Supplementary Materials: Globally Optimal Pose from Orthographic Silhouettes

Agniva Sengupta[1,2] Dilara Kuş[1,2] Jianning Li[2] Stefan Zachow[2]
[1]Freie Universität Berlin [2]Zuse Institute Berlin

Contents

Summary

Given below is the appendix to our article 'Globally Optimal Pose from Orthographic Silhouettes'. It contains proof of theorems and lemmas, algorithms, experimental details, qualitative results, and some additional clarifications. Code, dataset and results are in https://agnivsen.github.io/pose-from-silhouette/

1. Proof of Theorem 1

We complete the proof in three parts, first for a single triangle, second for two triangles, and finally with a generalisation to any finite collection of triangles:

Single triangle. Let $\Delta_a = (\mathbf{P}_{a,1}, \mathbf{P}_{a,2}, \mathbf{P}_{a,3}) \in \mathbb{R}^{3\times 3}$ be a triangular 3D surface. Rotation of Δ_a by any arbitrary rotation matrix $\mathbf{R} \in \mathbb{SO}(3)$ can be orthographically projected to obtain $\Delta_{\Pi,a} = \Pi_{\mathrm{O}}(\mathbf{R}\Delta_a) = (\mathbf{p}_{a,1}, \mathbf{p}_{a,2}, \mathbf{p}_{a,3}) \in \mathbb{R}^{2\times 3}$ whose area can be described by Heron's formula:

$$\begin{aligned} A\big(\Delta_{\Pi,a}\big) = H(d_{a,(1,2)}, d_{a,(2,3)}, d_{a,(3,1)}) &= \sqrt{s_a(s_a - d_{a,(1,2)})(s_a - d_{a,(2,3)})(s_a - d_{a,(3,1)})}, \\ \text{where:} \quad s_a &= \frac{1}{2}(d_{a,(1,2)} + d_{a,(2,3)} + d_{a,(3,1)}), \\ \text{and} \quad d_{a,(x,y)} &= \|\mathbf{p}_{a,x} - \mathbf{p}_{a,y}\| = \|\mathbf{R}_{\{1,2\},:}(\mathbf{P}_{a,x} - \mathbf{P}_{a,y})\|. \end{aligned} \tag{1}$$

In eq. (1), all operations, including L_2-norm and square-root of non-negative values, are Lipschitz-continuous (L-c), thus $A\big(\Delta_{\Pi,a}\big)$ is L-c w.r.t the elements of $\mathbf{R}$. Even in the case of a degenerate triangle caused by the projection of Δ_a to a straight line in the projective plane, causing:

$$d_{a,(l,l')} + d_{a,(l',l'')} = d_{a,(l'',l)},\ \forall (l, l', l'') \in \{(1,2,3), (2,3,1), (3,1,2)\}, \tag{2}$$

eq. (1) is still applicable. Thus, $A\big(\Delta_{\Pi,a}\big)$ is L-c w.r.t any rotation parameterisation which admits an L-c map to $\mathbb{SO}(3)$, since any composition of L-c is still L-c. Similarly, $A\big(\Delta_{\Pi,a}\big)$ would also be L-c w.r.t composition of multiple disjoint rotations as long as each rotation is L-c w.r.t each other. Therefore, sequence of rotations that produce a L-c trajectory through rotation manifold results in a L-c change of orthographic Area-of-Silhouettes (AoS).

Two triangles. The case of two projected triangles is slightly more involved. The general shape describing the overlap of two 2D triangles is a hexagon with up to 3 degenerate vertices. We consider two triangles Δ_a and Δ_b with an overlapping projected region $\Delta_{\Pi,a\cap b} = (\mathbf{p}_{ab,1}, \ldots, \mathbf{p}_{ab,6}) \in \mathbb{R}^{2\times 6}$, including the three possibly degenerate vertices, given that degenerate intersections (edge-collinearity or higher-order intersections) are measure-zero. We assume the vertices of $\Delta_{\Pi,a\cap b}$ to be in clockwise winding order, without loss of generality. Thus the area of $\Delta_{\Pi,a\cap b}$ is given as:

$$\tilde{A}(\Delta_{\Pi,a\cap b}) = O_\Delta \sum_{l=1}^{4} H(d_{a,(1,l+1)}, d_{a,(l+1,l+2)}, d_{a,(l+2,1)}) = O_\Delta G(\Delta_{\Pi,a\cap b}) \tag{3}$$

where O_Δ is a flag indicating overlap of projected triangles $\Delta_{\Pi,a}$ and $\Delta_{\Pi,b}$ such that:

$$O_\Delta = \begin{cases} 1 & \text{If} \quad \Delta_{\Pi,a} \cap \Delta_{\Pi,b} \neq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The L-c-ness of $G(\Delta_{\Pi,a\cap b})$ is easy to verify by noticing that infinitesimal rotation applied to Δ_a and/or Δ_b could be well approximated by an infinitesimal 2D displacement field $\mathbf{t}_{ab} \in \mathbb{R}^{2\times 6}$ and the area of the polygon $\Delta_{\Pi,a\cap b} + \mathbf{t}_{ab}$ can therefore only change infinitesimally.

Next, we show that $G(\Delta_{\Pi,a\cap b})$ being L-c implies $\tilde{A}(\Delta_{\Pi,a\cap b})$ is L-c. Assume there exists two infinitesimal[1] rotations $\mathbf{R}_a$ and $\mathbf{R}_b$ such that the intersection of $\Pi_O(\mathbf{R}_a\Delta_a)$ and $\Pi_O(\mathbf{R}_b\Delta_b)$ is $\tilde{\Delta}_{\Pi,a\cap b}$. Thus, due to $G(\Delta_{\Pi,a\cap b})$ being L-c, there must exist a $K \geq 0$ such that:

$$\|G(\tilde{\Delta}_{\Pi,a\cap b}) - G(\Delta_{\Pi,a\cap b})\| \leq K\|\mathbf{t}_a - \mathbf{t}_b\|, \tag{5}$$

where $\mathbf{t}_a$ and $\mathbf{t}_b$ are the infinitesimal translations approximating the effect of $\mathbf{R}_a$ and $\mathbf{R}_b$ (resp.) on the projective plane. Now suppose the overlap flag for $\Delta_{\Pi,a\cap b}$ be $O_\Delta = 1$ and $\tilde{\Delta}_{\Pi,a\cap b}$ be $\tilde{O}_\Delta = 0$, i.e., the combined action of $\mathbf{R}_a$ and $\mathbf{R}_b$ caused $(\Delta_{\Pi,a} \cap \Delta_{\Pi,b})$ to become an empty set. Thus, we can say:

$$\begin{aligned} \|\tilde{A}(\tilde{\Delta}_{\Pi,a\cap b}) - \tilde{A}(\Delta_{\Pi,a\cap b})\| &= A(\Delta_{\Pi,a\cap b}) = G(\Delta_{\Pi,a\cap b}) \\ &= \|G(\tilde{\Delta}_{\Pi,a\cap b}) - G(\Delta_{\Pi,a\cap b})\| \leq K\|\mathbf{t}_a - \mathbf{t}_b\|, \end{aligned} \tag{6}$$

[1] W.r.t the Frobenius metric $\|\mathbf{R}_a^\top \mathbf{R}_b\|_{\mathrm{F}}$

thus proving $\tilde{A}$ to be L-c, i.e., intersection of two polygons with L-c-translating vertices has area that is L-c in the vertex positions. Finally, the total area of the combined projections of Δ_a and Δ_b, which could be anything between a triangle and a dodecagon (including the possibility of two non-overlapping triangles), is given by:

$$A^*(\Delta_{\Pi,a\cup b}) = A(\Delta_{\Pi,a}) + A(\Delta_{\Pi,b}) - \tilde{A}(\Delta_{\Pi,a\cap b}), \tag{7}$$

and the sum/difference of L-c functions are L-c (carefully disambiguate that $A^*(\Delta_{\Pi,a\cup b})$ being L-c does not imply its smoothness, especially at the point of overlap when O_Δ transitions from 0 to 1; nonetheless, the product of O_Δ and G in $\tilde{A}$ remains L-c).

Finite set of triangles. Next, we consider the case of a set of N triangles $\{\Delta_1, \ldots, \Delta_N\}$ with a fixed rigid orientation w.r.t. each other. Thus the combined area of the projections of $\{\Delta_1, \ldots, \Delta_N\}$ can be given by:

$$A^*(\Delta_{\Pi,\cup_{i=1}^N}) = \sum_{i=1}^{N} A(\Delta_i) + \tilde{A}(\Delta_{\Pi,\cap_{i=1}^N}) - \Big(\tilde{A}(\Delta_{\Pi,1\cap N}) + \sum_{i=1}^{N-1} \tilde{A}(\Delta_{\Pi,i\cap i+1})\Big), \tag{8}$$

where $A^*(\Delta_{\Pi,\cup_{i=1}^N})$ and $\tilde{A}(\Delta_{\Pi,\cap_{i=1}^N})$ denotes the area of union and intersection of the projection of all N triangles including all combinatorial possibilities. Importantly, $\tilde{A}(\Delta_{\Pi,\cap_{i=1}^N})$ is one or more polygon(s) with potentially large number of vertices, say N^*, but any infinitesimal rotation $\mathbf{R}$ applied to $\{\Delta_1, \ldots, \Delta_N\}$ can also be approximated by the displacement field $\mathbf{t}^* \in \mathbb{R}^{2\times N^*}$, thus $\tilde{A}(\Delta_{\Pi,\cap_{i=1}^N})$ is L-c w.r.t rotation of $\{\Delta_1, \ldots, \Delta_N\}$. Thus $A^*(\Delta_{\Pi,\cup_{i=1}^N})$ is L-c, completing the proof.

2. Proof of Lemma 1

The condition $\langle \mathbf{v}, (0,0,1)^\top \rangle = \langle \mathbf{v}_x, (0,0,1)^\top \rangle$ implies $\mathbf{v}$ and $\mathbf{v}_x$ are on the same latitude in $\mathcal{S}_\pi$, thus $F\big((\alpha, \mathbf{v}^\top)^\top\big)$ to $F\big((\alpha, \mathbf{v}_x^\top)^\top\big)$ represents an in-plane rotation of the orthographic projection of $\mathbf{Q}$ (since it is always projected to the XY-plane), and thus the AoS remains constant.

3. Proof of Theorem 2

There exists two feasible set in $\mathcal{S}_\pi$, $\mathcal{F}_A \subseteq \mathcal{S}_\pi$ and $\mathcal{F}_E \subseteq \mathcal{S}_\pi$, such that for all elements of $\mathcal{F}_A$, the Hausdorff distance based optimality condition:

$$H(\tilde{\mathbf{G}}, \mathbf{G}^*) \sim 0, \tag{9}$$

is obeyed and for all elements of $\mathcal{F}_E$, the area signature match is satisfied (i.e., the criterion $|A(\mathbf{G}^*) - A(\tilde{\mathbf{G}})| \leq \epsilon_{xy}$). Clearly, $\mathcal{F}_A \subseteq \mathcal{F}_E$. Assumption 1 with theorem 1 guarantees $\mathcal{F}_E, \mathcal{F}_A \neq \varnothing$. Due to lemma 1 and the sufficiency of $\mathcal{D}_\pi$, intersection of $A(\mathbf{G}^*)$ with $\mathcal{A}$ contains at least one point, i.e., $U_\mathcal{A} \neq \varnothing$. Consequently, the maximum separation of elements in $\tilde{C}$ from $\mathcal{F}_E$ is bounded by a function of sampling resolution and thresholds, i.e., $\propto \max(\epsilon_{xy}, \epsilon_z)$. Thus, under sufficient sampling, there must exist elements of $\tilde{C}$ for which, the distance to $\mathcal{F}_A$ is given by $\epsilon_o \propto \max(\epsilon_{xy}, \epsilon_z)$ (since $\mathcal{F}_A \subseteq \mathcal{F}_E$). Therefore $\lim_{\epsilon_{xy},\epsilon_z \to 0} \epsilon_o = 0$. Finally, we complete the proof asserting that $(\epsilon_{xy}, \epsilon_z)$ can be made arbitrarily small by increasing the sampling resolution of $\mathcal{A}$, as the points on the iso-contours are detected numerically. Thus, the proposed method recovers a globally optimal solution in the limit of infinite sampling density and vanishing thresholds.

4. Proof of Lemma 2

$\Pi : \mathbb{R}^3 \setminus \{Z = 0\} \mapsto \mathbb{R}^2$ is L-c (and C^∞) and composition of L-c functions are L-c, since the decomposition $\Pi = \Pi_\mathrm{O} \circ \tilde{\Pi}$ holds if $\tilde{\Pi}$ gives the homogeneous normalised image coordinates following perspective projection.

5. Algorithms

We present the algorithms for pre-computing $\mathcal{A}$ from any given semi-dense point cloud in algorithm 1. The same strategy applies to the pre-computation of $\mathscr{E}$ with elliptic aspect ratio instead of area, thus, not repeated separately. With such pre-computed $\mathcal{A}$ and $\mathscr{E}$, the pose estimation algorithm is given in algorithm 2; the given tunable parameter values correspond to the one used in our experiments.

Algorithm 1 Computation of $\mathcal{A}$ from Input $\mathbf{Q}$

Input: $\mathbf{Q} \in \mathbb{R}^{3\times M}$
Sample $\mathbf{D} \in \mathbb{R}^{2\times Q}$ such that $\mathbf{D} \subset \mathcal{D}_\pi$
for $q = 1$ to Q **do**
  $\mathbf{R}_q \leftarrow F\big(G(\mathbf{D}_q)\big)$
  $\tilde{\mathbf{G}} \leftarrow \tilde{S}(\mathbf{Q}, \mathbf{R}_q, \mathbf{0})$
  $\mathcal{A}(\mathbf{D}_q) \leftarrow A(\tilde{\mathbf{G}})$
end for
Output:$\mathcal{A}$

Algorithm 2 Estimating pose from silhouettes and $\mathcal{A}$

Initialize: $\epsilon_\cap^{(0)} \leftarrow 10^{-2}$, $\epsilon_z^{(0)} \leftarrow 2\times 10^{-2}$, $\epsilon_{xy} \leftarrow 10^{-2}$, $\epsilon_e \leftarrow 10^{-2}$, $\epsilon_H \leftarrow 10^{-2}$, $t \leftarrow 0$, $\alpha_{\min} \leftarrow \infty$, $\lambda_c \leftarrow 10^2$
Compute: $\mathbf{t}_{\text{opt}}$ with $\mathbf{t} = C(\tilde{S}(\mathbf{Q}, \mathbf{I}_3, \mathbf{0})) - C(\mathbf{G}^*)$
while $\alpha_{\min} \geq \epsilon_H^{(t)}$ **do**
  Compute: $U_{\mathcal{A}}$ with ϵ_{xy} such that eq. (3) is satisfied
  Compute: $U_{\mathscr{E}}$ with ϵ_e such that eq. (6) is satisfied
  Compute: $U_{\mathcal{A}\cap\mathscr{E}}$ with $\epsilon_\cap \leftarrow \epsilon_\cap^{(t)}$ such that eq. (7) is satisfied
  Compute: $\tilde{C}$ following eq. (5) with $U_{\mathcal{A}} \leftarrow U_{\mathcal{A}\cap\mathscr{E}}$ and $\epsilon_z \leftarrow \epsilon_z^{(t)}$
  if $|\tilde{C}| > \lambda_c$ **then**
    while $|\tilde{C}| > \lambda_c$ **do**
      Select $\hat{\mathbf{d}} \sim$ Uniform$(\tilde{C})$ and set $\tilde{C} \leftarrow \tilde{C} \setminus \{\hat{\mathbf{d}}\}$
    end while
  end if
  for all $\mathbf{R}_r \in \tilde{C}$ **do**
    $\tilde{\mathbf{G}} \leftarrow \tilde{S}(\mathbf{Q}, \mathbf{R}_r, \mathbf{t})$
    $\alpha_r \leftarrow H(\tilde{\mathbf{G}}, \mathbf{G}^*)$
  end for
  $\alpha_{\min} \leftarrow \min\{\alpha_r\}$, and $\mathbf{R}_{\text{opt}} \leftarrow \mathbf{R}_{r'}$ with r' such that $\alpha_{r'} = \alpha_{\min}$
  $\epsilon_\cap^{(t+1)} \leftarrow 2\epsilon_\cap^{(t)}$, $\epsilon_z^{(t+1)} \leftarrow \epsilon_z^{(t)} + 10^{-2}$
  $t \leftarrow t + 1$
end while
Output: globally optimal pose $(\mathbf{R}_{\text{opt}}, \mathbf{t}_{\text{opt}})$

6. Postel to $\mathbb{SO}(3)$ Mapping

A scalar-vector pair $(\alpha, \mathbf{v}^\top)^\top \in \mathcal{S}_\pi$ can be mapped to its equivalent rotation matrix as follows:

$$\mathrm{F}\big((\alpha, \mathbf{v}^\top)^\top\big) = \begin{bmatrix} \big(\mathbf{v}_2^2 + \mathbf{v}_3^2\big)\cos(\alpha) - \mathbf{v}_2^2 - \mathbf{v}_3^2 + 1 & -\sin(\alpha)\,\mathbf{v}_3 - \mathbf{v}_1\mathbf{v}_2\,(-1+\cos(\alpha)) & \sin(\alpha)\,\mathbf{v}_2 - \mathbf{v}_1\mathbf{v}_3\,(-1+\cos(\alpha)) \\ \sin(\alpha)\,\mathbf{v}_3 - \mathbf{v}_1\mathbf{v}_2\,(-1+\cos(\alpha)) & \big(\mathbf{v}_1^2 + \mathbf{v}_3^2\big)\cos(\alpha) - \mathbf{v}_1^2 - \mathbf{v}_3^2 + 1 & -\sin(\alpha)\,\mathbf{v}_1 - \mathbf{v}_2\mathbf{v}_3\,(-1+\cos(\alpha)) \\ -\sin(\alpha)\,\mathbf{v}_2 - \mathbf{v}_1\mathbf{v}_3\,(-1+\cos(\alpha)) & \sin(\alpha)\,\mathbf{v}_1 - \mathbf{v}_2\mathbf{v}_3\,(-1+\cos(\alpha)) & \big(\mathbf{v}_1^2 + \mathbf{v}_2^2\big)\cos(\alpha) - \mathbf{v}_1^2 - \mathbf{v}_2^2 + 1 \end{bmatrix} \tag{10}$$

7. Additional Experimental Results

We offer some of the additional experimental details and plots below, focussing on the ones that could not be included in the main article due to space restrictions.

7.1. Details of N*l*R, N*l*-P*a*R, and M*s*-GO

We offer below, the implementation details of Non-linear Refinement (N*l*R), Non-linear Project-and-Refine (N*l*-P*a*R), and Multi start - Global Optimization (M*s*-GO) that were used as an 'intuitive' baseline in section 4.1 of the main article. Recall that the key Pose-from-Silhouette (P*f*S) problem that we aim to solve (see eq. 2 of main article for notations) is given as:

$$\min_{\mathbf{R}\in\mathbb{SO}(3), \mathbf{t}\in\mathbf{R}^2} H(\tilde{\mathbf{G}}, \mathbf{G}^*), \quad \text{s.t.:} \quad \tilde{\mathbf{G}} = \tilde{S}(\mathbf{Q}, \mathbf{R}, \mathbf{t}). \tag{11}$$

The three methods N*l*R, N*l*-P*a*R, and M*s*-GO solves eq. (11) as described in the three next paragraphs.

N*l*R. In this approach, we map elements of $\mathbb{SO}(3)$ to its corresponding Lie algebra $\mathfrak{so}(3)$ via the logarithmic map and solve for:

$$\min_{\mathbf{r}\in\mathfrak{so}(3),\mathbf{t}\in\mathbf{R}^2} H(\tilde{\mathbf{G}},\mathbf{G}^*), \quad \text{s.t.:} \quad \tilde{\mathbf{G}} = \tilde{S}(\mathbf{Q},\exp(\mathbf{r}),\mathbf{t}). \tag{12}$$

where $\mathbf{r} \in \mathfrak{so}(3)$ is the Lie algebraic parametrization of rotation and the exponential map $\exp(\cdot)$ is via Rodrigue's rotation formula [5]. Equation (12) is solved using Levenberg-Marquardt (LM) initialized to unit pose.

N*l*-P*a*R. In this approach, we solve eq. (11) in two steps iteratively, the *first* step establishes correspondences between $\tilde{S}(\mathbf{Q},\mathbf{R}_l,\mathbf{t}_l)$ and $\mathbf{G}^*$ (where $\mathbf{R}_l,\mathbf{t}_l$ are the rotation and translation at l-th iteration), and the *second* step solves the problem:

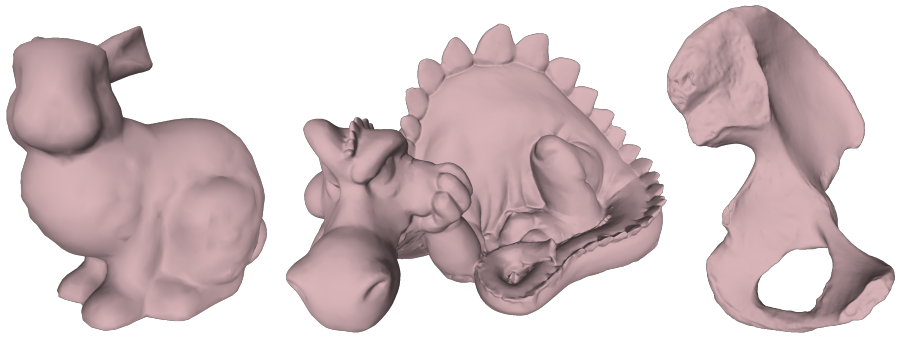
$$\begin{aligned}\min_{\mathbf{r}_l\in\mathfrak{se}(3),\mathbf{t}_l\in\mathbb{R}^2} \sum_{k'=1}^{|\mathbf{G}^*|} \|\mathbf{s}^*_{k'} - \tilde{\mathbf{s}}_{l,k''}\|_2^2, \qquad (k',k'') \in \Omega_l, \quad \mathbf{s}^*_{k'} \in \mathbf{G}^*, \quad \tilde{\mathbf{s}}_{l,k'} \in \tilde{\mathbf{G}}_l, \\ \text{s.t.:} \qquad \tilde{\mathbf{G}}_l = \tilde{S}(\mathbf{Q},\exp(\mathbf{r}_l),\mathbf{t}_l).\end{aligned} \tag{13}$$

where Ω_l is the correspondence between $\mathbf{G}^*$ and $\tilde{\mathbf{G}}_l$ established in the *first* step. Note that in eq. (13), we succeed in restricting the Hausdorff distance in eq. (11) to squared Euclidean distances solely due to the presence of correspondences from the *first* step. The iterations are initialised from unit pose till the difference of successive costs drop below 10^{-4}.

M*s*-GO. In this approach, eq. (11) is solved directly using the MATLAB `MultiStart` function which closely follows [11].

7.2. 3D Models for Orthographic Silhouettes

We show the 3 object models Stanford Bunny (SB), Phlegmatic Dragon (PD), and Pelvic Bone (PB) in fig. 1. SB and PD are genus-0 while PB is genus-1.



(a) (b) (c)

Figure 1. We show 3D meshes of the three 3D models: (a) SB, (b) PD, and (c) PB

7.3. Error Histograms for GlOptiPoS and GlOptiPoS+ on Orthographic Silhouettes

Error histograms for the three models SB, PD, and PB using Globally Optimal P*f*S (GlOptiPoS) and GlOptiPoS + Non-linear Refinement (GlOptiPoS+) are given in fig. 2.

7.4. Exemplar Noisy Orthographic Silhouettes

We show some examples of the noisy silhouettes of PB in fig. 3.

7.5. Effect of Depth Prior on Accuracy

The effect of varying depth prior on accuracy with the *Ape* and *AutoGPS* object models from *B*ino*c*ular *O*bject *T*racking (B*c*OT) dataset are shown in fig. 4.

7.6. Partial Baselines

We offer limited comparisons with two P*f*S methods that remain directly incomparable due to non-homogeneous input requirements and data assumptions.

Deep Active Contours. The Deep Active Contours (DAC) [12] framework is fundamentally different needing: I) texture along silhouettes, II) accurate pose initialisation, and III) temporal continuity; thus, DAC is a pose *tracking*, rather than a global pose *estimation* approach. We were unable to fully replicate the implementation beyond the author's provided test data despite contacting them. Silhouettes and estimated object diameters used by DAC were not provided, preventing comparison using ADD(-S) or Area Under Curve (AUC) metrics. We thus report indicative results based on the 5*cm*–5°(↑) and 2*cm*–2°(↑) metrics (from eq. 14, [12]): the authors report 94.0% and 65.5% (resp.) for DAC, while GlOptiPoS+ Perspective

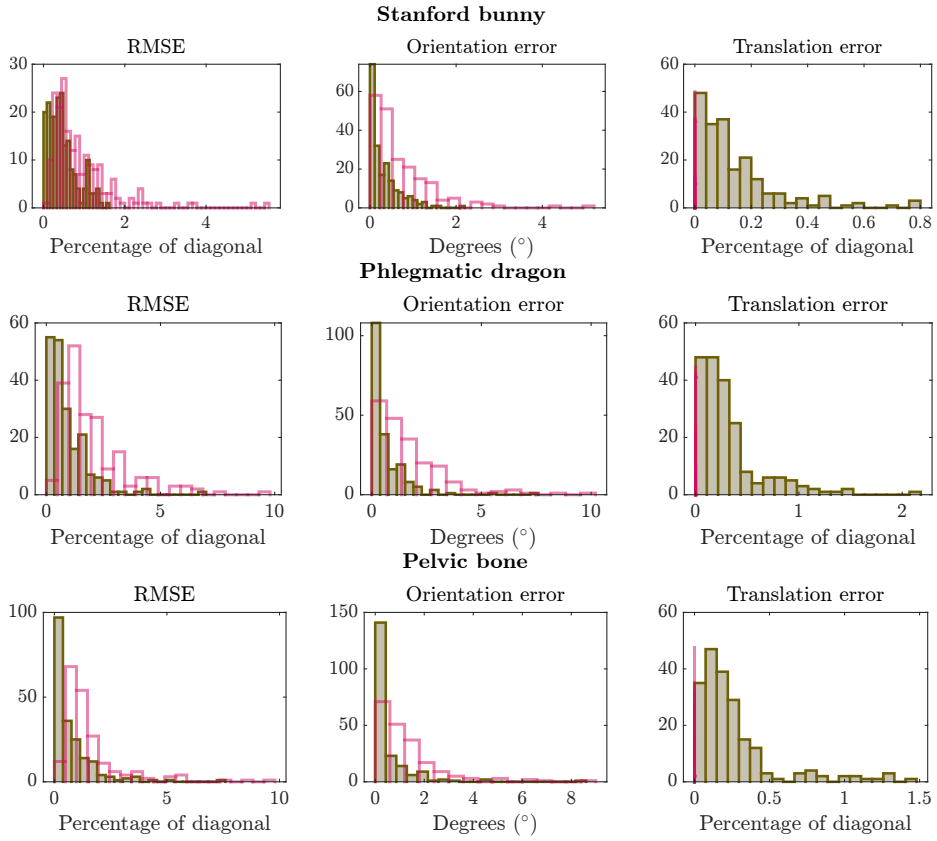


Figure 2. Root Mean Square Error (RMSE), Orientation Error (OE), and Translational Error (TE) for PD, SB, and PB; red-hollow bars are from GlOptiPoS, solid-filled bars are from GlOptiPoS+

(a) Low noise (b) Medium noise (c) High noise

Figure 3. Examples of the three noise settings used in the experiments in the paragraph 'effect of noisy silhouettes on accuracy' of the main article; we show here an example each of (a) *low*, (b) *medium*, and (c) *high* noise

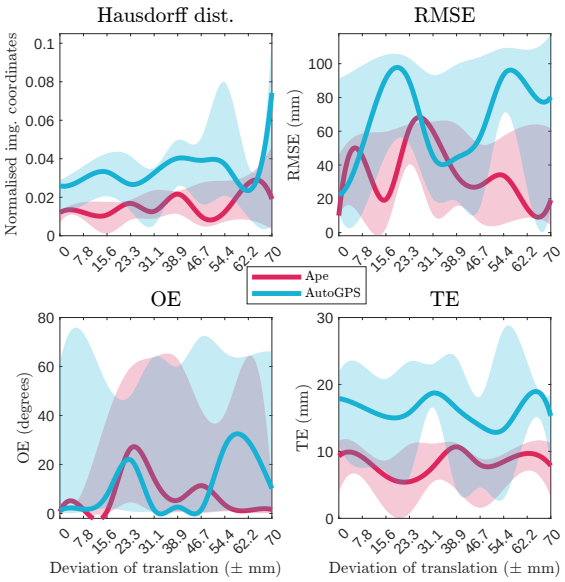


Figure 4. $H(\mathbf{G}_\Pi, \mathbf{G}^*_\Pi)$, RMSE, OE, and TE for *Ape* and *AutoGPS*; the shaded area spans the range of values, the central bold curve passes through the median values

(GlOptiPoS$_\Pi$+) achieves 93.8% and 88.8% under similar conditions (values should be interpreted as approximate indicators, experimental setups are potentially non-uniform).

Perspective-1-Ellipsoid. Perspective-1-Ellipsoid (P$_1$E) [6] estimates object/camera pose from silhouettes by fitting ellipses, yielding not a single solution but a continuous trajectory of feasible poses. An example on the *cat* object from the B*c*OT dataset is shown in fig. 5, illustrating these trajectories. In contrast, our approach (tables 2, 3, and 4 of the main paper) recovers a unique pose for *cat* with quantified accuracy. A uniform comparison is therefore non-trivial and of limited relevance given the differing problem formulations.

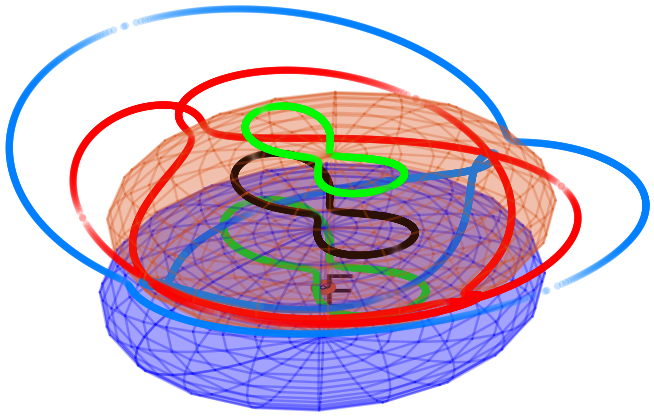


Figure 5. Results from P$_1$E on *cat* object of B*c*OT dataset: the blue ellipsoid shows the ellipsoid fitted to the *cat* object, the black curve shows the family of solutions to the camera centre recovered by P$_1$E (c.f., our method always necessarily recovers a point in space); refer [6] for details

7.7. Memory Requirements

The memory requirements for the storing the learned shape signatures (PARS and PEARS combined) is trivial in modern computing standards, e.g.: all three shapes of SB, PD, and PB requires storage space of $\sim$1 MB and all objects of B*c*OT requires storage space of $<$2 MB at their experimental resolution – thus, a detailed discussion about memory usage scalability is uninteresting.

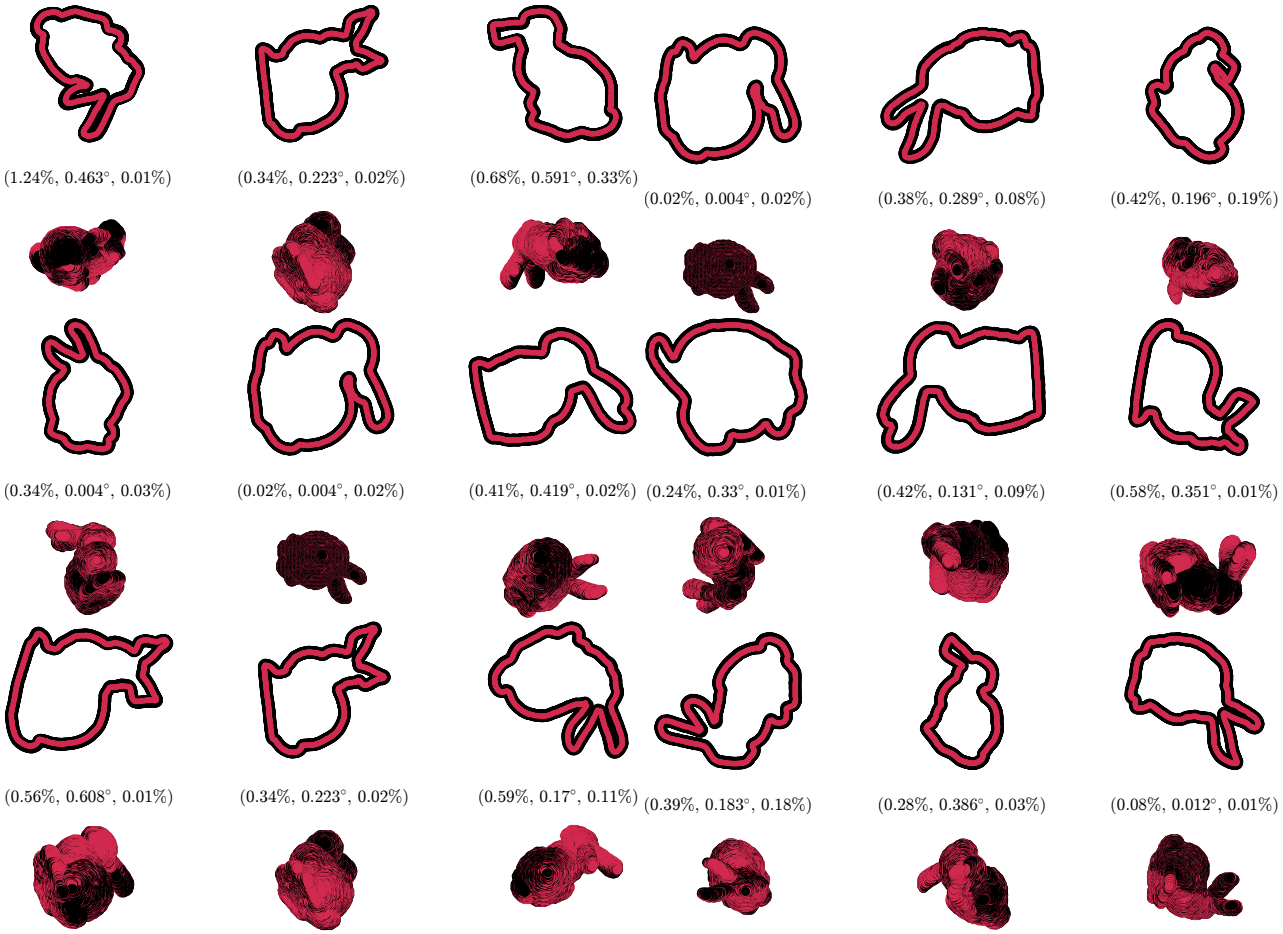


Figure 6. Qualitative results of SB. 1st, 3rd, and 5th row denotes Groundtruth (Gt) (black) and estimated (red) silhouettes; 2nd, 4th, and 6th row denotes Gt (black) and estimated (red) object pose in $\mathbb{R}^3$. Each silhouette image corresponds to the template pose shown below it, with accuracy metrics displayed in between. There is a strong overlap between Gt and estimated points/curves, confirming the very high accuracy of GlOptiPoS+

7.8. Additional Qualitative Results

We show many randomly sampled qualitative results for the three test objects SB, PD, and PB in fig. 6 fig. 7, and fig. 8 (resp.), corresponding to the results in table-I of the main article. We show many randomly sampled qualitative results for the assymmetric objects in BcOT dataset on the *complex_movable_handheld* sequence (which was chosen since it is one of the most challenging sequences in the dataset) in fig. 9, corresponding to the results in section 4.2 of the main article.

7.9. Ambiguities in Symmetric Shapes

Figure 10a shows results from ambiguous silhouettes, including constant projected area (sphere) and rotational symmetries (other spherical harmonics), highlighting camera poses which are the non-unique optima for the PfS problem.

7.10. Empirical Analysis of Hausdorff Distance and Area Difference between Silhouettes

Between silhouettes, small Hausdorff distance implies small difference in AoS but *not* vice versa. Effect of sampling/discretization density and segmentation noise on Hausdorff distance and AoS is consistent with intuition – increasing sampling density causes both Hausdorff distance and AoS to decay while increasing noise causes both the metric to grow; see fig. 10b for an example study on Squeezed Balloon (SB).

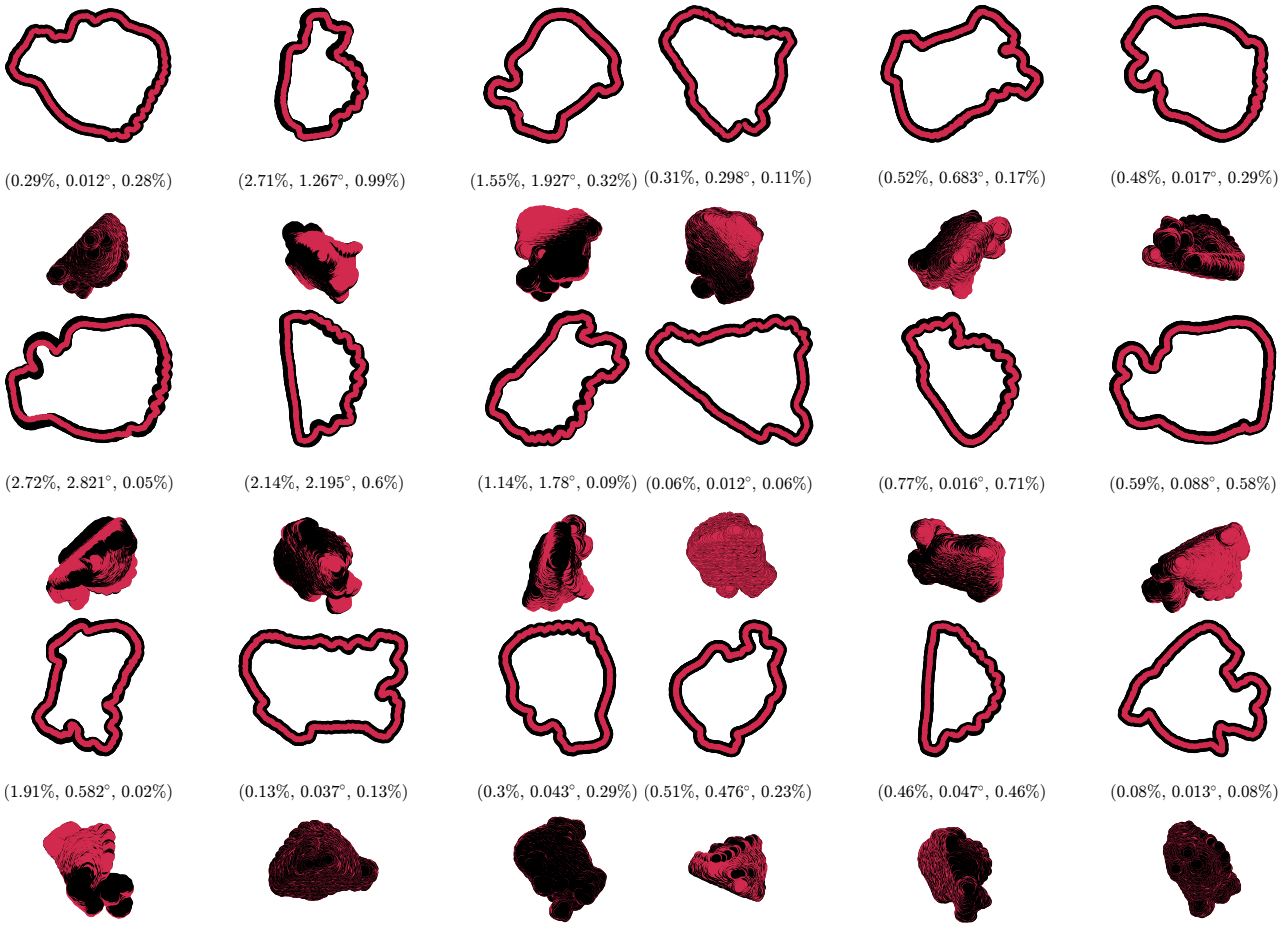


Figure 7. Qualitative results of PD. Same layout and colour scheme as fig. 6

7.11. Examples with Thin Shell Objects

Figure 10c shows an example of GlOptiPoS+ from an admissible thin shell object, a paper from the Bramante39M [1, 10] sequence. The histograms show the error metrics for the unambiguous cases while the 3D plots on the left show some examples of ambiguous silhouettes in such thin shell objects.

8. Further Details and Clarifications

The following notes offer concise clarifications on areas that may benefit from expanded detail.

8.1. Scope of the Method

The proposed method in our article deals with only the P*f*S problem, meaning recovering the 5-Degrees-of-Freedom (D*o*F) pose from orthographic silhouettes (3 from rotation in $\mathbb{SO}(3)$ and 2 for translation in $\mathbb{R}^2$ – for orthographic silhouettes, translation along Z-axis is irrecoverable) and 6-D*o*F pose from perspective silhouettes (3 from rotation in $\mathbb{SO}(3)$ and 3 for translation in $\mathbb{R}^3$). Our method **does not** 'reconstruct' shapes from silhouettes, which are usually in the scope of the similarly-named but very different problem-statement of Shape-from-Silhouette (S*f*S) [2, 3, 7]. We request careful disambiguation between the P*f*S and S*f*S problems. S*f*S typically reconstructs a shape from one or multiple views of its silhouettes while the proposed P*f*S method determines the 5-D*o*F/6-D*o*F pose of a known shape from a *single input silhouette*. Extension of our method to multi-view fusion is trivial: poses from independent views can simply be averaged, therefore remains uninteresting.

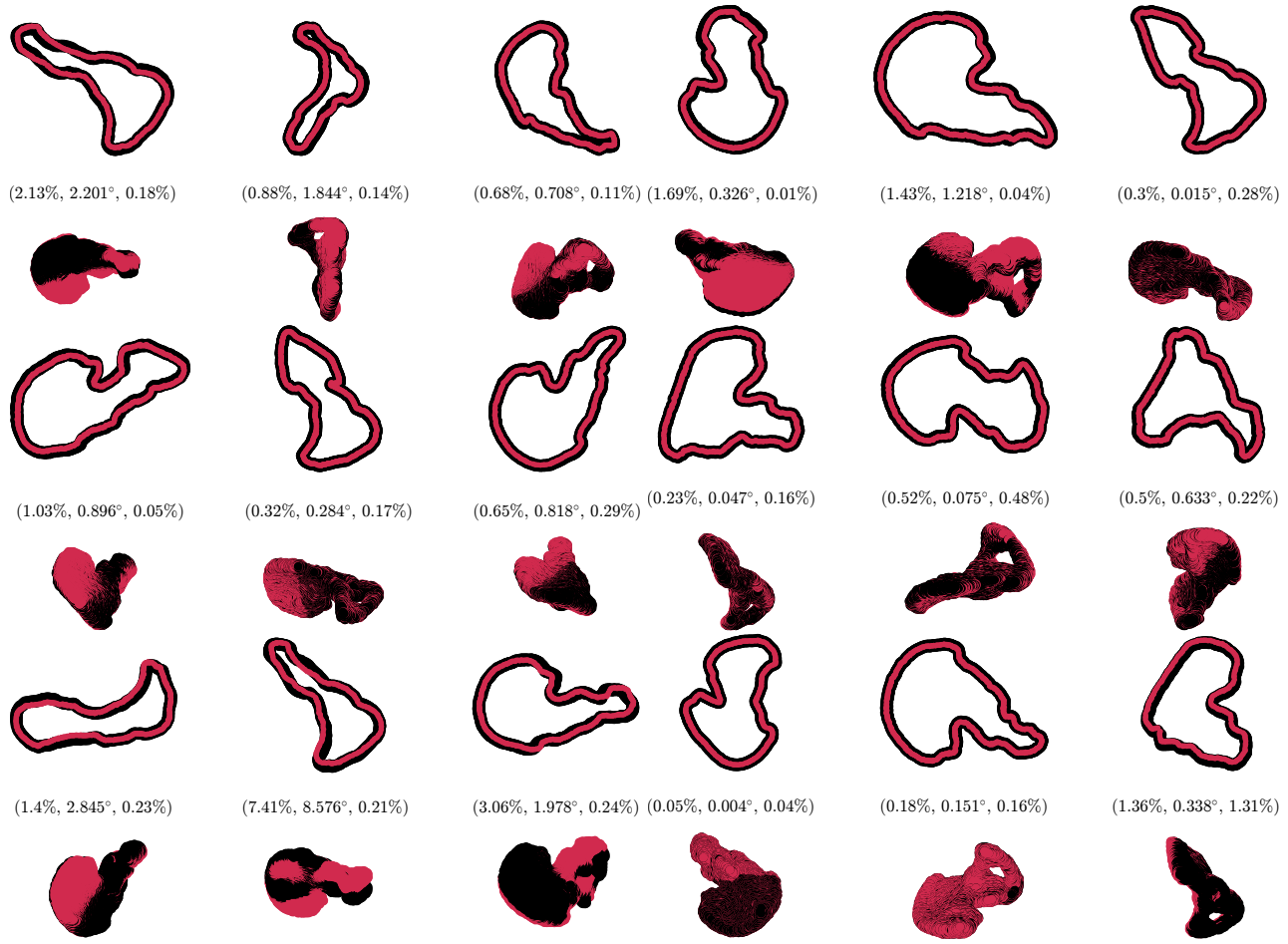


Figure 8. Qualitative results of PB. Same layout and colour scheme as fig. 6

8.2. Lack of Global Optimality in Perspective Projection

We explain below, some of the details regarding global optimality with GlOptiPoS Perspective (GlOptiPoS$_\Pi$) and GlOptiPoS$_\Pi$+.

Why is GlOptiPoS$_\Pi$ not globally optimal? Perspective projection causes A*o*S to vary with translation, unlike orthographic projections, where A*o*S is invariant to translation. Thus, unlike the orthographic case, where the learned shape signature (A*o*S) is $\mathbb{R}^2 \mapsto \mathbb{R}^+$ (i.e., the Postel disc to A*o*S), the minimal search space for the perspective case would have been $\mathbb{R}^5 \mapsto \mathbb{R}^+$ (i.e., 2 D*o*F for rotation and 3 D*o*F for translation to A*o*S). Therefore, although global optimality with perspective projection is 'possible' in theory, this is intractable in practice because Branch-and-Bound (B*n*B)-like approaches must systematically partition the feasible set and evaluate cost on each region; its worst-case complexity grows exponentially with the problem dimension [9].

Real-world usage of GlOptiPoS$_\Pi$, despite lacking global optimality. Indeed, due to the reasons described in the previous paragraph, GlOptiPoS$_\Pi$ lacks guaranteed global optimality; but nonetheless, through our experiments, we demonstrated the following: I) our proposed method is practically implementable and usable in challenging real-world scenarios, as demonstrated on B*c*OT dataset, and II) our proposed method significantly outperforms the compared baseline method, thus demonstrating an advancement of the state-of-the-art in even the perspective case. Moreover, to our knowledge, none of the existing state-of-the-art methods offer any hint of global optimality in perspective based P*f*S (recall: [4] is a stochastic approach, it cannot guarantee convergence to the global optimum and it can and does often get trapped in locally optimal solutions).

8.3. Deformable Objects

All the methods we propose in this article are inapplicable to deformable objects, object rigidity is an underlying assumption all throughout. The proposed method can in theory be extended to deformable objects by increasing dimensionality of the search space, but suffers from exponentially worsening cost bottleneck (analogous to the effect of introducing translation to the search space, but additional D*o*Fs arising out of object deformations could be significantly higher than 3). Interestingly, an older method [8] appears to have achieved positive results on articulated objects (or human shapes) using a comparable approach but was obviously limited to local solutions following an approximate initialization.

8.4. Occlusion

Although occlusion-aware P*f*S is appealing, pose recovery from occluded silhouettes is ill-posed and often unsolvable; no existing method (including ours) provides a principled solution in this regime. Accordingly, we do not view occlusion handling as a natural future direction of our core approach. A more viable direction is to pair our method with modern segmentation or silhouette-completion models capable of inferring missing or occluded contour regions prior to pose estimation.

8.5. Mesh Resolution

In all our experiments, the mesh resolution is fixed and selected empirically at a level that guarantees accurate numerical extraction of the object's boundary upon projection. The resolutions used throughout the paper correspond to the maximum practical density we observed to be necessary for stable silhouette computation across all tested shapes; increasing resolution further produced no measurable improvement in accuracy or consistency. Consequently, the wall-clock runtimes reported in Section 4.1 of main paper can be treated as representative absolutes rather than lower bounds dependent on mesh refinement. Overall, mesh resolution is not a sensitive or limiting factor for our method, and plays no meaningful role in its scalability.

References

[1] Adrien Bartoli and Agniva Sengupta. Camera pose in sft and nrsfm under isometric and weaker deformation models. *Computer Vision and Image Understanding*, page 104488, 2025.

[2] Thomas J Cashman and Andrew W Fitzgibbon. What shape are dolphins? building 3d morphable models from 2d images. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):232–244, 2012.

[3] Kong-Man Cheung, Simon Baker, and Takeo Kanade. Shape-from-silhouette across time part i: Theory and algorithms. *International Journal of Computer Vision*, 62(3):221–247, 2005.

[4] Xiao Cui, Nan Li, Chi Zhang, Qian Zhang, Wei Feng, and Liang Wan. Silhouette-based 6d object pose estimation. In *International Conference on Computational Visual Media*, pages 157–179. Springer, 2024.

[5] Karin Erdmann and Mark J Wildon. *Introduction to Lie algebras*, volume 122. Springer, 2006.

[6] Vincent Gaudillière, Gilles Simon, and Marie-Odile Berger. Perspective-1-ellipsoid: Formulation, analysis and solutions of the camera pose estimation problem from one ellipse-ellipsoid correspondence. *International Journal of Computer Vision*, 131(9):2446–2470, 2023.

[7] Gloria Haro. Shape from silhouette consensus. *Pattern Recognition*, 45(9):3231–3244, 2012.

[8] N R Howe. Silhouette lookup for automatic pose tracking. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 15–22. IEEE, 2004.

[9] Sebastian Nowozin, Christoph H Lampert, et al. Structured learning and prediction in computer vision. *Foundations and Trends® in Computer Graphics and Vision*, 6(3–4):185–365, 2011.

[10] Agniva Sengupta and Adrien Bartoli. Convex solutions to sft and nrsfm under algebraic deformation models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[11] Zsolt Ugray, Leon Lasdon, John Plummer, Fred Glover, James Kelly, and Rafael Martí. Scatter search and local nlp solvers: A multistart framework for global optimization. *INFORMS Journal on computing*, 19(3):328–340, 2007.

[12] Long Wang, Shen Yan, Jianan Zhen, Yu Liu, Maojun Zhang, Guofeng Zhang, and Xiaowei Zhou. Deep active contours for real-time 6-dof object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14034–14044, 2023.

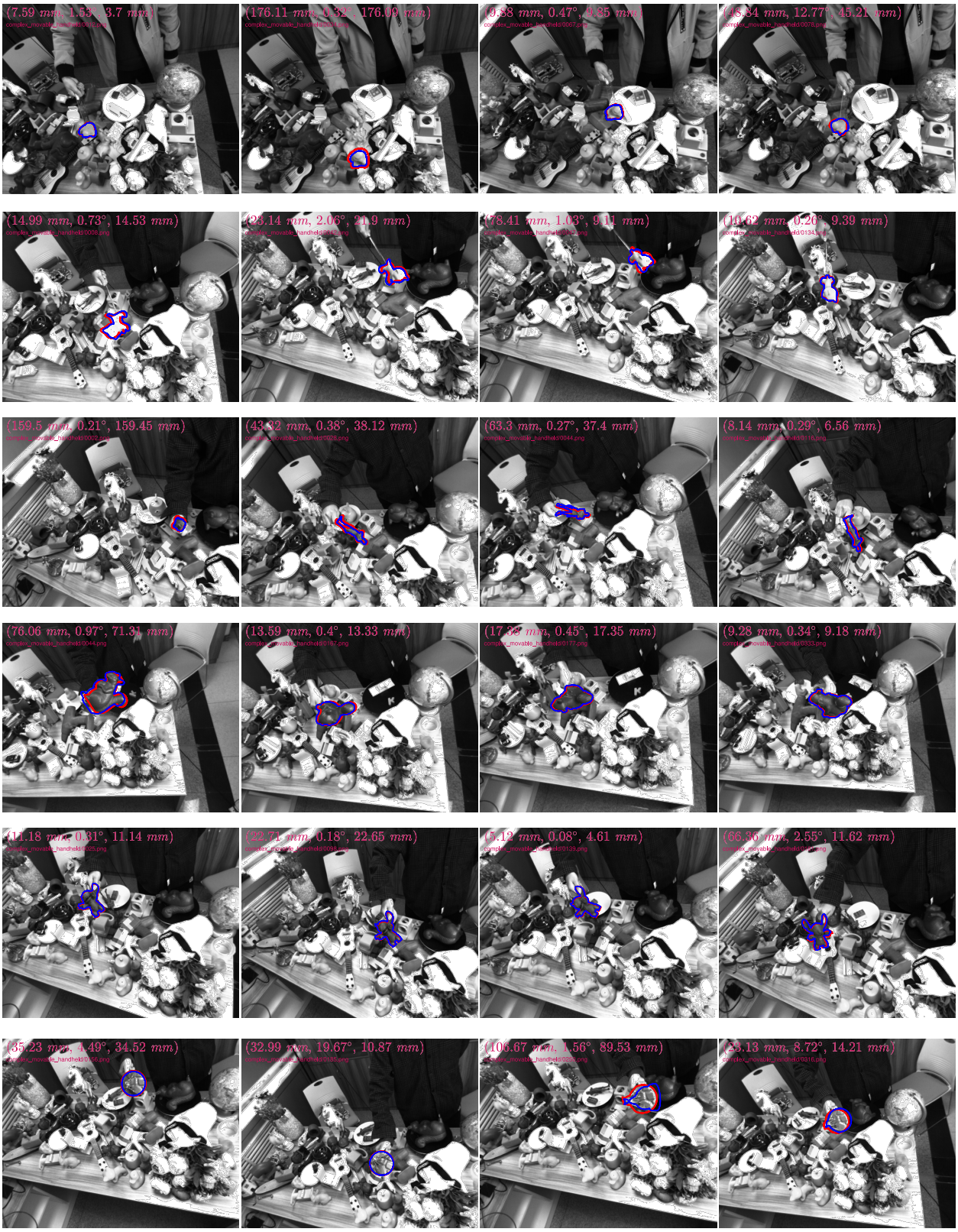


Figure 9. Qualitative results from the '*complex_movable_handheld*' sequence of the B*c*OT dataset, shown for the asymmetric objects – the six rows correspond to *Ape*, *Cat*, *Jack*, *Squirrel*, *Stitch*, and *Vampire Queen* (resp.) – across four randomly sampled image frames for each object; the triplet of values shown in the top-left of images correspond to (RMSE, OE, TE) in units (*mm*, degrees, *mm*). The **red** curves correspond to the G*t* and input silhouette while the **blue** curve corresponds to the silhouette obtained from GlOptiPoS$_{\Pi}$+

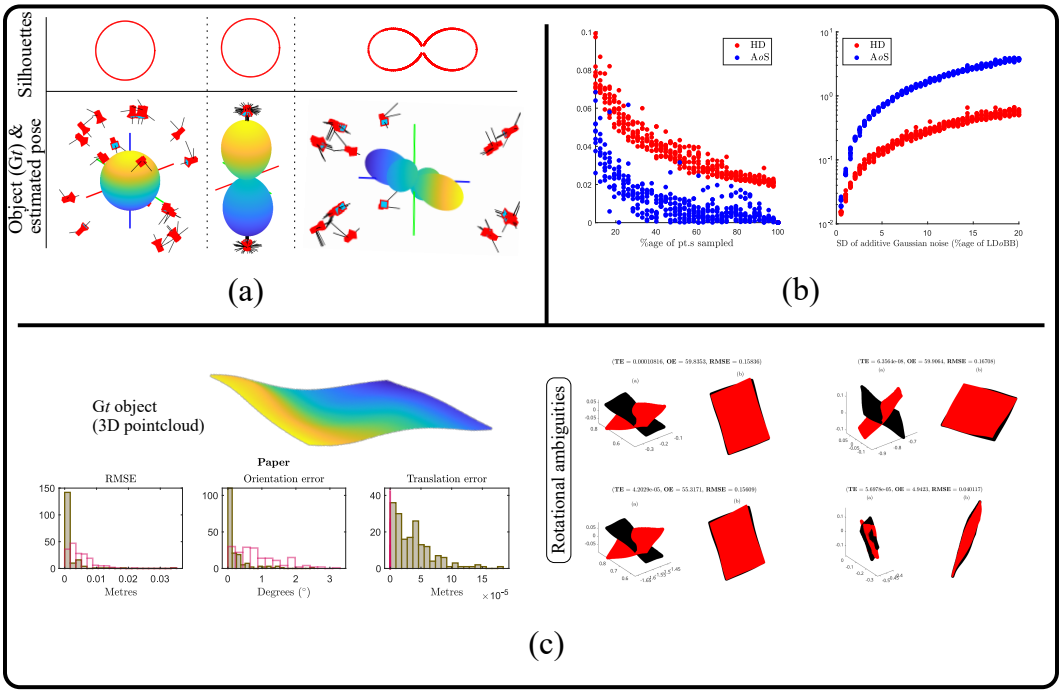


Figure 10. (a) Camera pose from GlOptiPoS+ for three symmetric objects; shown as red virtual cameras with black axes. The shapes are 3 of the spherical harmonics; depending on the silhouette, the estimated pose shows varying patterns of ambiguity over repeated experiments, (b) variation of Hausdorff distance (HD) and A*o*S with pointcloud density (left) and segmentation noise on silhouettes (right) for SB; both parameters improve with increased density and worsens with noise, as expected intuitively, and (c) experiments on thin shell object - a reconstructed, deformed sheet of paper. *Top left*: shows the 3D shape of the paper; *bottom left*: shows the histogram of RMSE, OE, and TE for 200 experiments (colour scheme follows fig. 2) – excluding cases of rotational ambiguity; *right*: shows four examples of rotational ambiguity due to the shape of the paper - the left sub-plots show 3D G*t* (black) and estimated (red) pose, right subplots show the projected pointclouds (same colour scheme).