A_1 B_1}$ and $\overline{A_2 B_2}$ of the two ellipses directly yields $L' = (u, v)^T$.

The third coordinate w is obtained as the intersection of hyperbolae in 3D space as shown below. Without loss of generality, consider the 3D coordinate system whose origin is at the center of the ellipse, and X and Y axes are respectively the major and minor axes of the ellipse. Then, the ellipse is represented in the following form.

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \quad (a > b) \tag{5}$$

In 3D space, there exists an inscribed sphere tangent to the conical surface and the plane, regardless of the position or the radius of $\mathcal{S}$. The cross section of the inscribed

[1] Each sphere is placed so that the minimum distance between a light source and the rear-projection screen is larger than the distance between the center of the sphere and the screen. Under this configuration, the cast shadow of the sphere is always an ellipse [2].

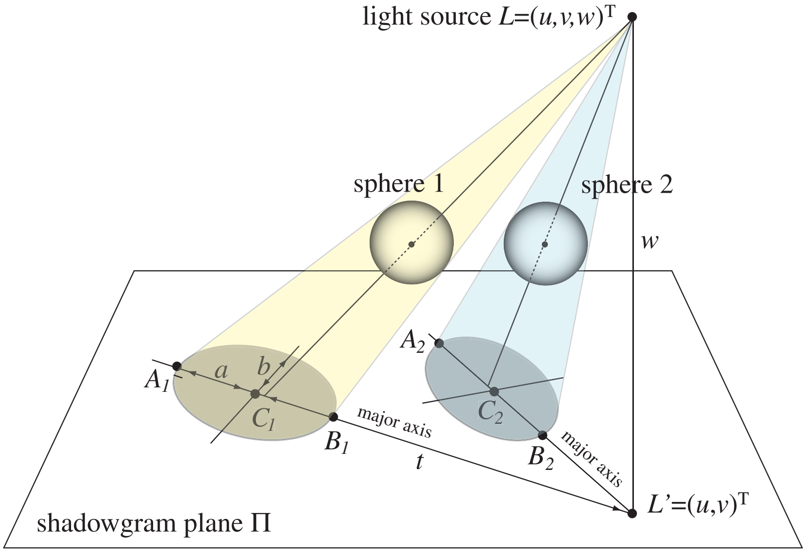


Figure 5: Source position $L = (u, v, w)^T$ is recovered using the elliptical shadowgrams of two spheres. The radii and positions of the spheres are unknown. The major axes of the ellipses intersect the screen at $L' = (u, v)^T$. The w component is obtained using Equation (8).

sphere by the plane that includes the apex of the cone and the major axis of the ellipse is shown in Figure 6. The center of the inscribed sphere is shown by R. The other symbols are corresponding to those in Figure 5. The center of the ellipse C is the origin of the coordinate system.

The inscribed sphere is tangent to XY-plane at a focus of the ellipse R', hence

$$\overline{CR'} = \sqrt{a^2 - b^2}. \tag{6}$$

Using the symmetry of triangles,

$$\overline{LA} - \overline{AR'} = \overline{LB} - \overline{BR'}. \tag{7}$$

Let the position of the apex be $L = (t, 0, w)$ in this coordinate system, then we can solve w with respect to t as:

$$w = \sqrt{\frac{b^2t^2}{a^2 - b^2} - b^2}, \tag{8}$$

where a and b are the semimajor and semiminor axes of one of the ellipses, and t is the length between L' and the center of the ellipse.

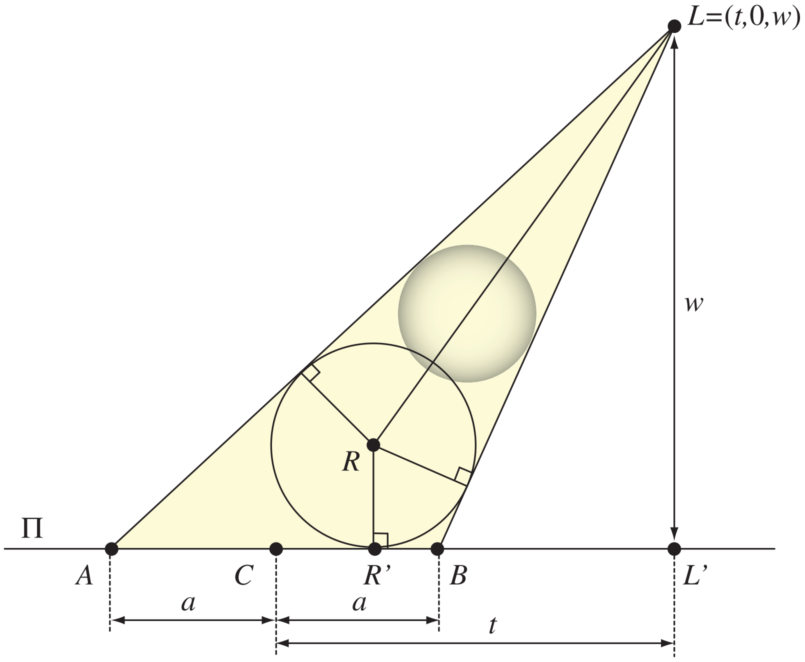


Figure 6: The cross section of a right circular conical surface formed by the light rays emanating from a point light source L and tangent to a calibration sphere.

Note that more than two spheres may be used for a robust estimate of the source position. The above method is completely automatic and does not require the knowledge of the radii of the spheres, the exact locations at which they are placed in the scene, or point correspondences.

3.2 Sensitivity to Silhouette Blurring

This technique for estimating the source position can be sensitive to errors in measured silhouettes. Due to the finite size of the light bulb, the shadowgram formed may be blurred, making it hard to localize the boundary of the silhouette. The extent of blurring depends on the relative distances of the screen and source from the object. To show the sensitivity of the technique, we performed simulations with spheres. We blurred the simulated silhouettes (effective resolution 480×360 pixels) with 5×5 and 10×10 averaging kernels, and estimated the 3D coordinates of the light source. Figure 7 presents u and w components of the source positions reconstructed using three spheres. Observe that the estimation becomes poor when the shadowgram is close to a right circle. In turn, the visual hull of a tree

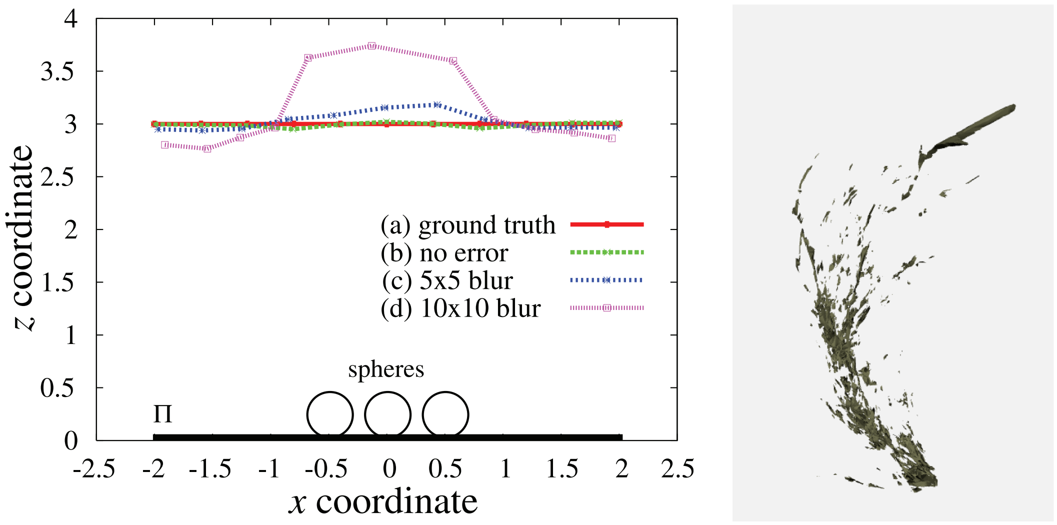


Figure 7: Source positions (u, w) are estimated using three calibration spheres. The sizes and positions of the spheres and screen are shown in the plot. Each plot shows 11 source positions obtained from (a) ground truth, (b) accurate shadowgrams, and (c)-(d) shadowgrams blurred using 5×5 and 10×10 averaging filters. On the right is the visual hull of a branch reconstructed from 50 light sources. The poor result demonstrates the need for better algorithms for reconstructing intricate shapes.

branch computed from the erroneous source positions is woefully inadequate. Thus, better algorithms for improving the accuracy of light source positions are crucial for obtaining 3D models of intricate shapes.

4 Epipolar Geometry

Analogous to the scenario of binocular stereo, we define the epipolar geometry between a pair of shadowgrams that are generated by placing the point source in two locations (L_1 and L_2 in Figure 8). Here, the locations of the point source are analogous to the centers-of-projection of the stereo cameras. The baseline connecting the two light sources L_1 and L_2 intersects the shadowgram plane Π at the epipole E_{12}. When the light sources are equidistant from the shadowgram plane Π, the epipole is at infinity. Based on these definitions, we make two key observations that do not hold for binocular stereo: since the shadowgrams are coplanar, (a) they share the *same epipole* and (b) the points on the two shadowgrams corresponding to the same scene point lie on the *same epipolar line*.

Let $L_i = (u_i, v_i, w_i)^T$ and $L_j = (u_j, v_j, w_j)^T$ be the 3D coordinates of the two light sources, and E_{ij} be the homogeneous coordinate of the epipole on the plane Π, defined by L_i and L_j. Then, the observations (a) and (b) are written as:

$$\boldsymbol{M}_{ij} \, E_{ij} = 0 \tag{9}$$

$$\boldsymbol{m}_i^T \boldsymbol{F}_{ij} \boldsymbol{m}_j = 0 \tag{10}$$

In Equation (9), $\boldsymbol{M}_{ij}$ is a 2×3 matrix composed of two plane equations in the rows

$$\boldsymbol{M}_{ij} = \begin{pmatrix} -\Delta v & \Delta u & u_i v_j - u_j v_i \\ -\Delta u \Delta w & -\Delta v \Delta w & (u_i \Delta u + v_i \Delta v)\Delta w - w_i(\Delta u^2 + \Delta v^2) \end{pmatrix} \tag{11}$$

where, $\Delta u = u_j - u_i$, $\Delta v = v_j - v_i$, and $\Delta w = w_j - w_i$. In Equation (10),

$$\boldsymbol{F}_{ij} = [E_{ij}]_\times \tag{12}$$

is the *fundamental matrix* that relates two corresponding points $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$ between shadowgrams. $[E_{ij}]_\times$ is the 3×3 skew symmetric matrix for which $[E_{ij}]_\times \boldsymbol{x} = E_{ij} \times \boldsymbol{x}$ for any 3D vector $\boldsymbol{x}$.

The camera geometry in coplanar shadowgram is a special case of the parallax geometry [16, 4] where the image deformation is decomposed into a planar homography and a residual image parallax vector. In our system, however, the homography is exactly known to be an identity, which allows us to recover the epipolar geometry *only* from acquired images accurately regardless of the number of views or the complexity of the shadowgram contours.

4.1 Algorithm for estimating epipolar geometry

Consider the plane in Figure 8 that includes the baseline and is tangent to the surface of an object at a *frontier point* F. The intersection of this plane and the shadowgram plane Π forms an epipolar line that can be estimated as one that is cotangent to the two shadowgrams (at T_1 and T_2 in Figure 8). Two such epipolar lines can then be intersected to localize the epipole [4].

Figure 9(a) illustrates the simplest case of two convex shadowgrams overlapping each other. There are only two cotangent lines that touch the shadowgrams at the top and bottom region, resulting in a unique epipole E. When the convex shadowgrams do not overlap each other, four distinct cotangent lines are possible, generating six candidate epipoles, as shown by dots in Figure 9(b). Only two of these four cotangent lines pass through the actual epipole, hence, the other two are false detections. Indeed, the false detections correspond to infeasible cases where the light source is located between the object and the

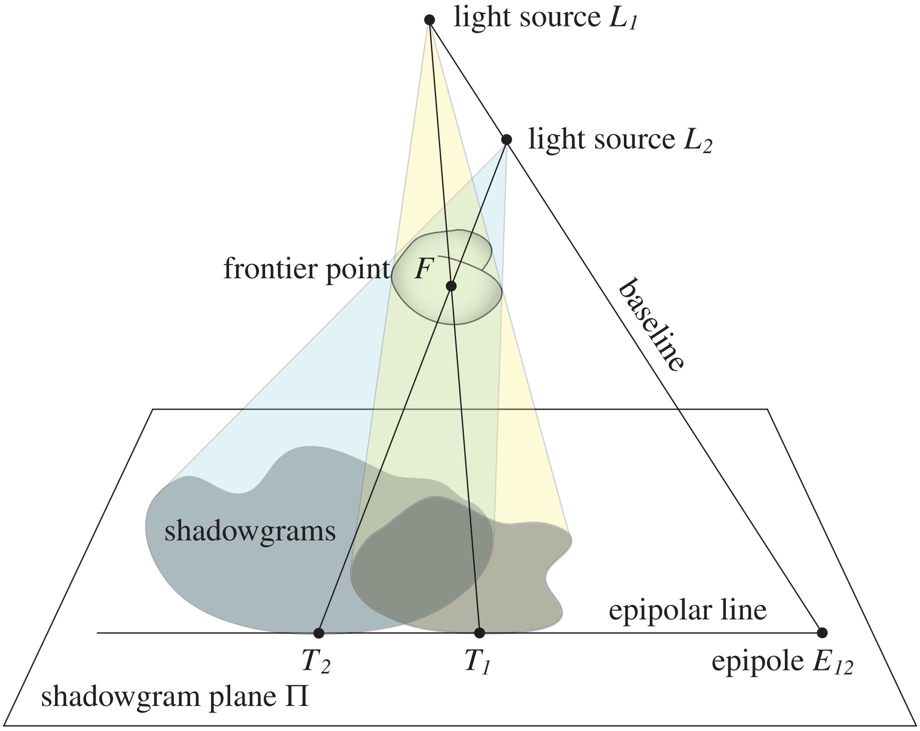


Figure 8: Epipolar geometry of two shadowgrams. The baseline connecting the two sources L_1 and L_2 intersects the shadowgram plane Π at an epipole E_{12}. Suppose an epipolar plane that is tangent to the surface of an object at a frontier point F, then the intersection of the epipolar plane and the shadowgram plane Π is an epipolar line. The epipolar line can be estimated as a line that is co-tangent to the shadowgrams at T_1 and T_2.

screen, or behind the screen. We can detect actual epipolar lines by choosing the cotangent lines where the epipole does not appear between the two points of shadowgram tangency.

When shadowgrams are non-convex, the number of cotangent lines can be arbitrarily large depending on the complexity of the shadowgram contours. Figure 9(c) illustrates the multiple candidates of cotangent lines at the point of tangency T. In this case, we compute the convex polygon surrounding the silhouette contour as shown in Figure 9(d) and prove the following proposition (see Appendix A for the proof):

Proposition 1 *If silhouette contours are consistent in that they can be generated from a physical 3D object, then the convex polygons obtained from the silhouette contours are also consistent.*

Using Proposition 1, the problem of estimating epipolar lines is reduced to the case of

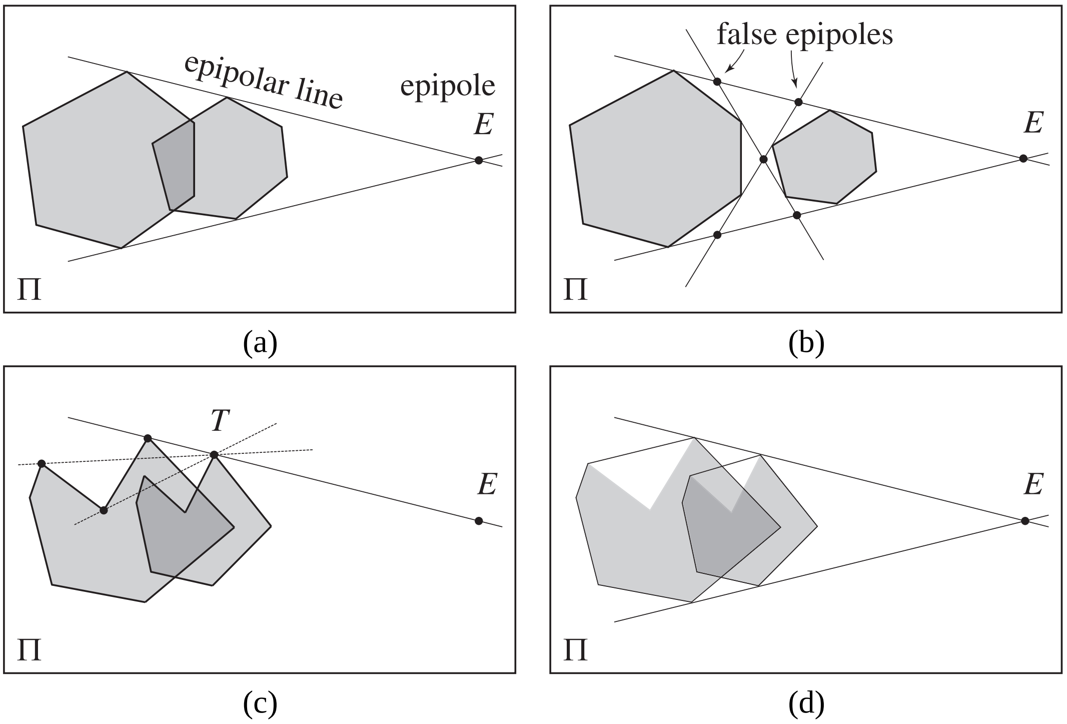


Figure 9: Localization of the epipole. (a),(b) If two shadowgrams are convex, a maximum of four co-tangent lines and six intersections are possible. Considering that the object and the light source are on the same side with respect to the screen, the epipole can be chosen uniquely out of the six intersections. (c),(d) If the shadowgrams are non-convex, the epipole is localized by applying the technique in (a) or (b) to the convex polygons of the original shadowgrams.

either (a) or (b). Thus, epipolar geometry can be reconstructed uniquely and automatically from only the shadowgrams. This capability of recovering epipolar geometry is independent of the shape of silhouette, and hence, the 3D shape of the object. Even when the object is a sphere, we can recover the epipolar geometry without any ambiguity. In traditional multi-view camera-based imaging, epipolar reconstruction requires at least seven pairs of correspondences [9]. Table 2 summarizes the differences between traditional imaging and coplanar shadowgrams in terms of recovering epipolar geometry.

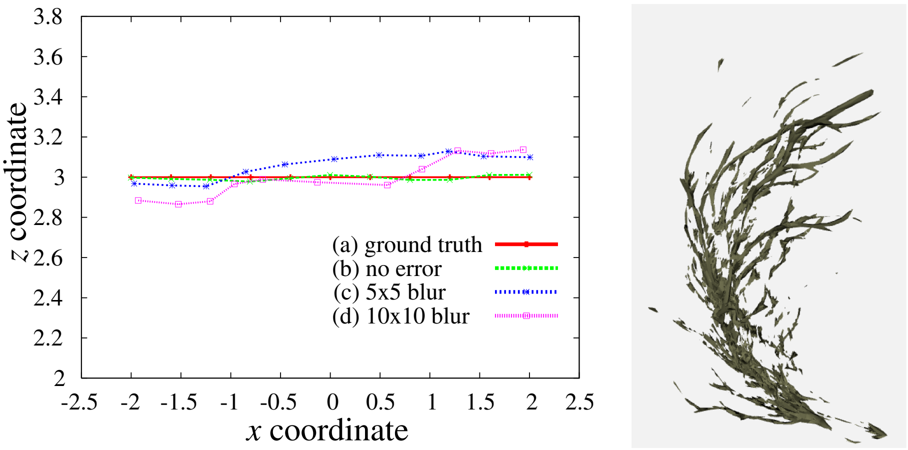


Figure 10: Initial light source positions in Figure 7 were improved by epipolar constraints in Equation (13). On the right is the visual hull reconstructed from the improved source positions.

4.2 Improving accuracy of source locations

The error in the light source positions reconstructed using spheres can be arbitrarily large depending on the localization of the elliptical shadowgram for each sphere. This error can be reduced by relating different light source positions through the epipolar geometry. Let the set of epipoles E_{ij} be estimated from all the source pairs L_i and L_j. The locations of the sources are improved by minimizing the expression in Equation (9) for each pair of light sources using least squares:

$$\{L_k^*\} = \underset{L_k}{\operatorname{argmin}} \sum_{i \neq j} \left\| \boldsymbol{M}_{ij} \, E_{ij} \right\|_2^2 \tag{13}$$

where $\|\cdot\|_2$ is the L2-norm of a vector. The source positions reconstructed from the shadowgrams of spheres are used as initial estimates. We evaluate this approach using the simulated silhouettes described in Figure 7. Figure 10 shows considerable improvement in accuracy obtained by enforcing the epipolar constraint in Equation (9). Compared to the result in Figure 7, collinearity in the positions of light sources is better recovered in this example.

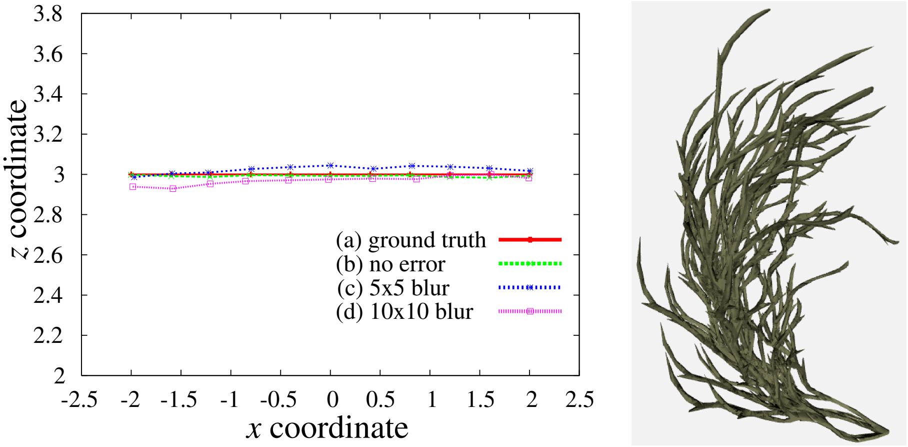


Figure 11: The light source positions reconstructed using epipolar constraint in Figure 10 were optimized by maximizing the shadowgram consistency in Equation (18). On the right is the visual hull reconstructed from the optimized source positions.

5 Using Shadowgram Consistency

While the epipolar geometry improves the estimation of the light source positions, the accuracy of estimate can still be insufficient for the reconstruction of intricate shapes (Figure 10). In this section, we present an optimization algorithm that improves the accuracy of all the source positions even more significantly.

5.1 Optimizing light source positions

Let V be the visual-hull obtained from the set of captured shadowgrams $\{S_k\}$ and the estimated projection matrices $\{\boldsymbol{P}(L_k)\}$. When V is re-projected back onto the shadowgram plane, we obtain the silhouettes S_k^V:

$$S_k^V = \boldsymbol{P}(L_k)\, V\,. \tag{14}$$

Due to the nature of the intersection operator, the re-projected silhouettes S_k^V always satisfy:

$$\forall k : S_k^V \subseteq S_k\,. \tag{15}$$

Only when the source positions are perfect, will the reprojected silhouettes match the acquired silhouettes. Thus, we can define a measure of silhouette mismatch by the sum of

Table 2: Differences between traditional multi-view camera-based imaging and coplanar shadowgrams in epipolar reconstruction. The traditional multi-view images require at least 7 point correspondences between the silhouette contours. Coplanar shadowgrams allow unique epipolar reconstruction irrespective of the shape of the 3D object.

Silhouette complexity	Convex	Non-convex		
#correspondences	2	< 7	≥ 7	$\gg 7$
Traditional multi-camera	impossible	impossible	not always	hard
Coplanar shadowgrams	possible	possible	possible	possible

possible — The epipolar geometry can be reconstructed uniquely.
not always — Possible if seven correspondences are found.
hard — Hard to find the correct correspondences in practice.
impossible — Impossible because of the insufficient constraints.

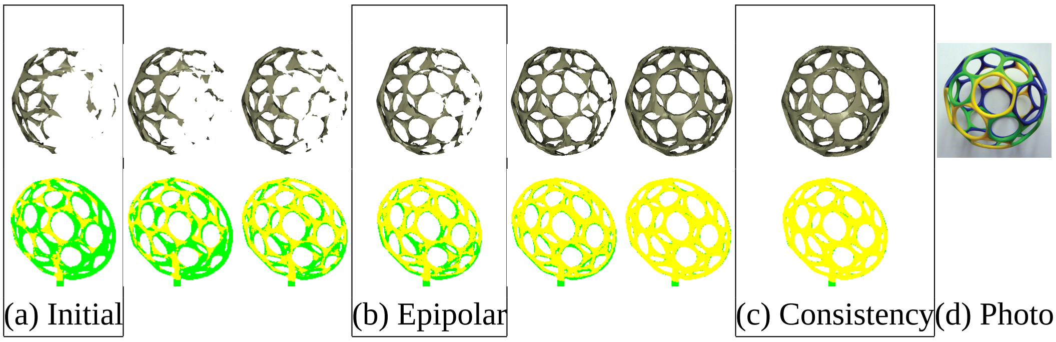

(a) Initial (b) Epipolar (c) Consistency (d) Photo

Figure 12: Reconstructed shape of a thin wire-frame object is improved with each iteration from left to right. (Top) Reconstructed visuals hull at the end of each iteration. (Bottom) The reprojection of the reconstructed visual hulls onto one of captured silhouette images. The reprojection and silhouettes are consistent at yellow pixels, and inconsistent at green. The boxed figures show the reconstruction from the light source positions (a) estimated from spheres, (b) improved by epipolar geometry, and (c) optimized by maximizing shadowgram consistency.

squared difference:

$$E^2_{reprojection} = \sum_k \sum_{\boldsymbol{x}} \left| S^V_k(\boldsymbol{x}) - S_k(\boldsymbol{x}) \right|^2 \tag{16}$$

where $\boldsymbol{x}$ is a pixel coordinate in silhouette image. We minimize the above mismatch by optimizing for the locations of the light sources. Unfortunately, optimizing Equation (16) solely is known to be inherently ambiguous owing to 4 DOF transformation mentioned in Section 3. To alleviate this issue, we simultaneously minimize the discrepancy between the optimized light source positions L_k and the initial source positions L_k^* estimated from the spheres (Section 3) and epipolar geometry (Section 4):

$$E_{initial}^2 = \sum_k \left\| L_k - L_k^* \right\|_2^2 \tag{17}$$

The final objective function is obtained by a linear combination of the two errors:

$$E_{total} = E_{reprojection}^2 + \alpha E_{initial}^2. \tag{18}$$

where, α is a user-defined weight. While the idea of minimizing silhouette discrepancy is well known in the traditional multi-view camera-based SFS [19, 23, 21, 10], the key advantage over prior work is the reduced number of parameters our algorithm needs to optimize (three per view for the light source position, instead of six per view for rotation and translation of the camera). In turn, this allows us to apply our technique to a much larger number of views than possible before.

5.2 Implementation

We use the signed Euclidean distances as the scalar-valued functions $S_k^V(\boldsymbol{x})$ and $S_k(\boldsymbol{x})$ in Equation (16). The intersection of silhouettes is computed for each 3D ray defined by a pixel in S_k, and then projected back to the silhouette to obtain S_k^V. This is a simplified version of image-based visual hull [13] and has been used in silhouette registration methods [10]. Equation (18) is minimized using Powell's gradient-free technique [14].

Due to the intricate shapes of the silhouettes, the error function in Equation (18) can be complex and may have numerous local minima. We alleviate this issue using the convex polygons of the silhouette contours described in Section 4. Given Proposition 1, we minimize Equation (18) using the convex silhouettes with $\{L_k^*\}$ as initial parameters. The resulting light source positions are in turn used as starting values to minimize Equation (18) with the original silhouettes. Using convex silhouettes, in practice, also speeds up convergence.

We evaluate this approach using the simulated silhouettes described in Figure 7 and 10. Compare the results in Figure 7 (using spheres to estimate source positions) and Figure 10 (enforcing epipolar constraints) with those in Figure 11. The final reconstruction of the tree branch is visually accurate highlighting the performance for our technique.

Table 3: The models used in our experiment. The detail of each experiment is shown in the corresponding figure. The size of the shadowgram indicates the average size of shadowgrams. The mismatch between the input shadowgrams and those generated by reprojecting the estimated visual hull is shown in reprojection error. For simulation data, the ratio between the volumes of the ground truth and the reconstructed visual hulls is shown in volumetric error.

	model	Figure	views	shadowgram size	reprojection err.	volumetric err.
Simulation	coral	13	84	530×270	2.2%	0.15%
	seaweed	14	49	334×417	3.2%	0.21%
	bicycle	15	61	635×425	2.3%	0.12%
	spider	16	76	356×354	1.3%	0.08%
Real	polygon-ball	17	45	126 × 116	3.2%	—
	wreath	18	122	674×490	5.2%	—
	palm-tree	19	56	520×425	4.8%	—
	octopus	20	53	451 ×389	4.6%	—

6 Results

In this section, we demonstrate the accuracy of our techniques using both simulated and real experimental data. Table 3 summarizes the data set used in the experiment. All results of 3D shape reconstructions shown in this paper are generated by exact polyhedral visual hull method proposed by Franco and Boyer [7]. The acquired 3D shape is then rendered using Autodesk Maya rendering package.

6.1 Reconstruction of Visual Hulls and 3D Source Positions

Simulation data:

We have chosen four objects with complex structure in our simulations — a coral, a seaweed (also used in Figure 7, 10, and 11 in the main paper), a bicycle, and a spider. The seaweed and coral objects have many thin sub-branches with numerous occlusions. The bicycle object is composed of very thin structures such as spokes, chains, and gears. The spider object is composed of both thick and thin structure. The simulation experiments with known ground truth shape and source positions are shown respectively in Figure 13, 14, 15, and 16.

Each of the figures is organized as follows: (a) A set of coplanar shadowgrams of the object is generated by a shadow simulator implemented by Direct 3D graphics library. (b) The positions of light source are perturbed with random noise with $\sigma = 5\%$ of the object size, and the silhouettes are blurred by 3×3 averaging filters. (c) The positions of the light sources are recovered using epipolar geometry followed by the maximization of silhouette consistency in (d). For each of (b), (c), and (d), the top row shows one of captured silhouette images (in green), overlaid with the reprojection of the reconstructed visual hulls onto the silhouette (in yellow). The middle row shows the ground truth positions of light sources (in red) and the estimated positions (in yellow). The reconstructed 3D shape is shown at the bottom. Finally, (e) the ground truth 3D shape and (f) the reconstructed visual hull rendered by Maya is shown.

Real data:

We show the 3D shape reconstruction of four different objects — a polygon-ball (also used in the Figure 12 in the main paper), a wreath (Figure 1 and 2), a palm-tree, and an octopus. The wreath object has numerous thin needles which cause severe occlusions. The polygon-ball is a thin wiry polyhedral object. The palm-tree object is a plastic object composed of two palm trees with flat leaves. The octopus object is a relatively simple structure, but has complex surface reflection and large concavities. The results of reconstructing 3D shape are shown in Figure 17, 18, 19, and 20. Each figure is organized in the same way as those of simulation data, except that: The final reconstruction of source positions are presented in red in the middle row of (b), (c), and (d). The photograph of the object is shown in (e).

6.2 Convergence

Figure 12 illustrates the convergence properties of our optimization algorithm. Figure 12(a) shows the visual hull of the wiry polyhedral object obtained using the initial positions of light sources estimated from the calibration spheres. The reprojection of the visual hull shows poor and incomplete reconstruction. By optimizing the light source positions, the quality of the visual hull is noticeably improved in only a few iterations.

The convergence of the reconstruction algorithm is quantitatively evaluated in Figure 21. The error in light source positions estimated by the algorithm proposed in Section 5 is shown in the left plot. The vertical axis shows L2 distance between the ground truth and the current estimate of light source positions. After convergence, the errors in the light source positions are less than 1% of the sizes of the objects. The silhouette mismatch defined in Equation (16) is plotted on the right. On average, the silhouettes cover on the order of 10^5 pixels. The error in the reprojection of the reconstructed visual hulls is less than 1% of the silhouette pixels for the real objects.

7 Discussion of Limitations

Despite significant progress in optical scanning hardware [5, 12] and multi-view geometry [9, 17], reconstruction of intricate shapes remains an open problem. We believe this work is an initial step in the right direction. In this section, we discuss some limitations of our current system and propose possible extension of the coplanar shadowgram imaging system.

7.1 Multi-screen Coplanar Shadowgrams

A single screen cannot be used to capture the complete $360^\circ \times 360^\circ$ view of the object. For instance, it is not possible to capture the silhouette observed in the direction parallel to a shadowgram plane. This limitation can be overcome by augmenting the system with more than one shadowgram screen (or move one screen to different locations). The algorithm of the multi-screen coplanar shadowgram imaging can be divided into offline and online steps:

Off-line Calibration (one-time): This calibration can be done in several ways and we mention a simple one here. In the case of two-screen setup which is observed by a single camera, we only need to estimate the homography between each screen and image plane. The extra work required over the one-screen case is an additional homography estimation. The homographies in turn can be used to recover the relative transformation between the screens.

Online Calibration: In the two-screen setup, we can estimate the light source positions for each set of shadowgrams on one screen separately using the technique demonstrated in the paper. Finally, we merge the two sets of results using the relative orientation between the screens resulting from the off-line calibration.

In principle, it is possible to also optimize (minimize) the errors due to off-line calibration. However, the off-line intrinsic calibration of a camera and the screen-to-image homography can be done carefully, More importantly, it is independent of the complexity of the object and the number of source positions.

We have performed simulations with a bicycle object with two screen positions as shown in Figure 22. The bicycle was chosen since frontal and side views are both necessary to carve the visual hull satisfactorily. Combining the two sets of shadowgrams enlarges the coverage of source positions, which successfully reduces the stretching artifact of the reconstructed shape in Figure 23.

7.2 Other Directions

Another drawback of SFS techniques is the inability to model concavities on the object's surface. Combining our approach with other techniques, such as photometric stereo or multi-view stereo can overcome this limitation, allowing us to obtain appearance together with a smoother shape of the object. Finally, using multiple light sources of different spectra to speed up acquisition, and the analysis of defocus blur due to a light source of finite area are our directions of future work.

References

[1] Bruce Guenther Baumgart. *Geometric modeling for computer vision*. PhD thesis, Stanford University, 1974.

[2] William Henry Besant. *Conic sections, Treated geometrically*. Cambridge, 1890.

[3] Roberto Cipollaand, Kalle Åström, and Peter Giblin. Motion from the frontier of curved surfaces. In *Proc. International Conference on Computer Vision '95*, pages 269–275, 1995.

[4] Geoff Cross, Andrew W. Fitzgibbon, and Andrew Zisserman. Parallax geometry of smooth surfaces in multiple views. In *Proc. International Conference on Computer Vision '99*, pages 323–329, 1999.

[5] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proc. SIGGRAPH '96*, pages 303–312, 1996.

[6] Andrew Fitzgibbon, Maurizio Pilu, and Robert Fisher. Direct least squares fitting of ellipses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):476–480, 1999.

[7] Jean-Sébastien Franco and Edmond Boyer. Exact polyhedral visual hulls. In *Proc. the 15th British Machine Vision Conference*, pages 329–338, 2003.

[8] Yasutaka Furukawa, Amit Sethi, Jean Ponce, and David Kriegman. Robust structure and motion from outlines of smooth curved surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):302–315, 2006.

[9] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edition, 2004.

[10] Carlos Hernández, Francis Schmitt, and Roberto Cipolla. Silhouette coherence for camera calibration under circular motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2):343–349, 2007.

[11] David J. Kriegman and Peter N. Belhumeur. What shadows reveal about object structure. *Journal of the Optical Society of America*, 18(8):1804–1813, 2001.

[12] Marc Levoy, Kari Pulli, Brian Curless, Szymon Rusinkiewicz, David Koller, Lucas Pereira, Matt Ginzton, Sean Anderson, James Davis, Jeremy Ginsberg, Jonathan Shade, and Duane Fulk. The digital michelangelo project: 3D scanning of large statues. In *Proc. SIGGRAPH'00*, pages 131–144, 2000.

[13] Wojciech Matusik, Chris Buehler, Ramesh Raskar, Steven J. Gortler, and Leonard McMillan. Image-based visual hulls. In *Proc. SIGGRAPH 2000*, pages 369–374, 2000.

[14] William Press, Brian Flannery, Saul Teukolsky, and William Vetterling. *Numerical Recipes in C*. Cambridge University Press, 1988.

[15] Silvio Savarese, Marco Andreetto, Holly Rushmeier, Fausto Bernardini, and Pietro Perona. 3d reconstruction by shadow carving: Theory and practical evaluation. *International Journal of Computer Vision*, 71(3):305–336, 2005.

[16] Harpreet S. Sawhney. Simplifying motion and structure analysis using planar parallax and image warping. In *Proc. International Conference of Pattern Recognition*, pages 403–408, 1994.

[17] Steven M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. Computer Vision and Pattern Recognition 2006*, volume 1, pages 519–526, 2006.

[18] Gary S. Settles. *Schlieren & Shadowgraph Techniques*. Springer-Verlag, 2001.

[19] Sudipta N Sinha, Marc Pollefeys, and Leonard McMillan. Camera network calibration from dynamic silhouettes. In *Proc. Computer Vision and Pattern Recognition 2004*, volume 1, pages 195–202, 2004.

[20] Alvy Ray Smith and James F. Blinn. Blue screen matting. In *Proc. SIGGRAPH '96*, pages 259–268, 1996.

[21] Kwan-Yee K. Wong and Roberto Cipolla. Reconstruction of sculpture from its profiles with unknown camera positions. *IEEE Transactions on Image Processing*, 13(3):381–389, 2004.

[22] Shuntaro Yamazaki, Srinivasa Narasimhan, Simon Baker, and Takeo Kanade. On using coplanar shadowgrams for visual hull reconstruction. Technical Report CMU-RI-TR-07-29, Carnegie Mellon University, August 2007.

[23] Anthony J. Yezzi and Stefano Soatto. Stereoscopic segmentation. *International Jornal of Computer Vision*, 1(53):31–43, 2003.

Appendix

A Proof of Proposition 1

Proof Given a set X in 2D or 3D space, the convex set $\hat{X}$ of X is defined as

$$\hat{X} \overset{\text{def}}{=} \bigcup_{p_m, p_n \in X} \overline{p_m p_n}. \tag{19}$$

Suppose a visual hull V_S reconstructed from a set of silhouettes $\{S_i\}$ ($i = 1, \cdots, N$) is consistent, then

$$\forall i : P_i V_S = S_i \tag{20}$$

holds by definition. The convex polygons $\{\hat{S}_i\}$ of $\{S_i\}$ is written as

$$\hat{S}_i \overset{\text{def}}{=} \bigcup_{p_m, p_n \in S_i} \overline{p_m p_n}. \tag{21}$$

A visual hull $V_{\hat{S}}$ reconstructed from $\{\hat{S}_i\}$ is

$$V_{\hat{S}} = \bigcap_i^N P_i^{-1} \hat{S}_i. \tag{22}$$

Projecting both sides of Equation (22) to all silhouette views,

$$\forall i : P_i V_{\hat{S}} \subseteq \hat{S}_i, \tag{23}$$

If the equation in Equation (23) does not hold, there exists a 2D point p that is included in a silhouette, but not included in the back-projection of reconstructed visual hull to the silhouette views.

$$\exists i : P_i V_{\hat{S}} \subset \hat{S}_i \tag{24}$$

$$\Rightarrow \exists i \exists p : p \in \hat{S}_i \wedge p \notin P_i V_{\hat{S}} \tag{25}$$

From Equation (21), Equation (25) becomes

$$\exists i \exists p_1 \exists p_2 \exists p : p_1 \in S_i \wedge p_2 \in S_i \wedge p \in \overline{p_1 p_2} \wedge p \notin P_i V_{\hat{S}}. \tag{26}$$

From Equation (20), p_1 and p_2 have respectively corresponding 3D points q_1 and q_2 in V_S. Suppose the projection of p into 3D space intersects at q with a 3D line segment $\overline{q_1 q_2}$, then Equation (26) becomes

$$\exists q_1 \exists q_2 \exists q : q_1 \in V_S \wedge q_2 \in V_S \wedge q \in \overline{q_1 q_2} \wedge q \notin V_{\hat{S}}. \tag{27}$$

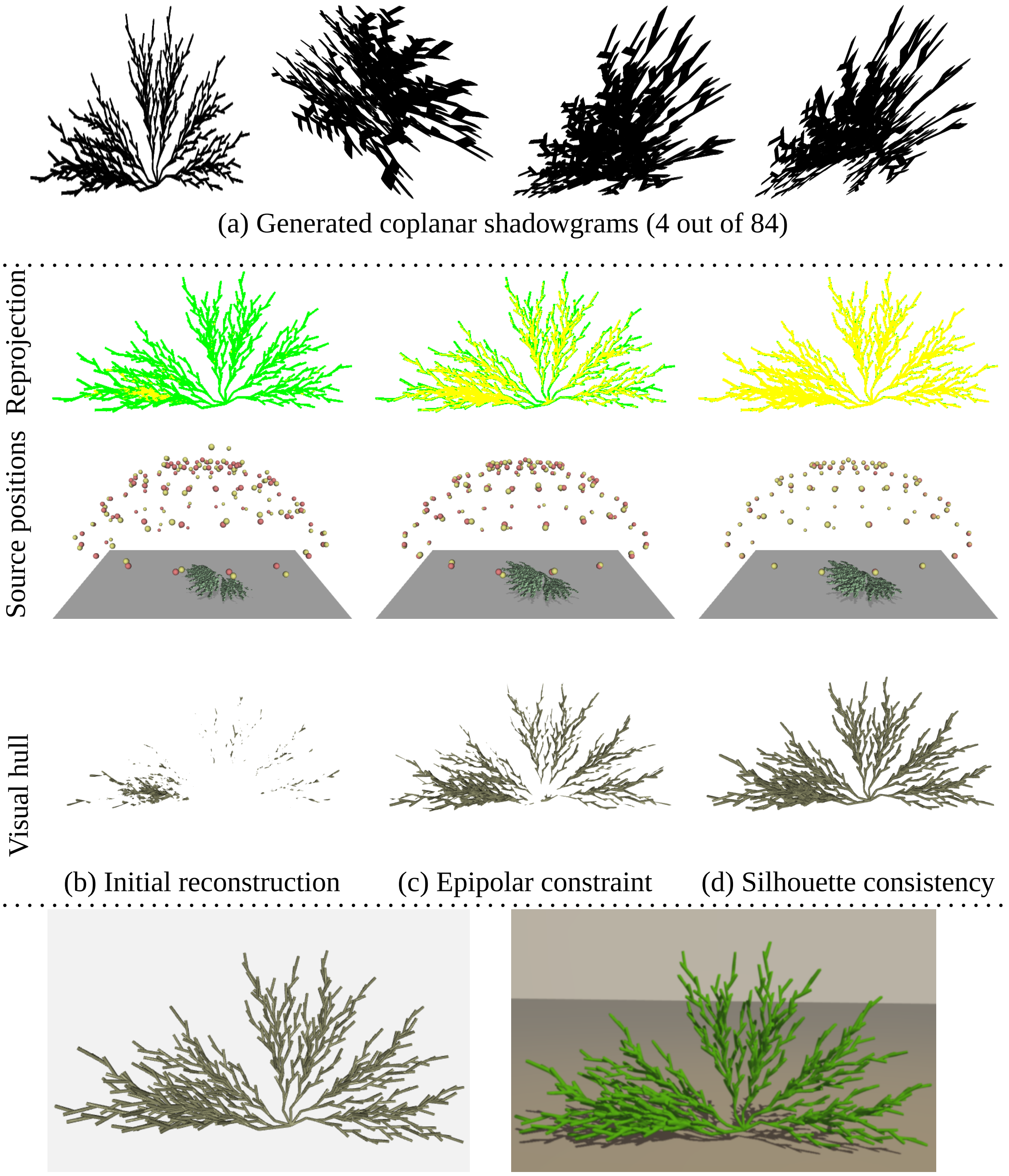
By projecting all terms in Equation (27) to silhouette views,

$$\forall j \exists p_1^j \exists p_2^j \exists p^j : p_1^i \in S_j \wedge p_2^i \in S_j \wedge p^i \in \overline{p_1^j p_2^j} \wedge p^i \notin \hat{S}_j, \tag{28}$$

where $p_1^j = P_j q_1$, $p_2^j = P_j q_2$ and $p^j = P_j q$. $P_j V_S$ is substituted with S_j by Equation (20). This is contradictory to the definition of $\hat{S}_i$ in Equation (21), which concludes the false hypothesis of Equation (24). Hence, the relation in Equation (23) gives the equation

$$\forall i : P_i V_{\hat{S}} = \hat{S}. \tag{29}$$

From Equation (22) and Equation (29), $V_{\hat{S}}$ is consistent by definition. ∎



(a) Generated coplanar shadowgrams (4 out of 84)

(b) Initial reconstruction

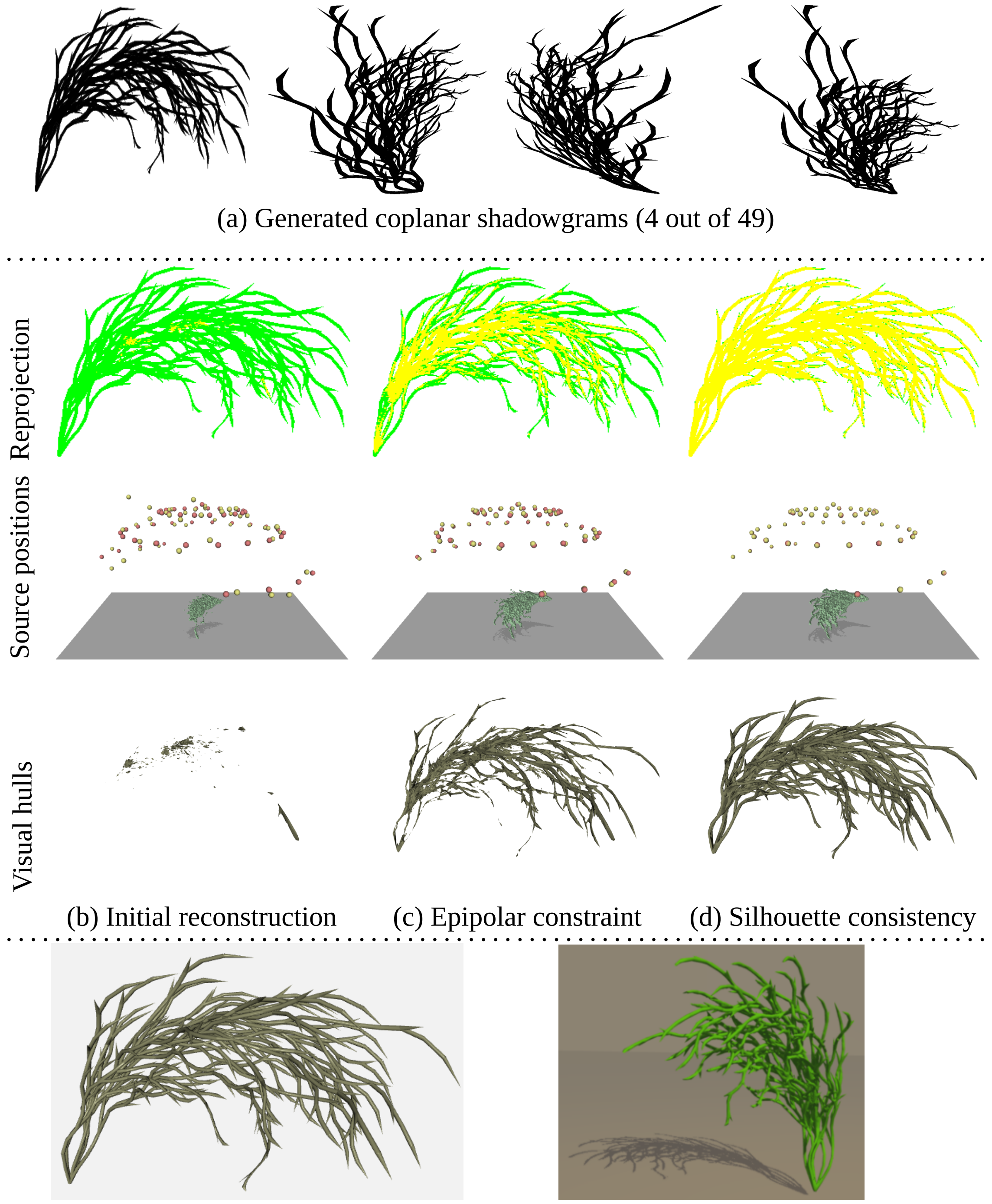
(c) Epipolar constraint

(d) Silhouette consistency

(e) Ground truth shape

(f) Rendering of reconstructed visual hull

Figure 13: Simulation with a **coral** object. (a) Eighty four coplanar shadowgrams of the object are generated with average resolution 530 × 270 pixels. (b) Initial reconstruction. (c) The reconstruction using epipolar geometry. (d) The reconstruction using silhouette consistency. (e) The ground truth 3D shape. The volume difference is 0.15% of the volume of the ground truth 3D shape. (f) Rendering of the reconsturcted shape. (Refer to main text for the detail of each figure.)

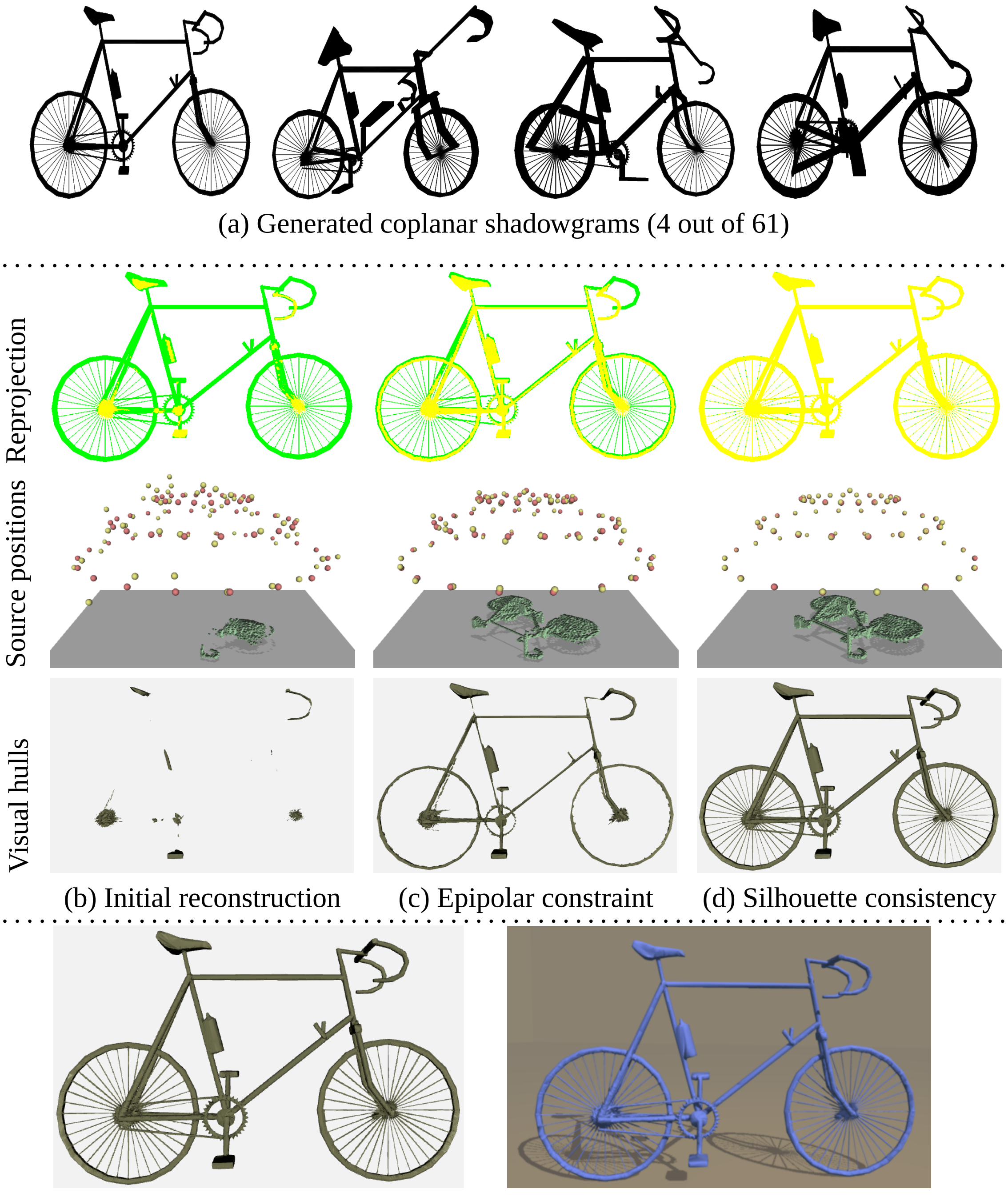


(a) Generated coplanar shadowgrams (4 out of 49)

(b) Initial reconstruction (c) Epipolar constraint (d) Silhouette consistency

(e) Ground truth shape (f) Rendering of reconstructed visual hull

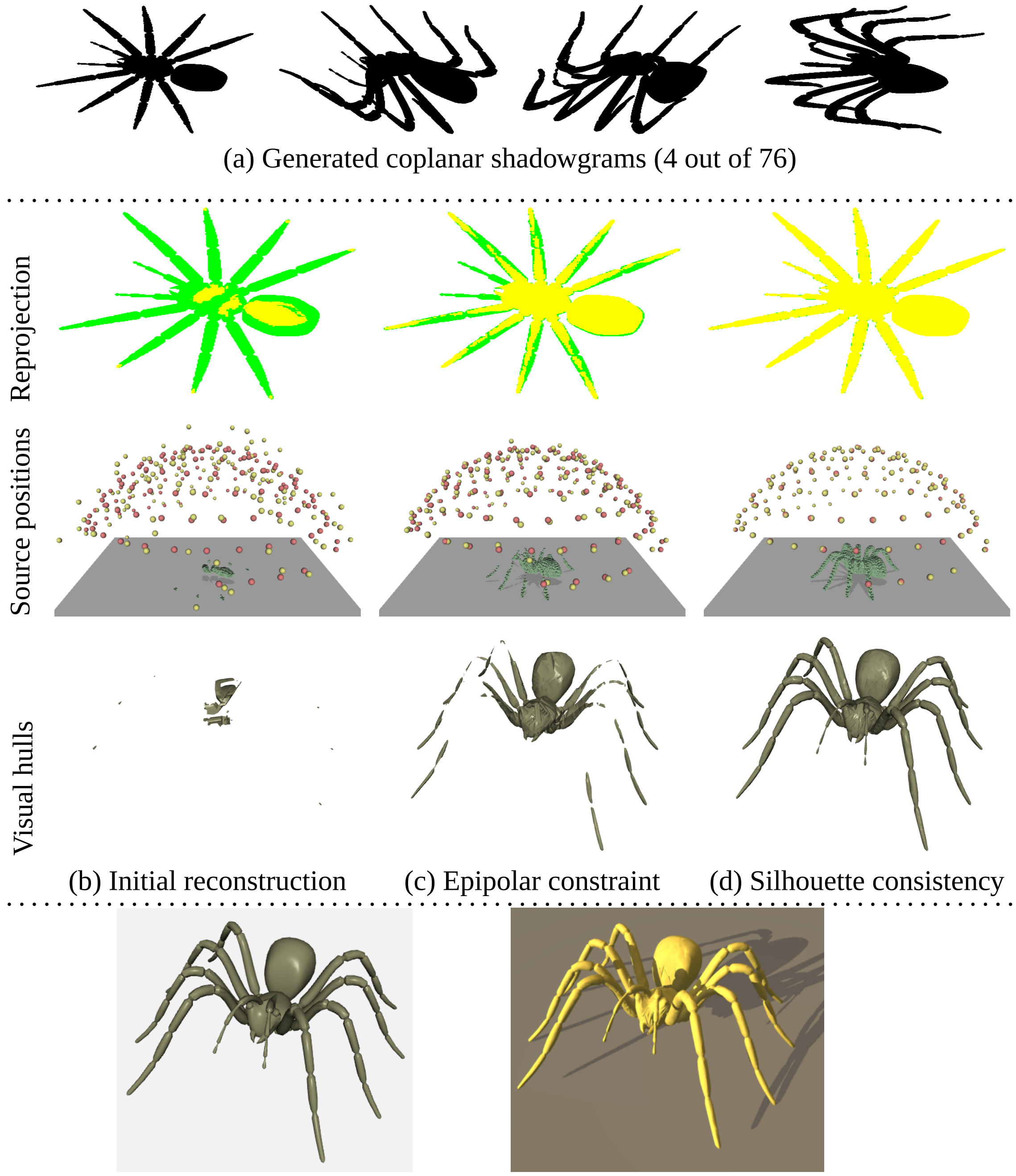
Figure 14: Simulation with a **seaweed** object. (a) Forty nine coplanar shadowgrams of the object are generated with average resolution 334×417 pixels. (b) Initial reconstruction. (c) The reconstruction using epipolar geometry. (d) The reconstruction using silhouette consistency. (e) The ground truth 3D shape. The volume difference is 0.21% of the volume of the ground truth 3D shape. (f) Rendering of the reconsturcted shape. (Refer to main text for the detail of each figure.)


(a) Generated coplanar shadowgrams (4 out of 61)

(b) Initial reconstruction (c) Epipolar constraint (d) Silhouette consistency

(e) Ground truth shape (f) Rendering of reconstructed visual hull

Figure 15: Simulation with a **bicycle** object. (a) Sixty one coplanar shadowgrams of the object are generated with average resolution 635 × 425 pixels. (b) Initial reconstruction. (c) The reconstruction using epipolar geometry. (d) The reconstruction using silhouette consistency. (e) The ground truth 3D shape. The volume difference is 0.12% of the volume of the ground truth 3D shape. (f) Rendering of the reconsturcted shape. (Refer to main text for the detail of each figure.)

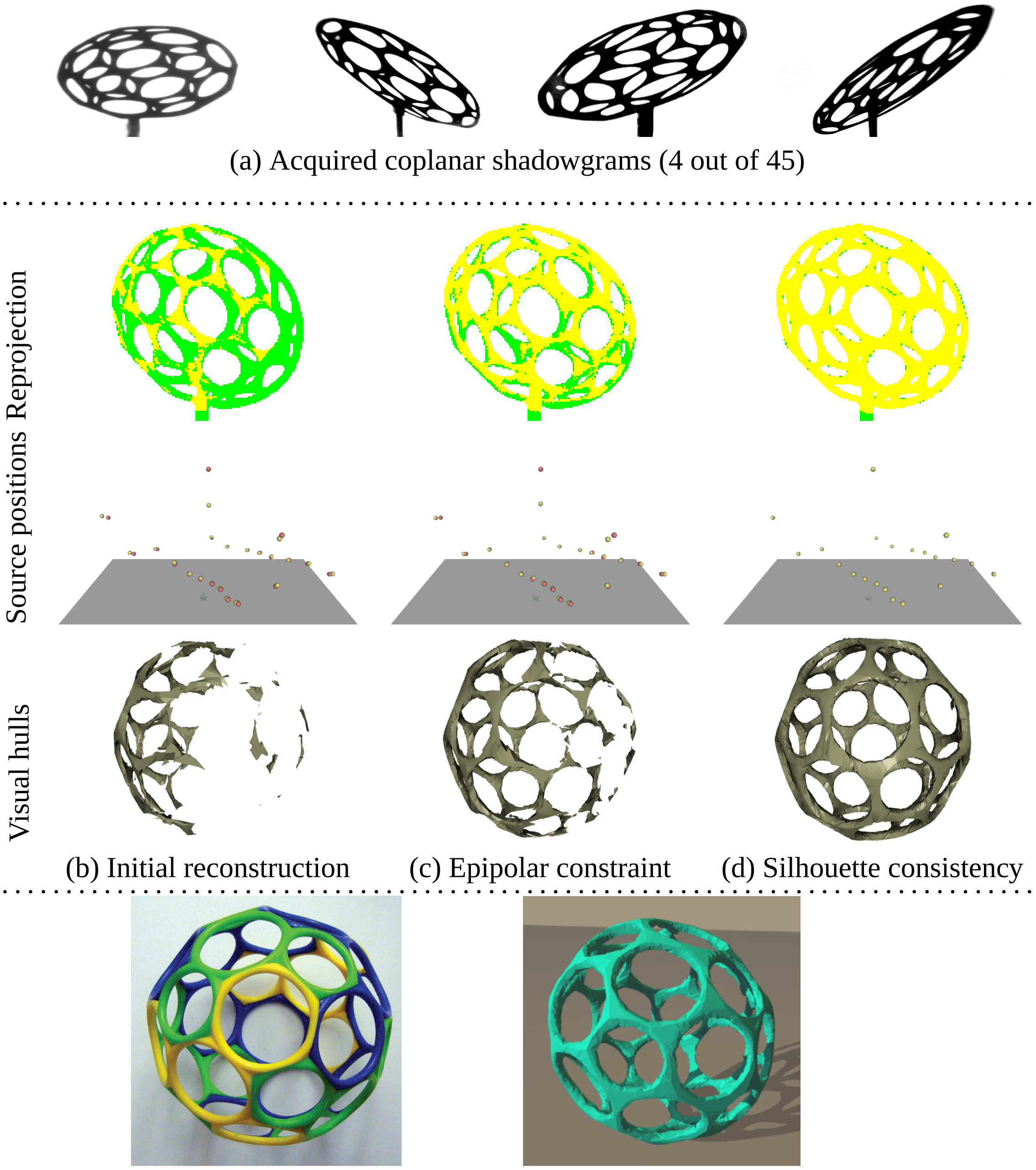


(a) Generated coplanar shadowgrams (4 out of 76)

(b) Initial reconstruction (c) Epipolar constraint (d) Silhouette consistency

(e) Ground truth shape (f) Rendering of reconstructed visual hull

Figure 16: Simulation with a **spider** object. (a) Seventy six coplanar shadowgrams of the object are generated with average resolution 356 × 354 pixels. (b) Initial reconstruction. (c) The reconstruction using epipolar geometry. (d) The reconstruction using silhouette consistency. (e) The ground truth 3D shape. The volume difference is 0.08% of the volume of the ground truth 3D shape. (f) Rendering of the reconsturcted shape. (Refer to main text for the detail of each figure.)

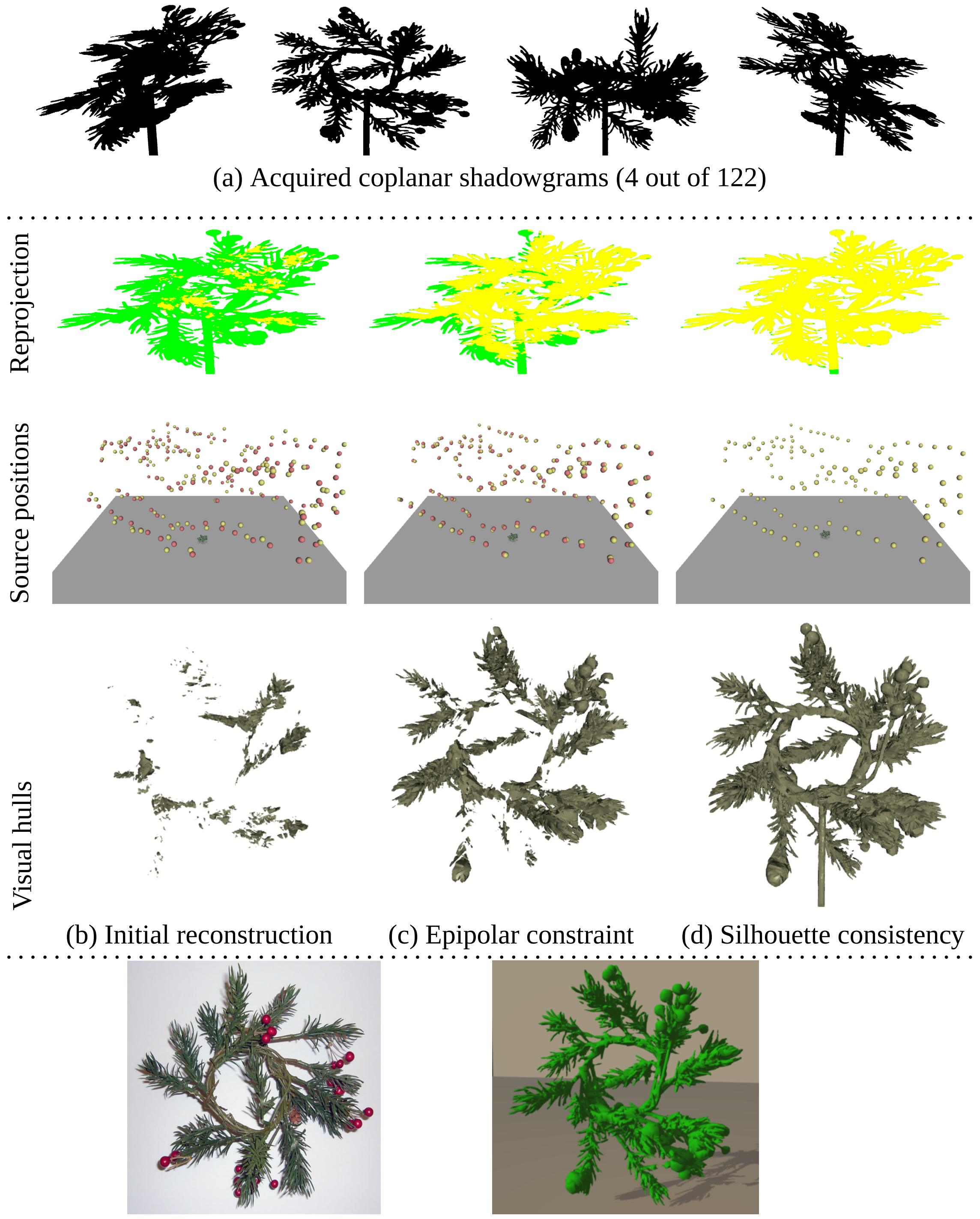


(a) Acquired coplanar shadowgrams (4 out of 45)

(b) Initial reconstruction (c) Epipolar constraint (d) Silhouette consistency

(e) Photograph (f) Rendering of reconstructed visual hull

Figure 17: Real experiment with a **polygon-ball**. (a) Forty five coplanar shadowgrams of the object are generated with average resolution 126×116 pixels. (b) Initial reconstruction. (c) The reconstruction using epipolar geometry. (d) The reconstruction using silhouette consistency. (e) Photograph of the object. (f) Photograph of the object. (Refer to main text for the detail of each figure.)

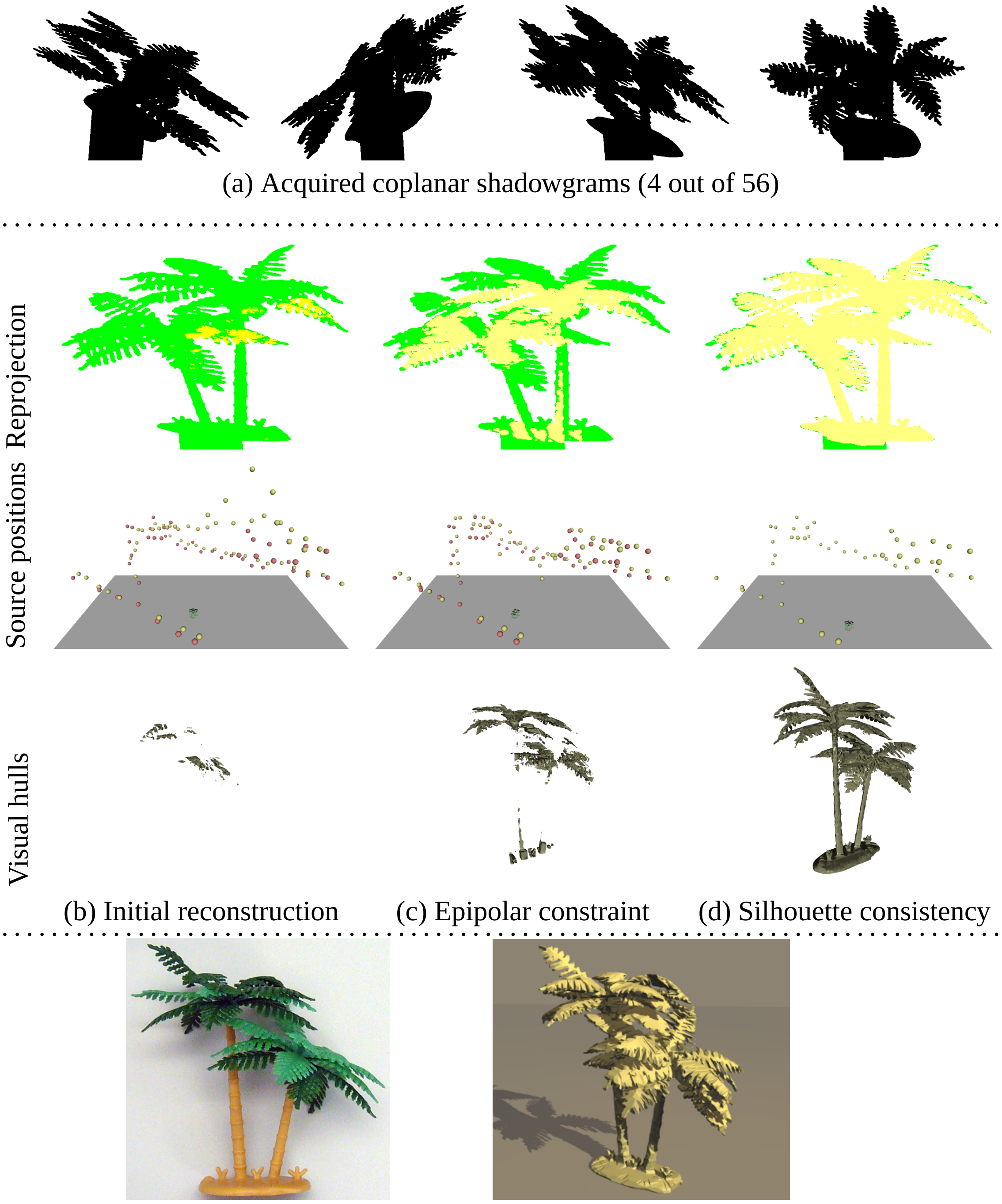


(a) Acquired coplanar shadowgrams (4 out of 122)

(b) Initial reconstruction (c) Epipolar constraint (d) Silhouette consistency

(e) Photograph (f) Rendering of reconstructed visual hull

Figure 18: Real experiment with a **wreath**. (a) 122 coplanar shadowgrams of the object are generated with average resolution 674 × 490 pixels. (b) Initial reconstruction. (c) The reconstruction using epipolar geometry. (d) The reconstruction using silhouette consistency. (e) Photograph of the object. (f) Photograph of the object. (Refer to main text for the detail of each figure.)

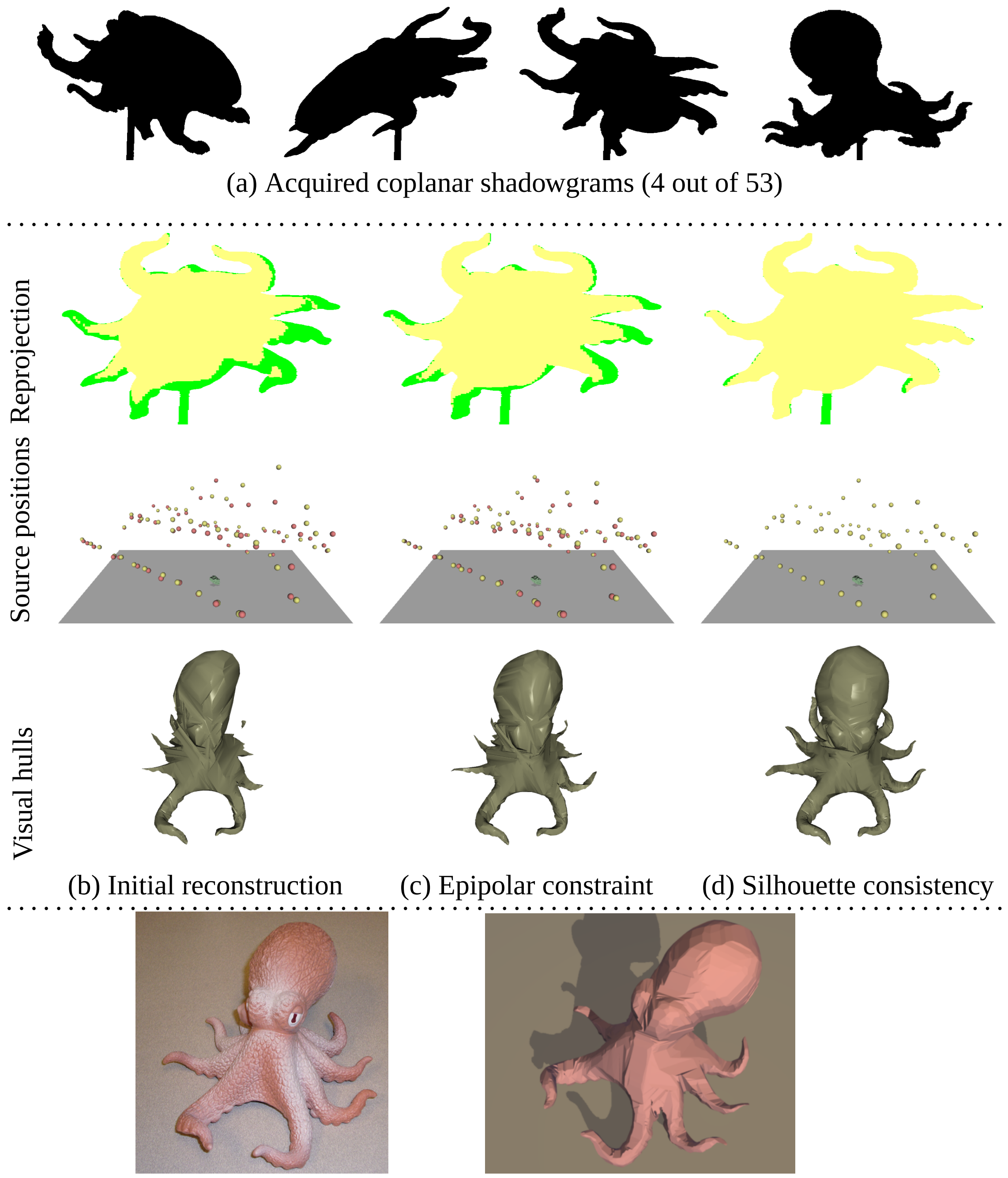


(a) Acquired coplanar shadowgrams (4 out of 56)

(b) Initial reconstruction (c) Epipolar constraint (d) Silhouette consistency

(e) Photograph (f) Rendering of reconstructed visual hull

Figure 19: Real experiment with a **palm-tree**. (a) Fifty six coplanar shadowgrams of the object are generated with average resolution 520 × 425 pixels. (b) Initial reconstruction. (c) The reconstruction using epipolar geometry. (d) The reconstruction using silhouette consistency. (e) Photograph of the object. (f) Rendering of the reconsturcted shape. (Refer to main text for the detail of each figure.)



(a) Acquired coplanar shadowgrams (4 out of 53)

(b) Initial reconstruction (c) Epipolar constraint (d) Silhouette consistency

(e) Photograph (f) Rendering of reconstructed visual hull

Figure 20: Real experiment with an **octopus**. (a) Fifty three coplanar shadowgrams of the object are generated with average resolution 451×389 pixels. (b) Initial reconstruction. (c) The reconstruction using epipolar geometry. (d) The reconstruction using silhouette consistency. (e) Photograph of the object. (f) Rendering of the reconsturcted shape. (Refer to main text for the detail of each figure.)

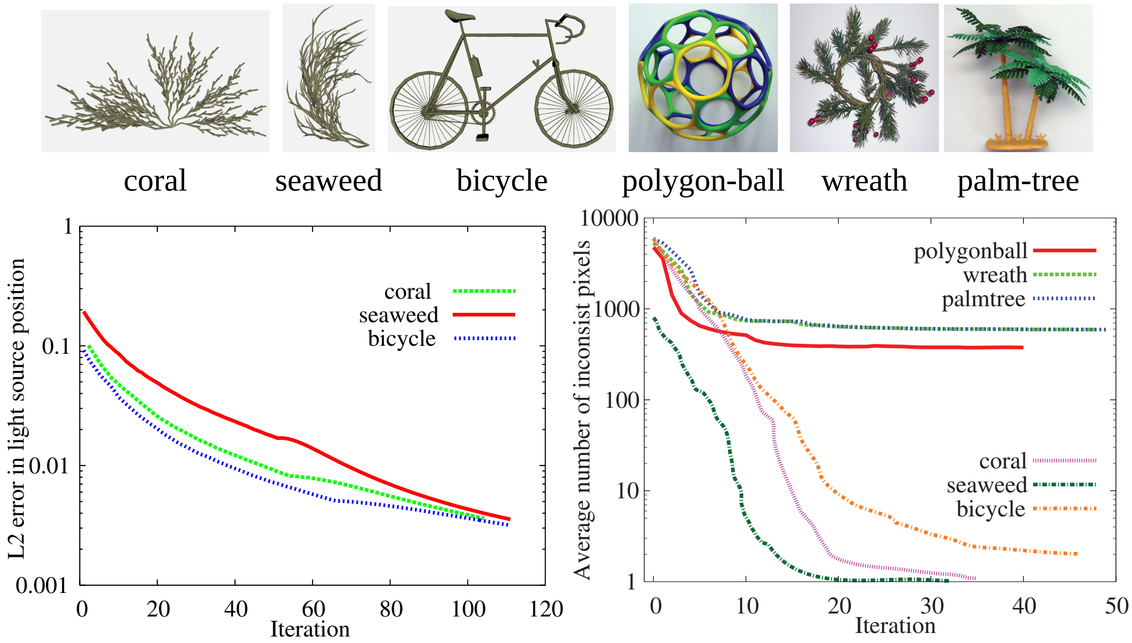


Figure 21: Convergence of error: (Left) Error in light source positions is computed using ground truth positions for simulation models. (Right) Error in shadowgram consistency. Both plots are in logarithmic scale.

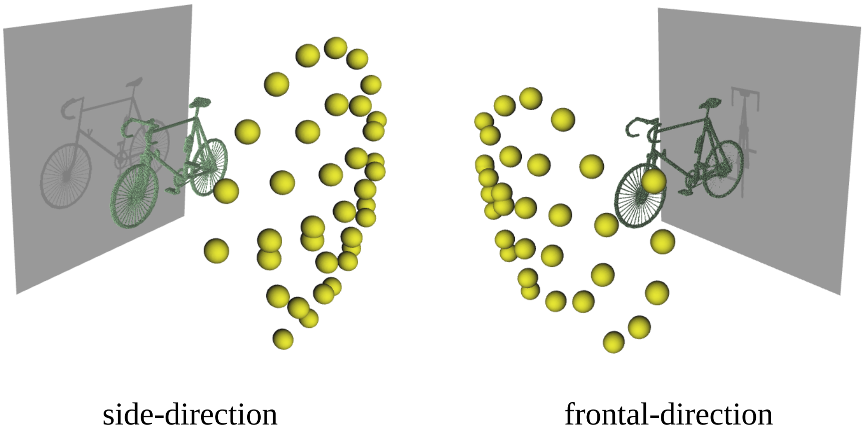


Figure 22: Two different configurations of coplanar shadowgrams of a bicycle object. Gray rectangle and yellow spheres indicate respectively a shadow screen and light source position. 36 light sources are used in both configurations. The screen is rotated by 90 degrees, while the object remains fixed. For the demonstration of the two-screen algorithm, a small number of light sources are used.

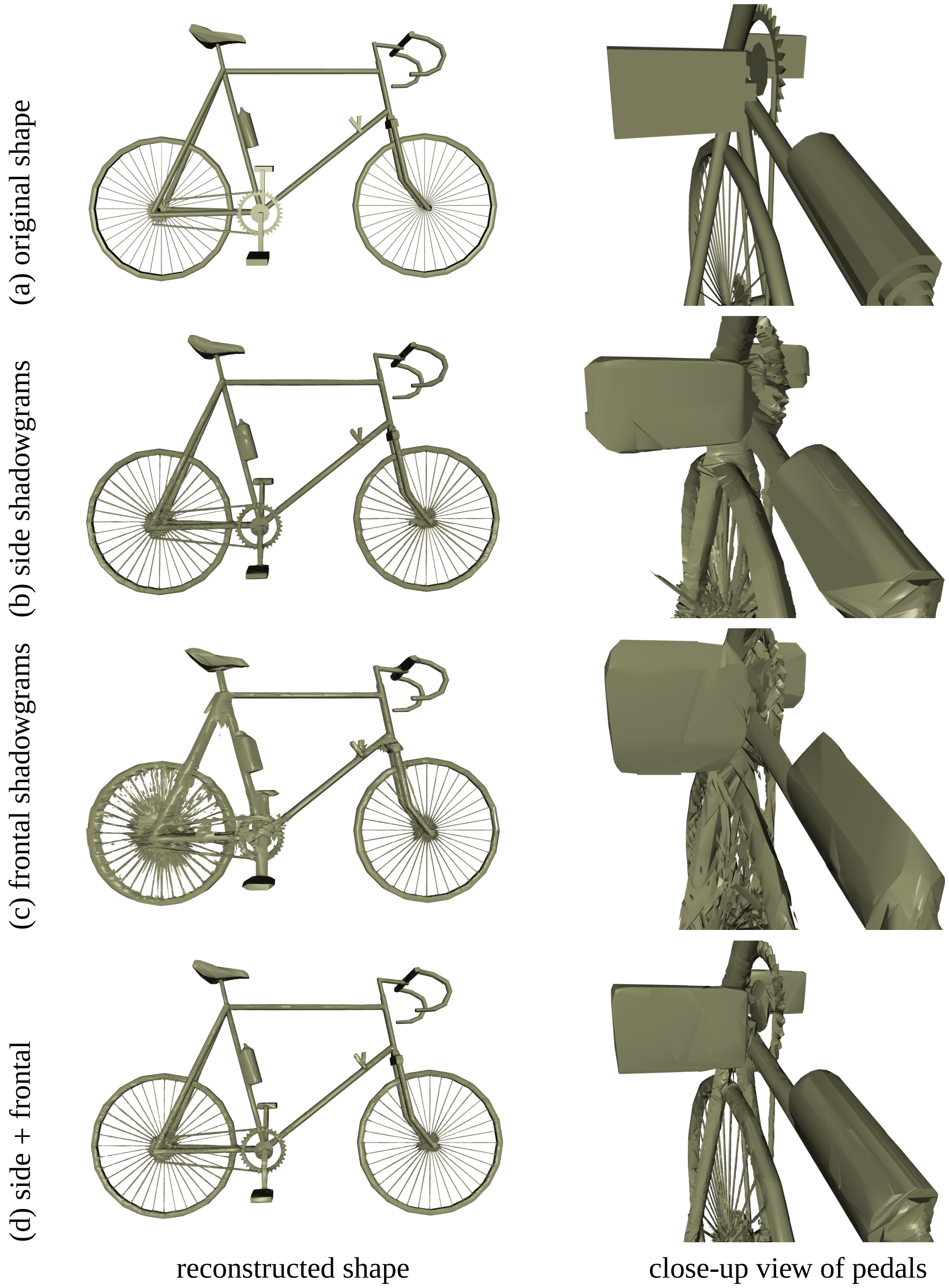


Figure 23: Comparison of shape reconstruction. We synthesized 36 coplanar shadowgrams of a 3D shape shown in (a). The visual hull of the object is reconstructed from: (b) the shadowgrams taken from side-direction (Figure 22 left) and (c) the shadowgrams taken from frontal-direction (Figure 22 right). The reconstructed shape is stretched into the direction perpendicular to a shadow screen due to the lack of views parallel to the screen. (d) Combining shadowgrams (b) and (c) enlarges the coverage of light source positions, which successfully reduces the stretching artifact in the reconstructed shape.