BioMTBee: Biologically Constrained Multi-View Template-Based 3D Reconstruction of Bumblebee

Yi Zhang[1], Minchen Ye[2] †, and Nenggan Zheng[3,4] ⋆

[1] Qiushi Academy for Advanced Studies (QAAS) and the College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310027, China
12521076@zju.edu.cn
[2] XtalPi Holdings Limited, Shanghai 201315, China
22251056@zju.edu.cn
[3] Qiushi Academy for Advanced Studies (QAAS) and the State Key Laboratory of Brain-Machine Intelligence, Zhejiang University, Hangzhou, Zhejiang 310027, China
[4] School of Computer and Information Engineering, Bengbu University, Bengbu, 233030, China zng@cs.zju.edu.cn

Abstract. Bumblebees play a key role in agriculture, ecology, and bio-inspired robotics, making accurate 3D surface reconstruction essential for quantitative behavioral analysis. However, their dark dense hairs, complex structure and flexible limbs prone to visual occlusion and annotation difficulties, resulting in difficulties in reconstruction. Current template-free or generic template-based models offer limited structural detail and biological realism, while high-quality bumblebee templates and reliable morphological priors are lacking. To address these challenges, we propose **BioMTBee**, a template-based 3D mesh reconstruction framework for bumblebees. We built a high-fidelity articulated mesh template from micro-CT scans to provide an accurate morphological prior. A multi-view 3D pose estimator with spatio-temporal filtering (BPST) is then introduced to extract robust 3D keypoints from 2D detections, guiding stable mesh template fitting. Building on this template, we combine 3D pose and shape supervision with biological constraints—bilateral symmetry, kinematic coupling, and temporal smoothness—to achieve anatomically consistent and temporally stable surface reconstructions. Experiments on 3D reconstruction from multi-view images show improvements in morphological accuracy and biological plausibility over representative baselines, capturing fast limb movements and fine structural and textural details, and also generalize well to cross-species reconstruction.

Keywords: 3D mesh reconstruction · Template-Based Modeling · Multi-View Vision

1 Introduction

Bumblebees are vital pollinators with wide significance in agriculture, ecology, and bio-inspired robotics. Their high pollination specificity and adaptability to

† Equal contribution.
⋆ Corresponding author.

flowering phenology have led to widespread use in greenhouse cultivation, fruit and vegetable production, and other crop systems, generating substantial economic benefits [47]. In natural ecosystems, they play a key role in maintaining biodiversity and ecological stability [7]. Inspired by their adaptive foraging strategies and sensorimotor coordination, bumblebees have emerged as model organisms to study perception–action coupling and advance bio-inspired intelligence [17, 34, 39, 41]. A detailed understanding of bumblebees motion and morphology can support research into their behavior, benefiting agriculture, ecosystems, bio-inspired artificial intelligence and robotics, and may offer new insights into occurrence, mechanisms and evolution of certain behavioral traits. [11].

Traditional animal motion capture approaches focus mainly on coarse-grained behavior classification [38, 42] or keypoint-based pose estimation [12, 36], which represents motion using sparse anatomical landmarks. Although these methods are effective for high-level behavioral categorization and analysis, their sparse spatial representations are insufficient to capture continuous surface deformation and subtle morphological variations [33, 52]. These limited coarse-grained classification or pose information hinder fine-grained motion analysis and realistic morphological reconstruction of bumblebees, which are essential to advance research on sensorimotor coordination and bio-inspired intelligence.

To capture more detailed animal motion, 3D mesh reconstruction, particularly non-rigid modeling, has become a key approach, as accurate surface recovery can reveal body movements and mesh deformations which are essential for behavioral and morphological analysis [49]. Recent template-free frameworks [14, 18, 50, 53, 54] and statistical shape templates [4, 24, 33, 52, 57, 59] have advanced the reconstruction of general animal geometry, but most rely on extensive ground-truth data that are difficult for insects to obtain, where their small size, complex structure, and frequent nonrigid deformations [26, 27] make 3D insect reconstruction highly challenging. Although Doan *et al.* [8] propose a scanner-based system for high-quality 3D reconstruction of insect specimens, their method is designed for static subjects and relies on camera motion to acquire dense multi-view observations, which limits its applicability in scenarios involving moving insects. Moreover, dense hair, highly articulated joints, and complex motions of bumblebees induce severe self-occlusions, posing challenges for photogrammetric and neural reconstruction methods [4, 29, 48].

High-fidelity 3D reconstruction of bumblebees remains particularly challenging due to their unique visual and biological characteristics: (1) *Visual Occlusion*: dense dark hairs and frequent self-occlusions severely degrade the reliability of image-based cues for recovering fine geometric details; (2) *Ground-Truth Acquisition Difficulty*: the miniature body size and highly flexible limbs make it extremely difficult to obtain accurate ground-truth annotations for supervising 3D mesh reconstruction, especially during motion; and (3) *Prior Deficiency*: the complex body structure requires anatomically precise articulated templates and strong biological constraints, yet such morphology-aware priors are largely lacking, while template-free or generic template-based methods often fail to preserve structural details and maintain alignment with the real object.

Considering these challenges, this paper introduces **BioMTBee**, a multi-view, template-based 3D mesh reconstruction framework for bumblebee surface capture in dynamic settings. To remedy the lack of reliable articulated priors, we construct a high-fidelity, CT-derived bumblebee template with explicit kinematics, providing an anatomically faithful basis for nonrigid surface recovery. To overcome the difficulty of ground-truth acquisition for bumblebees, we introduce a multi-view 3D pose estimator with spatiotemporal filtering (BPST), which robustly infers 3D keypoints from 2D detections and provides stable geometric guidance for mesh fitting under noise and occlusion. Based on the mesh template, we formulate a differentiable rendering–based optimization that jointly estimates pose, shape, and appearance while enforcing kinematic coupling, bilateral symmetry, and temporal smoothness, thus ensuring biologically plausible deformations despite frequent self-occlusions and rapid limbs movements. Experiments demonstrate that our work outperforms baselines in surface accuracy and biological consistency, reliably capturing fast limb dynamics and fine structural details for precise motion analysis in a practical 3D behavioral observation system. Moreover, consistent gains on a public mouse dataset indicate its cross-species generalization capability for accurate 3D mesh reconstruction.

2 Related Work

Template-free reconstruction. Template-free methods mainly rely on image supervision to directly learn 3D shape and motion from images or videos. They reconstruct deformable meshes by minimizing pixel, silhouette, or optical-flow losses [14, 18, 28], while dense features and motion cues are exploited to capture temporal dynamics [50,53,54]. Recent advances also explore alternative representations, such as segmentation-guided 3D modeling [46] and neural point-based rendering [20], to improve geometric fidelity and universality. These approaches have shown potential in cross-species generalization, unsupervised video learning, and animation generation [15, 27, 44, 55]. However, the absence of explicit structural or geometric priors often leads to coarse shapes and blurred textures [26]. Multi-view and synthetic-view constraints improve geometric consistency [3, 13], and dedicated multi-view scanners have been developed for static insect-scale modeling [8]. Nevertheless, for bumblebees with dense dark hairs and highly articulated limbs, frequent self-occlusion and limited visual cues make reliable geometric recovery under motion particularly challenging.

Reconstruction based on generic statistical shape templates. Researches such as SMAL [59], SMALST [57], hSMAL [24], and related works [4,58] reconstruct animal morphology by learning low-dimensional shape and pose representations from parameterized statistical models. VAREN [60]further improves hSMAL with a parametric 3D model of horses learned from real data. Extensions such as CSM [23] and Animal3D [52] primarily improve robustness across different viewing angles and poses, while AniMer [33] enables generalization within predefined animal families under the SMAL model. However, most previous work focuses on vertebrates, and limited research on insects—especially

bumblebees—has resulted in a scarcity of high-quality datasets, restricting the applicability of template-based reconstruction methods.

Reconstruction based on specific templates. To achieve more realistic reconstructions, recent studies have built animal-specific templates. ArMo [5] reconstructs mouse motion from multi-view depth data; MAMMAL [1] models multi-animal pig motion via temporal mesh optimization; Bolanos *et al*. [6] construct CT-derived virtual mouse models for biomechanical simulation; and Badger *et al*. [2] reconstruct bird morphology from single-view images. Although such templates improve accuracy, some rely on toys or handmade models and dense annotations, limiting anatomical fidelity and scalability. For small, highly flexible bumblebees, reliable ground-truth supervision is difficult to obtain, while their many kinematic degrees of freedom make generic optimization unstable and prone to implausible deformations (e.g., joint inversion or body collapse). This motivates the need for a high-resolution 3D template with multi-view supervision and biological constraints for realistic dynamic bumblebee reconstruction.

3 Method

Given synchronized multi-view videos, we reconstruct temporally consistent, high-fidelity bumblebee surfaces (Fig. 1). The task is challenging due to small body size, severe self-occlusion, and rapidly moving limbs. To obtain anatomically grounded geometry, we first build an articulated bumblebee mesh template from micro-CT. To provide supervision for deformation and recover motion under occlusion, we propose BPST to estimate 3D pose keypoints robust to self-occlusion, with masks extracted by HQ-SAM [19]. We then introduce learnable nonrigid deformation parameters to adapt the template to individuals. Finally, we formulate losses for data fidelity and biological plausibility and optimize all variables via differentiable rendering.

3.1 CT-based Articulated Mesh Template

Reconstructing articulated bumblebees without realistic priors is highly underconstrained. We therefore used computed tomography (CT) to scan a bumblebee specimen and constructed a 3D mesh template with micrometer-level precision.

Template construction. CT volumes were reconstructed into a dense surface mesh using standard biomedical image processing and Marching Cubes [31, 43]. The mesh was subsequently cleaned, simplified, and retopologized to obtain a stable template with $N = 4{,}070$ vertices and 8,136 faces, preserving major anatomical structures while removing fine-scale surface noise.

Articulated kinematic modeling. To enable skeleton-driven deformation, we implant a hierarchical bone system with $B = 63$ control bones organized as a thorax-rooted kinematic tree. Mesh deformation follows linear blend skinning with a weight matrix $\mathbf{W} \in \mathbb{R}^{N \times B}$.

Keypoint correspondence. We define $K = 41$ anatomical consistency keypoints (in Fig. 3) and establish a vertex-to-keypoint mapping matrix $\mathbf{M} \in \mathbb{R}^{K \times N}$, linking mesh vertices to 3D pose supervision during reconstruction.

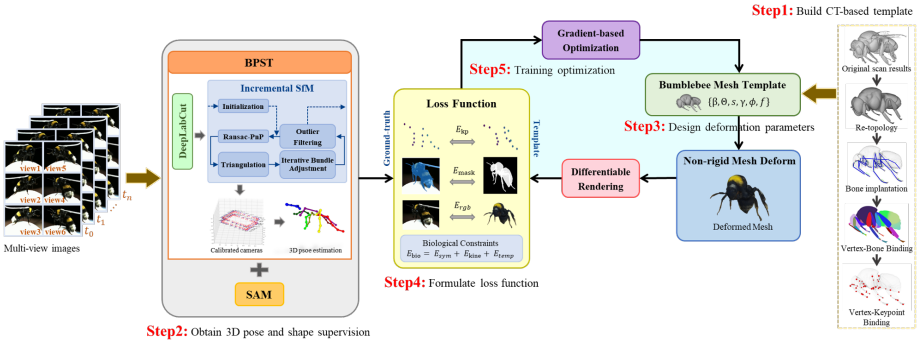

Fig. 1: Method overview. We design a biologically grounded reconstruction pipeline: (1) build an articulated micro-CT bumblebee mesh template for anatomical consistency; (2) estimate robust 3D pose keypoints via BPST with HQ-SAM masks to guide deformation; (3) learn nonrigid deformations to adapt the template to individuals; (4) enforce data fidelity and biological plausibility via loss design; (5) optimize all variables through differentiable rendering.

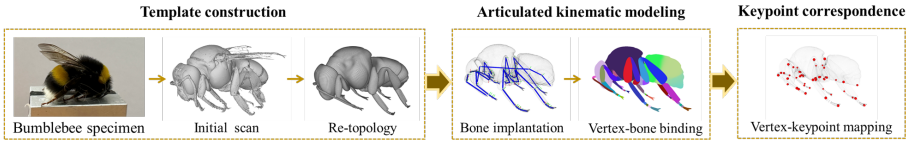

Fig. 2: The process of building the 3D bumblebee mesh model. After scanning a cryo-preserved bumblebee specimen, we obtain an initial mesh, retopologize it, implant a hierarchical bone system, and then bind mesh vertices both to bones and keypoints.

3.2 3D Pose Estimation and Shape Supervision

Reliable pose and shape supervision are crucial for accurate 3D mesh reconstruction. However, precise camera calibration for millimeter-scale settings is difficult, since slight shifts between calibration and recording can introduce projection errors, limiting robustness across different setups and experimental subjects. Existing pose estimation methods often fail on bumblebees due to their highly flexible limbs, causing false or missed detections during rapid movement. To address these issues, we adopt automatic calibration and propose a multi-view 3D pose estimator with spatiotemporal filtering (**BPST**) to provide stable multi-view pose supervision for dynamic reconstruction. Per-frame segmentation masks are further extracted to enforce shape silhouette consistency.

A DeepLabCut model [35] trained on our dataset (see Sec. 4.1) predicts 2D keypoints in each view, which are aggregated across frames for automatic multi-view calibration via incremental structure-from-motion (SfM) [40]. Starting from stereo initialization, additional views are incorporated using RANSAC–PnP [25] to estimate camera extrinsics by minimizing reprojection errors with outlier rejection, followed by bundle adjustment [56] for global refinement and accurate

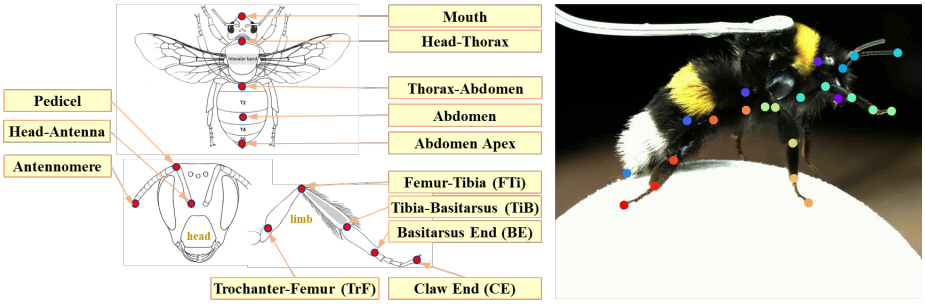

Fig. 3: Definition of bumblebee anatomical keypoints and annotation examples (two antennae and six limbs).The structural diagram is referenced from Dolan *et al.* [9].

triangulation of 3D keypoints. To mitigate self-occlusion and noisy detections during motion, we optimize the following objective:

$$L = L_{\text{rpj}} + \lambda_t L_t + \lambda_s L_s, \tag{1}$$

where L_{rpj} denotes a robust reprojection loss, L_t enforces temporal smoothness across frames, and L_s preserves segment length consistency to maintain anatomical proportions. Detailed calculation process of Eq. (1), including the definitions of each loss term L_{rpj}, L_t, L_s, are provided in the supplementary material.

For shape supervision, silhouettes are extracted from each view using HQ-SAM [19], with 2D keypoints serving as prompts. The resulting temporally stable 3D keypoints and view-consistent masks provide strong geometric constraints for subsequent nonrigid mesh reconstruction.

3.3 Mesh Template Deformation Parameterization

To fit the CT-based template to real bumblebee motion sequences, we introduce a set of learnable parameters $\{\beta, \Theta, s, \phi, \gamma, f\}$ to capture body state, articulated motion, and appearance. Let the kinematic tree contain B bones and the mesh consist of N vertices.

The *shape parameter* $\beta \in \mathbb{R}^{B\times 3}$ controls anisotropic scaling along the local axes of each bone, enabling part-wise deformation of mesh regions bound to anatomical structures and capturing individual morphological variations beyond the template.

The *pose parameters* $\Theta = \{\theta^{(t)} \in \mathbb{R}^{B\times 3}\}_{t=1}^{T}$ represent per-frame relative bone rotations in axis–angle form within the kinematic hierarchy. Global similarity transformation is modeled by scale $s \in \mathbb{R}^3$, rotation $\phi \in \mathbb{R}^3$, and translation $\gamma \in \mathbb{R}^3$. The per-vertex RGB appearance is denoted as $f \in \mathbb{R}^{N\times 3}$.

Let $\bar{\mathbf{V}}_i \in \mathbb{R}^{N\times 3}$ denote the i-th vertex in the bind pose. As defined in Sec. 3.1, W_{ib} is its normalized skinning weight satisfying $\sum_{b=1}^{B} W_{ib} = 1$. The deformed vertex position is computed as follows (see supplementary material for details):

$$\mathbf{V}_i = s\,\mathbf{R}(\phi)\left(\sum_{b=1}^{B} W_{ib}\,\mathbf{T}_b\,\mathbf{A}_b^{-1}\begin{bmatrix}\bar{\mathbf{V}}_i\\1\end{bmatrix}\right)_{1:3} + \gamma, \tag{2}$$

where $\mathbf{R}(\phi) \in \mathrm{SO}(3)$ is the global rotation. $\mathbf{A}_b \in \mathrm{SE}(3)$ is the bind-pose transformation of bone b. $\mathbf{T}_b \in \mathrm{SE}(3)$ is its transformation at the current pose.

3.4 Loss Function Formulation

We optimize the learnable parameters defined in Sec. 3.3 by minimizing the total loss $E(\beta, \Theta, s, \phi, \gamma, f)$ using gradient descent. The objective contains two terms:

$$E = E_{\text{data}} + E_{\text{bio}}, \tag{3}$$

where E_{data} denotes the loss of data fidelity and E_{bio} incorporates biological characteristics to achieve a realistic state.

Data loss. The data loss term incorporates the supervision of pose keypoints, shape silhouettes, and color appearance to optimize the alignment of pose, shape, and color between the template and the target object.

$$E_{\text{data}} = \lambda_{\text{kp}} E_{\text{kp}} + \lambda_{\text{mask}} E_{\text{mask}} + \lambda_{\text{rgb}} E_{\text{rgb}}. \tag{4}$$

Keypoint loss: $E_{\text{kp}} = \frac{1}{TK}\sum_{t=1}^{T}\sum_{k=1}^{K} \nu_k^{(t)} \left\| \hat{X}_k^{(t)} - X_k^{(t)} \right\|_1$. The 3D keypoints $X_k^{(t)}$ estimated from BPST (Sec. 3.2) are used to capture posture-induced surface deformations, as they provide robust and stable supervision for mesh reconstruction. $\hat{X}_k^{(t)} = MV^{(t)}$ denotes the reconstructed keypoints and $\nu_k^{(t)}$ indicates visibility.

Mask loss: $E_{\text{mask}} = \frac{1}{TC}\sum_{t=1}^{T}\sum_{c=1}^{C} \|\hat{S}_c^{(t)} - S_c^{(t)}\|_2$. Rendered silhouettes $\hat{S}_c^{(t)}$ are compared with segmentation masks $S_c^{(t)}$ extracted by HQ-SAM.

Color loss: $E_{\text{rgb}} = \frac{1}{TC}\sum_{t=1}^{T}\sum_{c=1}^{C} \hat{S}_c^{(t)} \odot \|\hat{I}_c^{(t)} - I_c^{(t)}\|_1$. The color loss enforces photometric consistency between the rendered appearance and the observed images. Here $\hat{I}_c^{(t)} = R(V^{(t)}, f, \pi_c)$ denotes differentiable renderings. $\odot$ denotes the element-wise (Hadamard) product used to weight the color error by $\hat{S}_c^{(t)}$.

Biological constraints. Bumblebees have a complex structure and highly flexible limbs (*e.g.* 27 degrees of freedom in the limbs), increasing the difficulty of data-driven mesh deformation. Like many insects, bumblebees have bilaterally symmetrical bodies [22, 45]. Proper setting of kinematic degrees of freedom has been shown to improve the accuracy and plausibility of pose estimation [30, 37, 51]. We incorporate symmetry loss, kinematic constraints, and temporal smoothness as biological priors to restrict joint motion, reduce unrealistic deformations, and enhance posture consistency across frames:

$$E_{\text{bio}} = \lambda_{\text{sym}} E_{\text{sym}} + \lambda_{\text{kine}} E_{\text{kine}} + \lambda_{\text{temp}} E_{\text{temp}}. \tag{5}$$

Symmetry loss: $E_{\text{sym}} = \frac{1}{|Z|}\sum_{(i,j)\in Z} \|\beta_i - \beta_j\|_2^2$. Bilateral limb pairs are constrained to maintain a symmetric morphology. Z denotes the set of symmetric body segment pairs, as the insect body is roughly symmetric about the midline.

Kinematic constraint: $E_{\text{kine}} = \frac{1}{TB}\sum_{t=1}^{T}\sum_{b=1}^{B}\left\|\max\left(0,\ L_b - \theta_b^{(t)},\ \theta_b^{(t)} - U_b\right)\right\|_2^2$. Joint angles are limited within species-specific anatomical ranges. Here, L_b and U_b denote the upper and lower bounds of the joints.

Temporal smoothness: $E_{\text{temp}} = \frac{1}{T-1}\sum_{t=1}^{T-1}\|\theta^{(t+1)} - \theta^{(t)}\|_2^2$. Pose variation across consecutive frames is regularized to ensure motion consistency.

3.5 Training Optimization

We adopt a two-stage scheme to decouple grid vertex movement caused by varying parameter changes. **Stage I** optimizes $\{\beta, s, \phi, \gamma, f\}$ on a subset of frames (no temporal terms) to align the template to the individual. With the shape and global parameters fixed, **Stage II** optimizes the pose parameters per-frame $\theta^{(t)}$. We implement in PyTorch/PyTorch3D [16] with Adam [21] and decay of the cosine learning-rate [32].

4 Experiments

4.1 Dataset Description

High-quality visual datasets of animals have greatly advanced behavioral analysis [10, 13]. To obtain clear and minimally occluded multi-view recordings of bumblebee motion, we build a synchronized multi-camera acquisition setup consisting of six Hikvision MV-CS016-10UC cameras (Sony IMX273 sensor; max 1440×1080), equipped with MVL-HF2524M-10MP macro lenses (25 mm), a working distance of approximately 250 mm and an effective precision of 0.035 mm/pixel. Image capture runs at up to 250 FPS with a minimum exposure time of 15 μs. The controlled studio setup keeps the bumblebee within the camera field of view while allowing rapid limb movements to be captured with clear motion details.

We record an individual (*ID*: F-FB-230701-5-3) and construct the Deep Bumblebee Studio dataset(DBS), containing six synchronized views of a continuous 2000-frame sequence at 200 FPS and 1000×800 resolution. From each view, 1800 frames are manually annotated with 2D keypoints by two experts and verified by a third; 1500 frames are used to train DeepLabCut and the remaining 300 frames per view are triangulated to obtain 3D ground-truth keypoints.

4.2 3D Pose Estimation Results

We evaluate joint localization accuracy, temporal smoothness, and anatomical stability on the DBS dataset to verify the effectiveness of BPST. Detailed comparisons and analyses are provided in the supplementary material. Let T be the number of frames and K the number of 3D keypoints. With predictions $\hat{\mathbf{X}}_k^{(t)}$ and ground truth $\mathbf{X}_k^{(t)}$, the Mean Per-Joint Position Error (MPJPE) is defined as:

$$\text{MPJPE} = \frac{1}{TK}\sum_{t=1}^{T}\sum_{k=1}^{K}\left\|\hat{\mathbf{X}}_k^{(t)} - \mathbf{X}_k^{(t)}\right\|_2. \tag{6}$$

We also report PA-MPJPE and 3D-PCK@δ with δ=0.35 mm.

Joint position accuracy. On the DBS test split, BPST achieves an MPJPE of 0.110 mm, PA-MPJPE of 0.107 mm and 3D-PCK of 98.3%, showing accurate multi-view triangulation and strong spatial consistency. In Fig. 4, BPST robustly recovers 3D poses for small, articulated bumblebees despite severe self-occlusion.

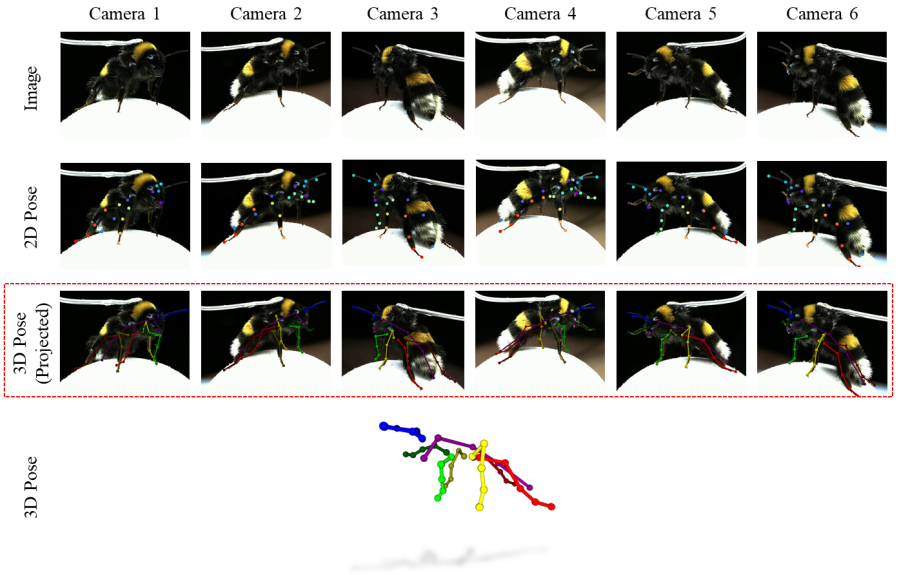

Fig. 4: 3D pose estimation results. BPST effectively recovers accurate 3D keypoints for small, highly articulated bumblebees, overcoming missed detection caused by occlusion in 2D pose estimation.

Table 1: Ablation of BPST components for 3D pose estimation. Lower is better for MPJPE/PA-MPJPE; higher is better for 3D-PCK.

Method	MPJPE (mm)↓	PA-MPJPE (mm)↓	3D-PCK (%)↑
Ours (w/o ρ)	0.125	0.121	97.9
Ours (w/o E_t)	0.119	0.116	98.0
Ours (w/o E_s)	0.115	0.112	98.1
Ours	**0.110**	**0.107**	**98.3**

Ablation. We remove (i) the robust reprojection term ρ, (ii) the temporal smoothness term E_t, and (iii) the segment-length constraint E_s. The results are shown in Tab. 1. Removing ρ causes the largest degradation (+0.015 mm

MPJPE), highlighting its role in suppressing outliers. Disabling E_t increases MPJPE by 0.009 mm and reduces temporal stability. Removing E_s leads to reduced anatomical coherence. The full model integrating all components achieves the best accuracy and stability.

Comparative Experiment. Compared with a standard DeepLabCut+3D triangulation pipeline, BPST improves joint position accuracy by 16.7%, temporal smoothness by 16.9%, and segment-length stability by 15.8%. More comparisons and analyses with baselines are provided in the supplementary material.

4.3 Mesh Reconstruction Results

Pose accuracy is evaluated by mapping mesh vertices to keypoints through the matrix $\mathbf{M}$ (Sec. 3.1) and reporting MPJPE, PA-MPJPE, and 3D-PCK@0.35 mm. Shape accuracy is measured by the multi-view mask IoU between rendered silhouettes $\hat{S}_c^{(t)}$ and ground-truth $S_c^{(t)}$:

$$\text{IoU} = \frac{1}{TC} \sum_{t=1}^{T} \sum_{c=1}^{C} \frac{\left| \hat{S}_c^{(t)} \cap S_c^{(t)} \right|}{\left| \hat{S}_c^{(t)} \cup S_c^{(t)} \right|}. \tag{7}$$

Reconstruction accuracy. Our method achieves an MPJPE of 0.152 mm, PA-MPJPE of 0.147 mm, 3D-PCK of 96.6%, and IoU of 0.758, demonstrating accurate and anatomically consistent mesh reconstruction. As shown in Fig. 5, The bumblebee exhibits rapid, large limb motions, while our method accurately captures these nonrigid deformations and reconstructs meshes closely aligned with the target geometry.

Table 2: Ablation on the bumblebee dataset for mesh reconstruction

Method	MPJPE (mm)↓	PA-MPJPE (mm)↓	3D-PCK (%)↑	IoU ↑
w/o E_{sym}	0.163	0.156	93.6	0.724
w/o E_{kine}	0.166	0.160	93.1	0.717
w/o E_{mask}	**0.148**	**0.144**	**96.9**	0.652
Full method	0.152	0.147	96.6	**0.758**

Ablation. As shown in Tab. 2, removing E_{sym} increases MPJPE/PA-MPJPE by 0.011/0.009 mm and reduces IoU by 3.4 points. Disabling E_{kine} further degrades performance (+0.014/0.013 mm MPJPE/PA-MPJPE, –4.1 IoU), showing that biological constraints effectively maintain anatomical validity. Eliminating E_{mask} causes the largest IoU drop (–10.6 points) with only a minor increase in MPJPE (+0.004 mm), indicating a trade-off between sparse keypoint and dense silhouette supervision. The full model achieves the best overall accuracy and structural consistency.

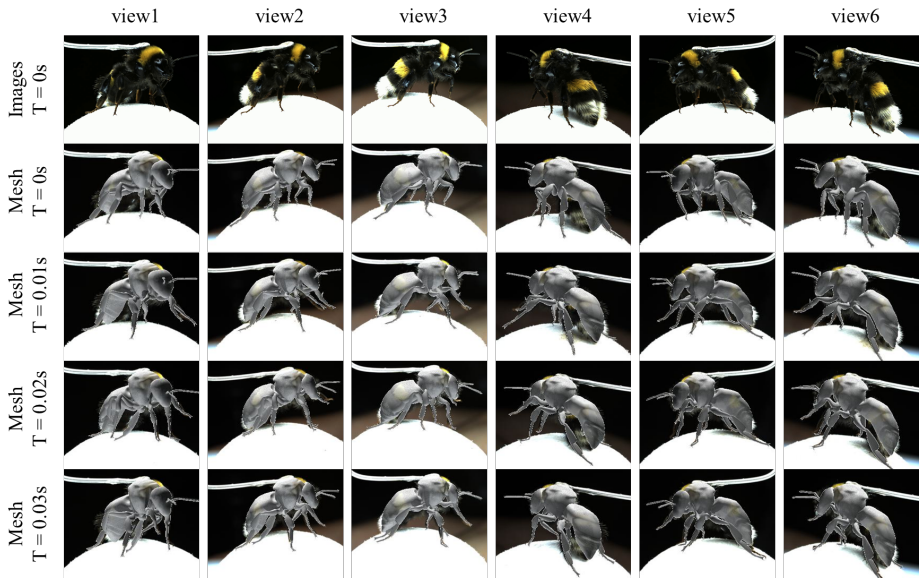

Fig. 5: Mesh reconstruction results of a bumblebee in motion. Despite rapid limb movements, our method accurately captures nonrigid deformations and reconstructs meshes closely aligned with the ground truth.

5 Discussion

5.1 Mesh Reconstruction Accuracy and Comparative Studies

We compare our method with three representative animal reconstruction pipelines: MAMMAL [1], DANNCE [10], and its temporal variant DANNCE-T [25]. In addition, we include representative template-free methods (e.g., Pad3r [28], fauna [27], and 3D SAM [46]) for qualitative comparison. All methods are trained and evaluated on the same splits described in Sec. 4.1.

Accuracy comparison. As shown in Table 3, our method achieves the best results on all metrics: MPJPE **0.152 mm** (2.1% of thorax length), PA-MPJPE **0.147 mm**, 3D-PCK **96.6%** and IoU **0.758**. This improves over MAMMAL by 0.037 mm MPJPE (19.6% relative), +4.6 pp PCK and +6.2 pp IoU; joint error is also lower than DANNCE/DANNCE-T by 0.077/0.067 mm. Figure 6 reports 3D-PCK in five limb joints (TrF, FTi, TiB, BE, CE). Our method performs best across all joints; at TrF (close to the body, often occluded), it reaches **84.8**% vs. 70.9/71.4/77.3% for DANNCE/DANNCE-T/MAMMAL.

Qualitative results. As illustrated in Fig. 7, we compare BioMTBee with MAMMAL and representative template-free methods (e.g., Pad3r [28], fauna [27], and 3D SAM [46]) on bumblebee reconstruction. BioMTBee produces more anatomically realistic results with accurate body proportions, thin appendage preservation, and coherent anatomical topology. In contrast, MAMMAL ex-

Table 3: Bumblebee dataset: accuracy comparison

Method	MPJPE (mm)↓	PA-MPJPE (mm)↓	3D-PCK (%)↑	IoU ↑
DANNCE	0.229	0.217	84.9	–
DANNCE-T	0.219	0.206	87.3	–
MAMMAL	0.189	0.178	92.0	0.696
Ours	**0.152**	**0.147**	**96.6**	**0.758**

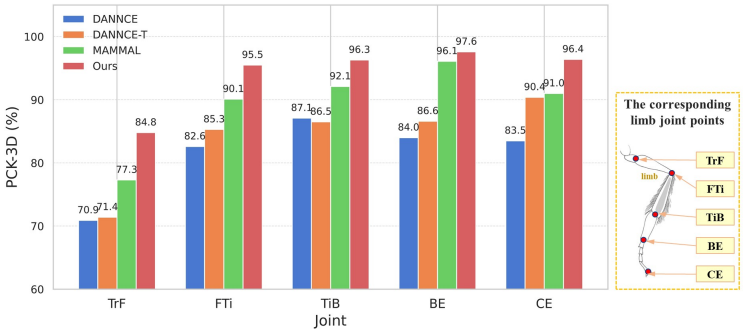

Fig. 6: 3D-PCK comparison across methods on bumblebee limb joints

hibits distorted abdomen proportions, while template-free approaches often generate inconsistent geometry and biologically implausible structures, particularly around fine articulated regions. These challenges hinder their suitability for insect motion analysis requiring accurate morphological and articulation modeling.

5.2 Biological Plausibility and Symmetry

We assess biological plausibility through two metrics: (i) *Joint-angle violation rate*, the percentage of frames where estimated joint angles exceed anatomical limits; (ii) *Bilateral symmetry*, the discrepancy in the left–right length of the corresponding limb segments.

Kinematic validity. Table 4 shows that the baseline MAMMAL exhibits high violation rates (e.g., FTi: 59.2% at 10°). Without E_{kine}, our model reduces violations but still exceeds 20% on proximal joints. With E_{kine}, the rates drop below 8% at $> 10°$ and nearly zero at $> 30°$, demonstrating that the kinematic constraint effectively enforces physiological motion limits.

Morphological symmetry. Table 5 reports the mean left–right discrepancies of the limb segments. MAMMAL shows strong asymmetry (hindlimb femur 1.02 mm, 18.9% of segment length). Without E_{sym}, the difference decreases to 0.427 mm; with E_{sym}, it further drops to 0.357 mm (–65%), indicating that the symmetry constraint maintains consistent bilateral morphology and mitigates the deformation bias.

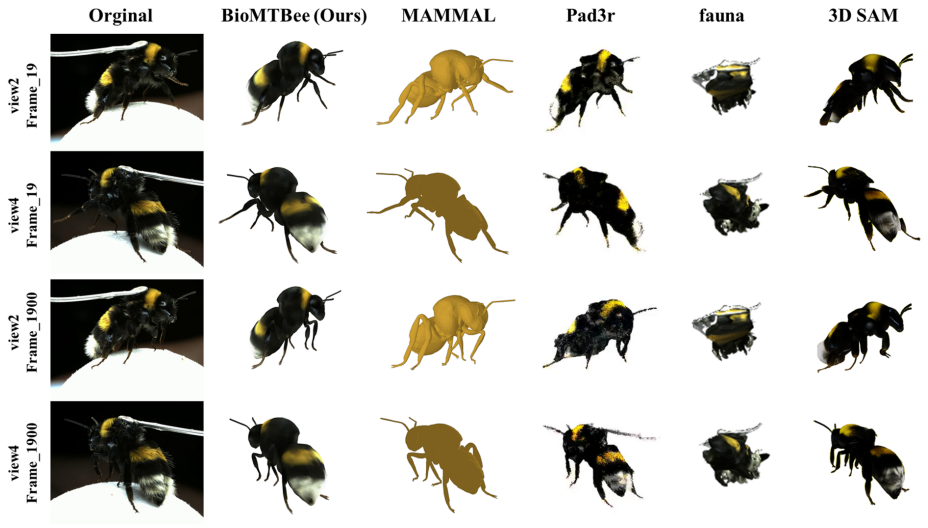

Fig. 7: Qualitative comparison on our DBS dataset. BioMTBee preserves more realistic body proportions, thin appendages, and anatomical topology than template-based and template-free baselines.

Table 4: Joint-angle violation rate (%) at three thresholds (lower is better)

Method	Violation $> 10°$				Violation $> 20°$				Violation $> 30°$			
	ThC	CTr	FTi	BE	ThC	CTr	FTi	BE	ThC	CTr	FTi	BE
MAMMAL	36.4	34.0	59.2	45.7	15.5	14.6	20.1	18.0	5.1	4.4	7.9	6.9
Ours (w/o E_{kine})	31.0	23.8	43.3	36.2	12.7	8.3	14.4	15.2	3.4	1.3	3.5	6.1
Ours	**3.2**	**1.8**	**7.9**	**6.5**	**1.6**	**0.9**	**5.4**	**4.7**	**0.0**	**0.0**	**0.5**	**0.3**

The kinematic and symmetry priors substantially reduce joint-angle violations and left–right discrepancies, yielding anatomically valid and stable reconstructions.

5.3 Category Outside Bumblebee: Mouse Dataset

We evaluate cross-species generalization on a public multi-view *mouse* dataset using a parametric skeletal mesh [6] with 140 movable joints, 14,522 vertices, and 22 keypoints. The global pose and joint-angle limits are adapted with a mouse-specific biomechanics model [37].

Table 6 shows that our approach achieves the best pose and shape accuracy (MPJPE 2.01 mm, PA-MPJPE 1.95 mm, 3D-PCK 93.6%, IoU 0.774), outperforming the other three reconstruction methods. Table 7 shows that removing E_{sym} increases MPJPE by 0.13 mm and reduces IoU by 2.7 pp; removing E_{kine}

Table 5: Bilateral left–right length discrepancy (mm; lower is better)

Method	Forelimb			Midlimb			Hindlimb		
	Femur	Tibia	Basitarsus	Femur	Tibia	Basi	Femur	Tibia	Basi
MAMMAL	0.244	0.063	0.178	0.408	0.097	0.141	1.020	0.096	0.418
Ours (w/o E_{sym})	0.168	0.069	0.126	0.420	0.069	0.103	0.427	0.063	0.318
Ours	**0.128**	**0.051**	**0.106**	**0.251**	**0.059**	**0.079**	**0.357**	**0.055**	**0.254**

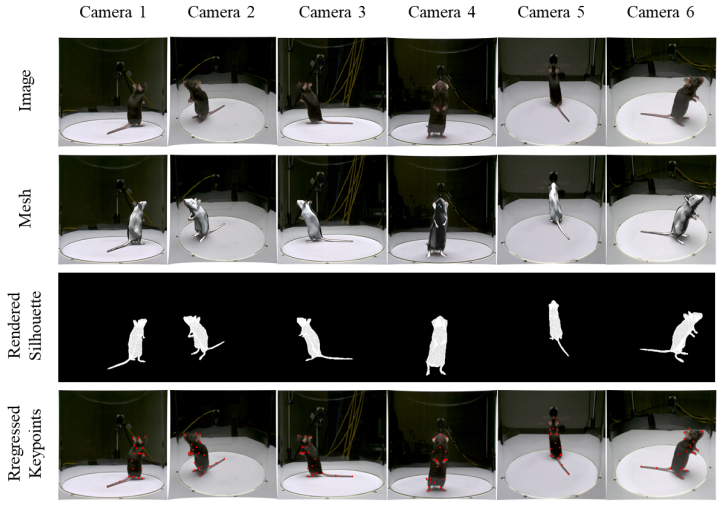

Fig. 8: Mesh reconstruction of mouse. Our method reconstructs mouse effectively owing to its capability to accurately reconstruct bumblebees with highly complex and flexible structures.

yields the largest pose error (+0.22 mm) and a 4.2 pp IoU drop; removing E_{mask} slightly improves pose errors but reduces IoU by 7.9 pp.

Multi-view renderings (Fig. 8) show close alignment among mesh, silhouettes, and keypoints, consistent with quantitative results. Since our method achieves high-precision reconstruction even in bumblebees with complex morphology and highly flexible limbs, it can be easily applied to larger and structurally simpler species, such as mice.

6 Conclusion

We presented **BioMTBee**, a multi-view, template-based system for high-precision 3D reconstruction of moving bumblebees. To address challenges including small body size, dense hair, self-occlusion, and rapidly moving limbs, BioMTBee integrates a CT-derived template, a spatio–temporal pose estimator (BPST), and biologically constrained differentiable rendering in a unified pipeline.

Table 6: Mouse dataset: accuracy comparison

Method	MPJPE (mm)↓	PA-MPJPE (mm)↓	3D-PCK (%)↑	IoU ↑
DANNCE	4.99	4.54	78.1	–
DANNCE-T	4.78	4.31	79.5	–
MAMMAL	2.43	2.25	90.2	0.751
Ours	**2.01**	**1.95**	**93.6**	**0.774**

Table 7: Mouse dataset: ablation of priors

Method	MPJPE (mm)↓	PA-MPJPE (mm)↓	3D-PCK (%)↑	IoU ↑
w/o E_{sym}	2.14	2.02	92.5	0.747
w/o E_{kine}	2.23	2.15	91.1	0.732
w/o E_{mask}	**1.93**	**1.88**	**94.0**	0.695
Full method	2.01	1.95	93.6	**0.774**

The articulated template encodes 63 bones with joint-specific degrees of freedom derived from bumblebee anatomy, forming a targeted parameter space for stable optimization. BPST provides temporally consistent 3D keypoints and enables automatic multi-view calibration from 2D detections. Biological constraints, including kinematic constraints, bilateral symmetry, and temporal smoothness, regularize non-rigid deformation and maintain anatomical validity. Experiments on the DBS dataset demonstrate accurate surface reconstruction (MPJPE 0.152 mm, PA-MPJPE 0.147 mm, 3D-PCK 96.6%, IoU 0.758) with strong temporal stability and minimal joint-angle violations. The method also generalizes to a public mouse dataset, indicating cross-species applicability and robustness.

BioMTBee provides a scalable and reliable solution for fine-grained insect motion capture, enabling quantitative studies in insect biomechanics, ecological monitoring, precision agriculture, and bio-inspired robotics. Future work will extend the framework to larger datasets and more diverse insect species, improving automation and generalization in real-world biological settings.

Acknowledgements

We gratefully acknowledge support from the Natural Science Foundation of Zhejiang Province under Grant LZ24F020003, the Shanghai Key Technology Research and Development Program under Grant 25XF3200500, the Excellent Research Innovation Team of Anhui Provincial Education Department under Grant 2024AH010018, the Leading Goose R&D Program of Zhejiang Province under Grant 2025C02045 and 2026C02A1226, the National Natural Science Foundation of China under Grant T2293723.

References

1. An, L., Ren, J., Yu, T., et al: Three-dimensional surface motion capture of multiple freely moving pigs using mammal. Nature Communications **14**, 7727 (2023). https://doi.org/10.1038/s41467-023-43483-w
2. Badger, M., Wang, Y., Modh, A., et al: 3d bird reconstruction: A dataset, model, and shape recovery from a single view. In: ECCV. pp. 1–17 (2020). https://doi.org/10.1007/978-3-030-58523-5_1
3. Bala, P.C., Eisenreich, B.R., Yoo, S.B.M., et al: Automated markerless pose estimation in freely moving macaques with openmonkeystudio. Nature Communications **11**(1), 4560 (2020). https://doi.org/10.1038/s41467-020-18441-5
4. Biggs, B., Boyne, O., Charles, J., et al: Who left the dogs out? 3d animal reconstruction with expectation maximization in the loop. In: ECCV. pp. 195–211 (2020). https://doi.org/10.1007/978-3-030-58621-8_12
5. Bohnslav, J.P., Osman, M.A.M., Jaggi, A., et al: Armo: An articulated mesh approach for mouse 3d reconstruction. bioRxiv (2023). https://doi.org/10.1101/2023.02.17.526719
6. Bolaños, L.A., Xiao, D., Ford, N.L., et al: A three-dimensional virtual mouse generates synthetic training data for behavioral analysis. Nature Methods **18**, 378–381 (2021). https://doi.org/10.1038/s41592-021-01103-9
7. D, G., GC, L., B., D.: Decline and conservation of bumble bees. Annual Review of Entomology **53**(1), 191–208 (2008). https://doi.org/10.1146/annurev.ento.53.103106.093454
8. Doan, T.N., Nguyen, C.V.: A low-cost digital 3d insect scanner. Information Processing in Agriculture **11**, 337–355 (2024). https://doi.org/10.1016/j.inpa.2023.03.003
9. Dolan, A.C., Delphia, C.M., O'Neill, K.M., Ivie, M.A.: Bumble bees (hymenoptera: Apidae) of montana. Annals of the Entomological Society of America **110**(2), 129–144 (2017). https://doi.org/10.1093/aesa/saw064
10. Dunn, T.W., Marshall, J.D., Severson, K.S., et al: Geometric deep learning enables 3d kinematic profiling across species and environments. Nature Methods **18**, 564–573 (2021). https://doi.org/10.1038/s41592-021-01106-6
11. Forrester, G.S., Hopkins, W.D., Hudry, K., Lindell, A.: Chapter 1 - insights into the evolution of lateralization from the insects. In: Cerebral Lateralization and Cognition: Evolutionary and Developmental Investigations of Behavioral Biases, vol. 238, pp. 3–31. Elsevier (2018). https://doi.org/10.1016/bs.pbr.2018.06.001
12. Graving, J.M., Chae, D., Naik, H., et al: Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning. eLife **8**, e47994 (2019). https://doi.org/10.7554/eLife.47994
13. Günel, S., Rhodin, H., Morales, D., et al: Deepfly3d, a deep learning-based approach for 3d limb and appendage tracking in tethered, adult *Drosophila*. eLife **8**, e48571 (2019). https://doi.org/10.7554/eLife.48571
14. H. Kato, Y.U., Harada, T.: Neural 3d mesh renderer. In: CVPR. pp. 3907–3916 (2018). https://doi.org/10.1109/CVPR.2018.00411
15. Jakab, T., Li, R., Wu, S., Rupprecht, C., Vedaldi, A.: Farm3d: Learning articulated 3d animals by distilling 2d diffusion. In: 2024 International Conference on 3D Vision (3DV). pp. 852–861 (2024). https://doi.org/10.1109/3DV62453.2024.00051
16. Johnson, J., Ravi, N., Reizenstein, J., et al: Accelerating 3d deep learning with pytorch3d. SIGGRAPH Asia 2020 Courses (2020). https://doi.org/10.1145/3415263.3419160

17. Kalavakonda, R.R., Dehghanzadeh, P., Huan, J., et al: Fusion intelligence: A paradigm for merging natural and artificial intelligence. IEEE Internet of Things Journal **12**(15), 30548–30563 (2025). https://doi.org/10.1109/JIOT.2025.3572367
18. Kanazawa, A., Tulsiani, S., Efros, A.A., Malik, J.: Learning category-specific mesh reconstruction from image collections. *arXiv:1803.07549* (2018). https://doi.org/10.48550/arXiv.1803.07549
19. Ke, L., Ye, M., Danelljan, M., et al: Segment anything in high quality. In: NeurIPS. vol. 36, pp. 29914–29934 (2024), https://proceedings.neurips.cc/paper_files/paper/2023/file/5f828e38160f31935cfe9f67503ad17c-Paper-Conference.pdf
20. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023), https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv:1412.6980* (2017)
22. Klingenberg, C.P., Badyaev, A.V., Sowry, S.M., Beckwith, N.J.: Inferring developmental modularity from morphological integration: analysis of individual variation and asymmetry in bumblebee wings. The American naturalist **157**, 11–23 (2001). https://doi.org/10.1086/317002.
23. Kulkarni, N., Gupta, A., Fouhey, D.F., Tulsiani, S.: Articulation-aware canonical surface mapping. In: CVPR. pp. 449–458 (2020). https://doi.org/10.1109/CVPR42600.2020.00053
24. Li, C., Ghorbani, N., Broomé, S., et al: hsmal: Detailed horse shape and pose reconstruction for motion pattern recognition. *arXiv:2106.10102* (2021). https://doi.org/10.48550/arXiv.2106.10102
25. Li, T., Severson, K.S., Wang, F., et al: Improved 3d markerless mouse pose estimation using temporal semi-supervision. International Journal of Computer Vision **131**, 1389–1405 (2023). https://doi.org/10.1007/s11263-023-01756-3
26. Li, Z., Amrani, A., Rai, S., Laga, H.: Advances and trends in the 3d reconstruction of the shape and motion of animals. *arXiv:2508.16062* (2025). https://doi.org/10.48550/arXiv.2508.16062
27. Li, Z., Litvak, D., Li, R., et al: Learning the 3d fauna of the web. In: CVPR. pp. 9752–9762 (2024). https://doi.org/10.1109/CVPR52733.2024.00931
28. Liao, T.H., Liu, H., Xu, Y., et al: Pad3r: Pose-aware dynamic 3d reconstruction from casual videos. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers (2025). https://doi.org/10.1145/3757377.3763836
29. Liu, R., Wu, R., Hoorick, B.V., et al: Zero-1-to-3: Zero-shot one image to 3d object. *arXiv:2303.11328* (2023). https://doi.org/10.48550/arXiv.2303.11328
30. Lobato-Rios, V., Ramalingasetty, S.T., Özdil, P.G., et al.: Neuromechfly, a neuromechanical model of adult drosophila melanogaster. Nature Methods **19**(5), 620–627 (2022). https://doi.org/10.1038/s41592-022-01466-7
31. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. SIGGRAPH Comput. Graph. **21**, 163–169 (1987). https://doi.org/10.1145/37402.37422
32. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. *arXiv:1608.03983* (2017). https://doi.org/10.48550/arXiv.1412.6980
33. Lyu, J., Zhu, T., Gu, Y., et al: Animer: Animal pose and shape estimation using family aware transformer. In: CVPR. pp. 17486–17496 (2025). https://doi.org/10.1109/CVPR52734.2025.01629

34. MaBouDi, H., Richter, J., Guiraud, M.G., et al: Active vision of bees in a simple pattern discrimination task. eLife **14**, e106332 (2025). https://doi.org/10.7554/eLife.106332
35. Mathis, A., Mamidanna, P., Cury, K.M., et al: Deeplabcut: Markerless pose estimation of user-defined body parts with deep learning. Nature Neuroscience **21**, 1281–1289 (2018). https://doi.org/10.1038/s41593-018-0209-y
36. Mathis, M.W., Mathis, A.: Deep learning tools for the measurement of animal behavior in neuroscience. Current Opinion in Neurobiology **60**, 1–11 (2020). https://doi.org/10.1016/j.conb.2019.10.008
37. Monsees, A., Voit, K.M., Wallace, D.J., et al: Estimation of skeletal kinematics in freely moving rodents. Nature Methods **19**(11), 1500–1509 (2022). https://doi.org/10.1038/s41592-022-01634-9
38. Ng, X.L., Ong, K.E., Zheng, Q., et al: Animal kingdom: A large and diverse dataset for animal behavior understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19023–19034 (2022). https://doi.org/10.1109/CVPR52688.2022.01844
39. Nityananda, V., Chittka, L.: Modality-specific attention in foraging bumblebees. Royal Society Open Science **2**(10), 150324 (2015). https://doi.org/10.1098/rsos.150324
40. OpenSfM Contributors: Opensfm: Incremental reconstruction algorithm documentation (2024), software documentation, available at https://opensfm.org/docs/reconstruction_module.html
41. P, G., van Leeuwen JL, FT, M.: Bumblebees compensate for the adverse effects of sidewind during visually guided landings. Journal of Experimental Biology **227**(8), jeb245432 (2024). https://doi.org/10.1242/jeb.245432
42. Rogers, M., Gendron, G., Valdez, D.A.S., et al: Meerkat behaviour recognition dataset. *arXiv:2306.11326* (2023). https://doi.org/10.48550/arXiv.2306.11326
43. Schindelin, J., Arganda-Carreras, I., Frise, E., et al: Fiji: an open-source platform for biological-image analysis. Nature Methods **9**(7), 676–682 (2012). https://doi.org/10.1038/nmeth.2019
44. Sinha, S., Shapovalov, R., Reizenstein, J., et al: Common pets in 3d: Dynamic new-view synthesis of real-life deformable categories. CVPR pp. 4881–4891 (2023). https://doi.org/10.1109/CVPR52729.2023.00473
45. Smits, C.M., Dutta, S., Jain-Sharma, V., et al: Maintaining symmetry during body axis elongation. Current Biology **33**(16), 3536–3543.e6 (2023). https://doi.org/10.1016/j.cub.2023.07.050
46. Team, S.D., Chen, X., Chu, F.J., et al: Sam 3d: 3dfy anything in images. *arXiv:2511.16624* (2025). https://doi.org/10.48550/arXiv.2511.16624
47. Velthuis, H.H.W., van Doorn, A.: A century of advances in bumblebee domestication and the economic and environmental aspects of its commercialization for pollination. Apidologie **37**(4), 421–451 (2006). https://doi.org/10.1051/apido:2006019
48. Vicente, S., Agapito, L.: Balloon shapes: Reconstructing and deforming objects with volume from images. In: International Conference on 3D Vision (3DV). pp. 223–230 (2013). https://doi.org/10.1109/3DV.2013.37
49. Wang, G., Yu, J., Liu, S., et al: Automatic fish weight estimation and 3d surface reconstruction with a lightweight instance segmentation model. Expert Systems with Applications **288**, 128275 (2025). https://doi.org/10.1016/j.eswa.2025.128275

50. Wu, S., Jakab, T., Rupprecht, C., Vedaldi, A.: Dove: Learning deformable 3d objects by watching videos. International Journal of Computer Vision **131**, 2623–2634 (2023). https://doi.org/10.1007/s11263-023-01819-5
51. Xia, Y., Zhou, X., Vouga, E., et al: Reconstructing humans with a biomechanically accurate skeleton. In: CVPR. pp. 5355–5365 (2025). https://doi.org/10.1109/CVPR52734.2025.00504
52. Xu, J., Zhang, Y., Peng, J., et al: Animal3d: A comprehensive dataset of 3d animal pose and shape. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9065–9075 (2024). https://doi.org/10.1109/ICCV51070.2023.00835
53. Yang, G., Sun, D., Jampani, V., et al: Lasr: Learning articulated shape reconstruction from a monocular video. In: CVPR. pp. 15980–15989 (2021). https://doi.org/10.1109/CVPR46437.2021.01572
54. Yang, G., Vo, M., Neverova, N., et al: Banmo: Building animatable 3d neural models from many casual videos. In: CVPR. pp. 2853–2863 (2022). https://doi.org/10.1109/CVPR52688.2022.00288.
55. Yao, C.H., Hung, W.C., Li, Y., et al: Lassie: Learning articulated shapes from sparse image ensembles via 3d part discovery. *arXiv:2207.03434* (2023). https://doi.org/10.48550/arXiv.2207.03434
56. Zhou, Q.Y., Park, J., Koltun, V.: Fast global registration. In: ECCV. pp. 766–782 (2016). https://doi.org/10.1007/978-3-319-46475-6_47
57. Zuffi, S., Kanazawa, A., Berger-Wolf, T., Black, M.: Three-d safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In: ICCV. pp. 5358–5367 (2019). https://doi.org/10.1109/ICCV.2019.00546
58. Zuffi, S., Kanazawa, A., Black, M.J.: Lions and tigers and bears: Capturing non-rigid, 3d, articulated shape from images. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3955–3963 (2018). https://doi.org/10.1109/CVPR.2018.00416
59. Zuffi, S., Kanazawa, A., Jacobs, D.W., Black, M.J.: 3d menagerie: Modeling the 3d shape and pose of animals. In: CVPR. pp. 5524–5532 (2017). https://doi.org/10.1109/CVPR.2017.586.
60. Zuffi, S., Mellbin, Y., Li, C., et al: Varen: Very accurate and realistic equine network. In: CVPR. pp. 5374–5383 (2024). https://doi.org/10.1109/CVPR52733.2024.00514