Multiview Retrieval of Referent Geometry from Articulated Silhouettes

Inês Valcorvo
Department of Visual Computing
Louriaçal Institute of Technology
Braga
Portugal
i.valcorvo@louriacal.example

Hajime Senokura
Division of Image Science
Hokuryo College of Engineering
Niigata
Japan

Amal Bensarif
Centre for Image and Performance Studies
Institut des Techniques d'Arveil
Lyon
France

Lotte van Wessem
Department of Visual Computing
Louriaçal Institute of Technology
Braga
Portugal

capstract:2610.00012v1 [cs.CV] 30 Oct 2026

***Abstract*—A silhouette typically constrains the geometry of the object that intercepts the light. When an articulated surrogate is used to depict an object, the cast shadow is evidence about two geometries: the occluder and its intended referent. We formulate referent retrieval for a calibrated screen and a catalog of three-dimensional animal models. Forty participants produced 2,880 hand-shadow poses while viewing prescribed orthographic projections of the models. A dual encoder aligns pose silhouettes and model geometry without requiring point correspondences between the two. Under a joint performer and model holdout, three instructed poses and three light positions yield exact retrieval in 111 of 144 sets, compared with 70 of 144 for a contour matcher (28.5 percentage points difference, 95% crossed-bootstrap interval 17.4 to 40.3). Oblique illumination decreases direct contour overlap but improves retrieval even when the instructed view is held fixed. An asymmetry in the separation of the fingers that represent ears helps distinguish two rabbit models whose central shadows are frequently interchangeable. The distinction has no counterpart in a reconstruction of the occluding volume.**

***Index Terms*—shape retrieval, cast shadows, silhouette matching, multiview vision**

I. Introduction

A calibrated silhouette supplies an occupancy constraint. Intersecting the viewing cones from several silhouettes gives the visual hull [1], [2]. The same construction applies to a cast shadow on a calibrated screen if the cone apex is placed at the light instead of the camera.

A shadow intended to depict a second object presents a different inverse problem. Consider an archive that acquires three views of a rabbit silhouette. The silhouettes describe a pair of human hands. They were made at the request of a curator who showed the performer a rabbit figurine. The figurine is missing from the acquisition and the original manifest is incomplete. A sufficiently dense visual hull could recover the approximate length of the person’s thumb. It would assign two ears to the thumb and index finger only through an additional model of what the performer was trying to show.

We ask whether a change in lighting can improve the retrieval of a referent. Two extended fingers can project to one extremum under the central lamp and separate when the lamp is displaced. The resulting shadow may be a worse match to the reference picture but reveal which rabbit the performer was shown. The camera motion that would ordinarily expose the back of the nearer ear cannot be applied to a rabbit that was never at the screen.

In our protocol a performer receives the same animal model at three specified orientations. The system receives the resulting binary shadows, the light positions, and the prescribed yaws. It retrieves one model from a catalog. Section III describes the holdout and the coordinate systems. Section IV compares learned retrieval with direct matching and uses prints of the referents to measure changes due to illumination alone.

II. Related Work

A. Silhouettes and cast shadows

Octree carving [3] and image-based rendering [2] use different representations of the same intersection. We use the planar transfer of Hartley and Zisserman [4] to map recorded silhouettes onto the screen. The print and hand conditions use the same observation equation.

Recordings of layered cutouts show that increasing light separation exposes appendage boundaries that overlap centrally [5]. Haseq and Takamori [6] fitted joint limits to photographs of performers following published hand diagrams. Several inequivalent pairs of wrists reproduced one rabbit diagram. We consulted the diagrams to select categories for which we could elicit a pose without demonstrating one.

B. Shape recognition and hand representations

Moment invariants [7] and shape contexts [8] provide baselines for direct contour retrieval. A hand tracker [9] or a para-

metric model such as MANO [10] could provide an anatomical bottleneck with different observations. Our input consists of silhouettes, so we compare with the occupancy of the hands estimated by cone intersection.

The model uses residual image features [11], point-set pooling [12], and a contrastive objective [13]. Corvelet and Grevlin [14] used repeated drawings to search a collection of paper cutouts. Ngari and Duprinel [15] reported systematic variation in hand posture across prescribed profiles. We distinguish changing the requested profile from relighting a fixed pose.

III. Acquisition

A. Referential stimulus

The catalog contains 60 closed meshes, ten each of bird, camel, deer, dog, rabbit, and turtle. Two appendage layouts per class recur across assets built from independent base meshes. The rabbit layouts differ in the connection between the ears. Variants sharing a base mesh stay in the same split. Each mesh stands on the same ground plane and is scaled to unit height. We render it orthographically at yaw $v \in \{0^\circ, 45^\circ, 90^\circ\}$ relative to its anatomical front, with elevation zero. The display omits class and model labels.

We adapted the elicitation protocol of Silveira Tomal and Nishihama [16]. Participants used either or both hands to depict the animal in the presented view, feet toward the lower edge. They saw their central shadow and the reference picture on adjacent displays and could revise their pose for 20 s. We recorded the pose at the limit regardless of resemblance. For the physical control we used opaque prints of the complete meshes.

B. Screen geometry and illumination

Let H be the volume occupied by the hands and let y be a coordinate on the receiving screen, transformed to the corresponding world point $Y(y)$. At lamp position ℓ the binary image has a simple geometric form under point illumination. Writing L_ℓ for the center of emission, the foreground condition is

$$S_\ell(y; H) = \mathbf{1}\{\exists\, t \in (0,1) \colon L_\ell + t[Y(y) - L_\ell] \in H\}. \tag{1}$$

The corresponding catalog mask $R_v(M)$ renders mesh M at yaw v. Eq. (1) describes the hands regardless of the animal represented. We preserve the screen up direction so that the image encodes which appendages were above their neighbors.

Five lights lay at 0°, $\pm 15^\circ$, and $\pm 30^\circ$ from the screen normal measured at the center of the performance region. A global-shutter camera captured one exposure per lamp in a 40 ms burst. We rectified the frames to 512 pixels per 480 mm of screen by checkerboard calibration [17]. Foreground lies below 0.55 of each pixel’s brightness in the corresponding blank exposure. We discarded components of 20 pixels or less and masked the fixed entry strip below the performance region. For the learned model we downsampled the calibrated region to 128 pixels square. Its position is fixed from setup, independent of which exposures the model receives.

C. Sampling and holdout

Forty adults at one site consented to release of masks and stimulus IDs. Twenty produced the training data with 36 meshes. Eight took part in development with 12 other meshes, and 12 in testing with the final 12. Every performer saw every mesh in their split at all three yaws. We randomized the order within performer and balanced the number of poses per yaw across two visits. Test performers did not know how many targets of each class were in the bank.

Table I counts complete captures. One bout comprises five exposures for one request yaw. A retrieval set groups three bouts for the same performer and target. We repeated a burst only after a lost hardware trigger. All sets entered the evaluation, including ones in which the recorded shadows resembled the stimulus poorly.

TABLE I
Acquisition splits. Each set contains three bouts and each bout contains five images, one per light.

Split	Performers	Meshes	Sets	Bouts	Images
Training	20	36	720	2,160	10,800
Development	8	12	96	288	1,440
Test	12	12	144	432	2,160
Total	40	60	960	2,880	14,400

IV. Retrieval Method

A shared image encoder maps each mask to 128 features. A fusion network takes these features and the coordinates $\sin \ell$ and $\cos \ell$ of its light, averages over observed positions, and produces a unit vector q_v. The mesh encoder takes 2,048 fixed surface samples and rotates them by yaw v around the catalog up axis before producing a unit vector $g_v(M)$. It sees no filename or class label. Aligning point clouds to their principal axes would introduce possible reflections relative to the annotated front.

Writing B for the set of observed instructed yaws and $\mathcal{C}$ for the candidate catalog, the retrieval rule is

$$\hat{M} = \arg\max_{M \in \mathcal{C}} \frac{1}{|B|} \sum_{v \in B} q_v^{\top} g_v(M). \tag{2}$$

Training maximizes the log softmax of the per-bout similarity over all 36 candidate mesh codes at the same yaw. The temperature is 0.10. A shadow with a curve for the top of a shell and no foreground below it is a positive example for the full turtle if the prompt specified that mesh. The loss does not depend on the area of agreement between the shadow and the referent.

We sample lamp subsets uniformly from six configurations during training: central alone, either oblique pair, central plus either pair, and all five. In a diagnostic condition we exchange the signed azimuth tags while leaving the image masks untouched. Appendix A gives the schedule.

A. Comparison methods

The direct matcher compares the central shadow at each requested yaw with the orthographic silhouette of each candidate at the same yaw. It computes shape-context distance on 96 samples of the exterior contour following Belongie et al. [8]. Inner holes do not contribute samples. Each contour is normalized by its mean interpoint distance. The cost combines a one-to-one minimum distance assignment and the intersection-over-union deficit after aligning vertical extent and foreground centroid. The development set selected a weight of 0.40 for the deficit in the sum of these two costs, after scaling each by its median training distance. A second matcher replaces the shape-context distance with Euclidean distance between signed log Hu moments [7]. Both use the same pooling across instructed views as Eq. (2), with distance in place of similarity.

The hull matcher carves a calibrated 96-cubed grid in each bout using the central and two outer masks and Eq. (1). It samples the surface of the resulting volume and compares it to the catalog point clouds using symmetric Chamfer distance. We align centroids and standardize height while preserving yaw and screen up. The learned central-only score uses the same checkpoint as the full model.

B. Outcomes and uncertainty

The primary outcome is top-1 retrieval with all three instructed views and the central and $\pm 30^{\circ}$ masks. The primary comparison is with contour matching, selected on development accuracy. Ties use a fixed catalog order. We also score class accuracy and symmetric mean nearest-point distance between predicted and true meshes, using fixed samples of 2,048 points per surface and normalizing by reference height. Exact retrieval gives zero distance, since the two point sets are identical.

We draw 12 performer rows and 12 mesh columns with replacement, using the same indices for paired comparisons [18], [19]. Each selected set contributes all relevant yaws. The primary interval uses the 2.5th and 97.5th percentiles from 10,000 draws. With two models per class the interval depends on the range of shapes in the bank. For the descriptive illumination series we plot 1.96 bootstrap standard errors on either side of each point estimate.

V. Results

A. Joint transfer

The full model retrieved 111 of 144 referents, 41 more than contour matching (28.5 percentage points, 95% interval 17.4 to 40.3). Table II reports class accuracy and the distance between predicted and true surfaces. The hand hull saw the oblique images yet gave the largest surface error of the methods tested.

TABLE II
Retrieval of held-out referents. Both count columns have denominator 144. Surface distance is in percent of reference height.

Method	Exact mesh	Correct class	Mean surface distance (%)
Hu moments	59	116	7.9
Contour matching	70	125	6.1
Hand visual hull	64	110	8.6
Learned, central only	91	132	3.8
Learned, $\pm 30^{\circ}$ only	105	134	2.6
Learned, central and $\pm 30^{\circ}$	**111**	**135**	**2.2**

Of the 33 full-model errors, 24 stayed within class. Deer had 17 correct matches, the fewest of the six classes. Rabbits and dogs each had 20 of 24. Fig. 1 compares exact retrieval by class. Six of the contour matcher’s 10 rabbit errors were within the category, compared with three of the four full-model errors. One performer accounted for two of the three.

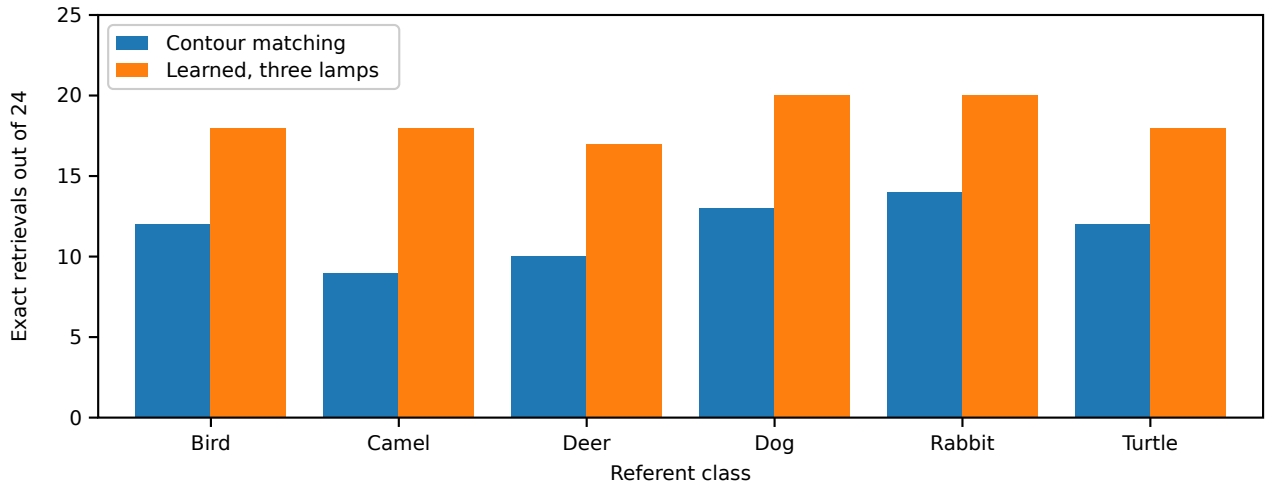


Fig. 1. Exact retrieval counts by target class, out of 24 reference sets per class.

B. *Illumination and depicted view*

Fig. 2 shows the effect of additional illumination. With the central mask present, adding the 15° pair recovered 104 targets, and adding the 30° pair recovered 111. Using the center alone recovered 91. For mask overlap we compare each signed light with the same requested view, after alignment by foreground height and centroid. At 30° the two lights averaged 0.44 intersection over union, versus 0.63 at the center.

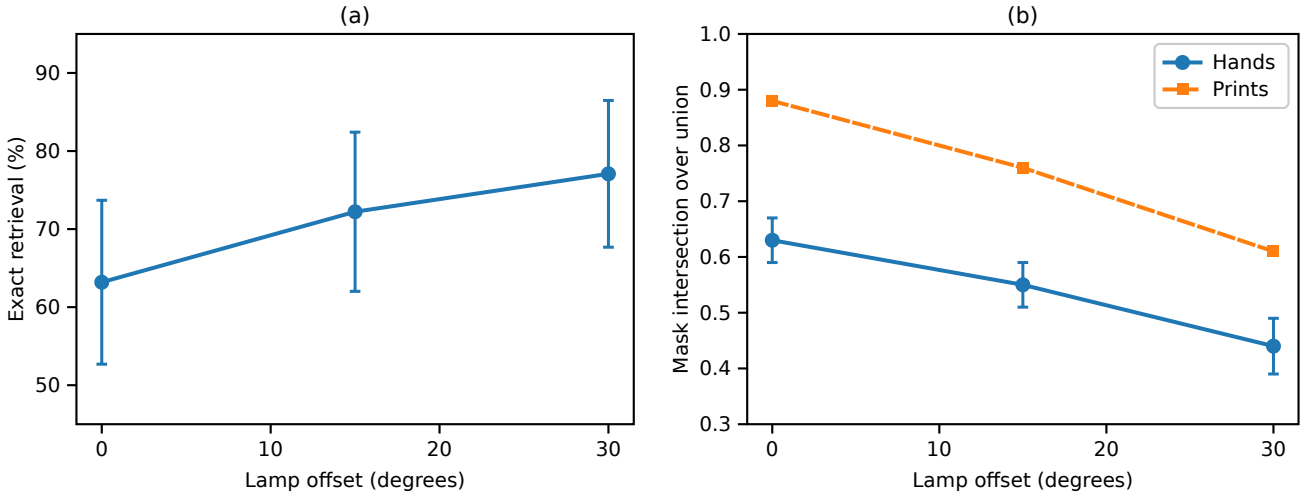


Fig. 2. Effect of illumination offset. Left, exact retrieval with the central lamp and both lamps at the given offset. At zero only the central lamp is used. Right, mean silhouette overlap for the individual views at the plotted offsets, with both sides pooled. Bars are 1.96 crossed-bootstrap standard errors in each direction.

The print control included 48 sets, four mountings of each test mesh. Contour matching retrieved 46 of 48 and the learned weights retrieved 36. Off-axis print masks also lost overlap with the prescribed view. The camera and lighting alone can therefore account for an overlay degradation. The controls had no effect on checkpoint selection.

Fig. 3 crosses requested yaw and illumination. At a single requested yaw of 45°, the central lamp recovered 71 targets and the full lamp set recovered 90. The all-yaw counts are 91 and 111. Swapping the two signed lamp tags reduced the latter to 99. An encoding of foreground area and aspect ratio alone recovered 52 targets.

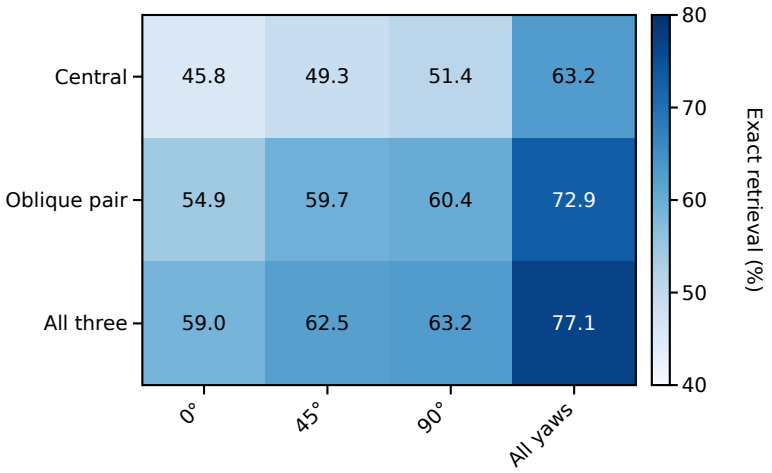


Fig. 3. Exact retrieval rates as percentages of the 144 held-out sets. Columns indicate the available instructed views, and rows indicate the light positions. The checkpoint is shared across rows.

C. Ear separation

The bank contains two rabbits at the same normalized height. Their normal projections at 90° overlap at 0.93 after alignment. The larger difference is at 45°, where the model with externally split ears has a second visible point near the crown. For 5 of the 12 performers shown that model at 90°, the two longest extensions in the central silhouette share one extremum. This occurred for 7 of 12 performers for the model with partially joined ears. The outcome is therefore a weak indication of which mesh was shown. At $\pm 30^\circ$, the first 5 sets included four with distinct extensions clear on at least one side. Among the other 7 sets, two had that property. We measured this descriptive effect on all relevant trials and retained the ones with merged outlines in the primary evaluation.

Two raters traced visible extensions to a common base. They were unaware of target identity and machine prediction. Both had to see two pixels of background between the extensions at some height for separation to count. The raters disagreed on one oblique view each in two sets, neither among the 12 sets with a merged central projection. All sets remained in the retrieval evaluation.

In one of the four cases in which the two tips became visible, the central silhouette ends at the tip of the longer finger. Under the left lamp the shorter one appears several pixels below it. Under the right lamp they overlap and the combined extension is wider. Replacing the oblique exposures with copies of the center, keeping the lamp tags, moved the intended model from first to second for this case. The second tip belonged to the performer's middle finger.

Table III gives results for the merged outlines. The separation is an observation made after target assignment. The counts cannot isolate its causal role in retrieval.

TABLE III
RABBIT RETRIEVAL AT THE 90° REQUEST AMONG SETS WITH ONE EXTREMUM IN THE CENTRAL IMAGE. THE SEPARATION COUNTS USE THE TWO OBLIQUE LAMPS.

Referent ears	Central merge	Separated off axis	Correct, central	Correct, three lamps
Split	5	4	2	4
Joined at base	7	2	4	5
Total	12	6	6	9

VI. DISCUSSION

An archive interested in the quality of the depiction might discard frames in which an extra limb emerges behind the shadow. Such frames can supply evidence about the referent even when they misrepresent its outline. Two finger arrangements can also yield nearly identical masks under a narrow angular baseline. In those cases a more accurate recovery of the occluder's projection may leave the referent equally uncertain.

Moving the lamp samples one posture. Changing the request introduces an independently chosen posture, possibly using different fingers for the same part. Combining both kinds of views as cone constraints would treat these successive postures as one solid. A record of lamp yaw alone would lose the protocol variable that distinguishes them.

The closed bank offers two choices per class and no absent-target outcome. Picking a different mesh of the right class can leave a small surface error while giving a false identification. We used a flat matte screen at one site. Larger lamps could broaden penumbrae and erase the gaps in Table III. For archival recordings without a prompt, the system would also need to estimate the requested yaw.

A learned distribution over shapes could extend the method to unknown referents. Some surfaces never enter the elicited

views. In our bank a rabbit’s underside is returned by retrieval even though we never asked a participant to depict it. An experiment that requested the ventral surface would first have to establish whether different performers can depict it with comparable postures.

Appendix A. Calibration and Optimization

Each session used 16 calibration board positions. For every lamp we fitted a source center from the shadows of eight spheres at known locations. Screen reprojection residual was 0.74 mm root mean square. On a held-back ninth sphere position the largest error was 1.6 mm. The carving comparison dilated foreground by one rectified pixel to allow for a penumbral edge. At two pixels it retrieved 63 references, one fewer than at one pixel.

Reference silhouettes used a 512-pixel square orthographic raster. The learned branch drew its surface point sets once per mesh and rotated them by yaw. The surface-distance metric used separate fixed samples, shared across methods.

We use a width-32 ResNet-18 with one input channel and a point encoder with layers of width 64, 128, and 256 followed by max pooling and a 128-dimensional output. Fusion uses a single 128-unit hidden layer. Adam [20] runs for 80 epochs with 48 bouts per batch, for 3,600 updates. The initial learning rate is 0.0003 and drops by a factor of ten after epochs 50 and 70. Augmentation translates each burst by up to four pixels and scales it isotropically by a factor between 0.95 and 1.05.

We selected the first epoch with the highest exact count on the 96 development sets. The primary run used seed 271 and epoch 73. Runs with seeds 419 and 613 recovered 109 and 113 test meshes. We excluded seed variation from the resampling analysis. Training took 4.6 GPU-hours in total on one RTX A5000. Median inference was 17 ms per set with cached mesh embeddings, excluding capture.

Appendix B. Accounting for Paired Trials

Both methods recovered 63 targets. The full model alone recovered 48, and the contour matcher alone recovered seven. Neither recovered the other 26. We resample the same row and column indices for both methods. Repeated indices count according to multiplicity. A resample can therefore contain fewer than six classes.

Removing the competing same-class item from the full catalog for each query raised the exact count from 111 to 129. Eighteen of the 24 within-class errors moved to the true mesh and six moved outside the class. The procedure uses the true class to select which item to remove. All primary results retain both models per class.

Acknowledgments

We thank the participants for returning for the second session and the archive staff at Arveil for preparing the reference meshes.

References

[1] A. Laurentini, “The Visual Hull Concept for Silhouette-Based Image Understanding,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 2, pp. 150–162, 1994.

[2] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan, “Image-Based Visual Hulls,” in *Proc. ACM SIGGRAPH*, 2000, pp. 369–374.

[3] R. Szeliski, “Rapid Octree Construction from Image Sequences,” *CVGIP: Image Understanding*, vol. 58, no. 1, pp. 23–32, 1993.

[4] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[5] I. A. Fieret and N. Bakhar, “Lateral Illumination and Boundary Recovery in Layered Silhouette Assemblies,” *Journal of Optical Image Records*, vol. 13, pp. 42–57, 2019.

[6] A. Haseq and E. Takamori, “Anatomical Consistency in Diagrams of Manual Shadow Figures,” *Transactions on Performance Documentation*, vol. 6, pp. 91–110, 2021.

[7] M.-K. Hu, “Visual Pattern Recognition by Moment Invariants,” *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[8] S. Belongie, J. Malik, and J. Puzicha, “Shape Matching and Object Recognition Using Shape Contexts,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[9] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, “Hand Keypoint Detection in Single Images Using Multiview Bootstrapping,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1145–1153.

[10] J. Romero, D. Tzionas, and M. J. Black, “Embodied Hands: Modeling and Capturing Hands and Bodies Together,” *ACM Transactions on Graphics*, vol. 36, no. 6, 2017.

[11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[12] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 652–660.

[13] A. van den Oord, Y. Li, and O. Vinyals, “Representation Learning with Contrastive Predictive Coding,” arXiv:1807.03748, 2018.

[14] N. Corvelet and A. M. Grevlin, “Part-Aware Search in a Collection of Cut-Paper Silhouettes,” in *Proc. Conference on Image Collections and Cultural Records*, 2020, pp. 81–89.

[15] T. W. Ngari and D. Duprinel, “Posture Variation Under Rotation of a Depicted Animal,” *Transactions on Performance Documentation*, vol. 8, pp. 135–151, 2023.

[16] V. Silveira Tomal and M. Nishihama, “A Protocol for Eliciting Silhouette Renditions from Orthographic Reference Images,” Louriaçal Institute of Technology, Tech. Rep. VC-22-04, 2022.

[17] Z. Zhang, “A Flexible New Technique for Camera Calibration,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[18] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

[19] A. B. Owen, “The Pigeonhole Bootstrap,” *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 386–411, 2007.

[20] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” in *Proc. International Conference on Learning Representations*, 2015.