arXiv:2506.01691v2 [cs.CV] 7 Aug 2025

STEERPOSE: SIMULTANEOUS EXTRINSIC CAMERA CALIBRATION AND MATCHING FROM ARTICULATION

A PREPRINT

Sang-Eun Lee
Graduate School of Informatics
Kyoto University

Ko Nishino
Graduate School of Informatics
Kyoto University

Shohei Nobuhara
Information and Human Sciences
Kyoto Institute of Technology

August 8, 2025

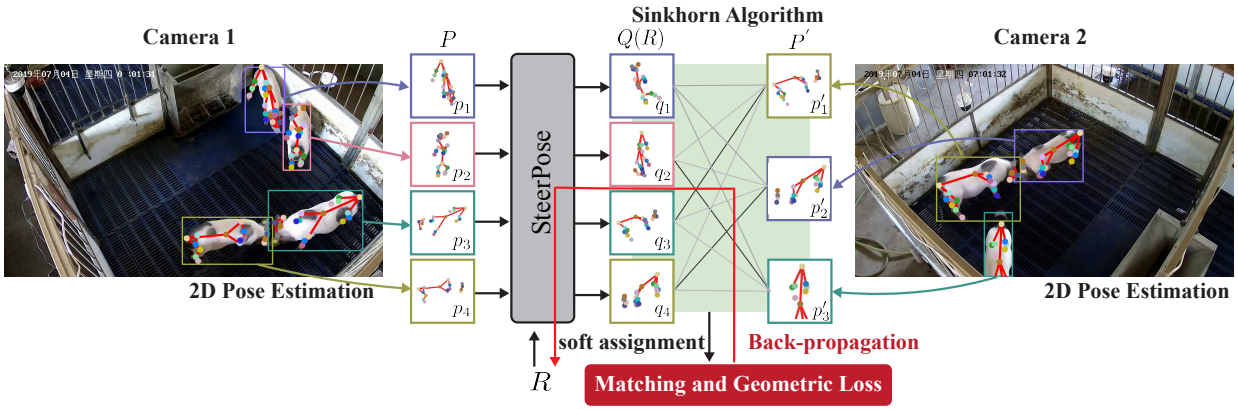


Figure 1: Camera calibration by *SteerPose*. Given a set of 2D poses P of targets with known articulation, SteerPose predicts their appearance under a relative camera rotation R as $Q(R)$ which can be matched against 2D poses P' observed in another view in a differentiable manner. Novel matching and geometric losses evaluate the validity of R as well as the matching between $Q(R)$ and P', and their gradients are back-propagated to R to optimize the relative rotation between the views and the matching simultaneously.

ABSTRACT

Can freely moving humans or animals themselves serve as calibration targets for multi-camera systems while simultaneously estimating their correspondences across views? We humans can solve this problem by *mentally rotating* the observed 2D poses and aligning them with those in the target views. Inspired by this cognitive ability, we propose *SteerPose*, a neural network that performs this rotation of 2D poses into another view. By integrating differentiable matching, *SteerPose* simultaneously performs extrinsic camera calibration and correspondence search within a single unified framework. We also introduce a novel geometric consistency loss that explicitly ensures that the estimated rotation and correspondences result in a valid translation estimation. Experimental results on diverse in-the-wild datasets of humans and animals validate the effectiveness and robustness of the proposed method. Furthermore, we demonstrate that our method can reconstruct the 3D poses of novel animals in multi-camera setups by leveraging off-the-shelf 2D pose estimators and our class-agnostic model.

1 Introduction

Motion capture is crucial for understanding human or animal behavior and social interactions, with a wide range of applications in fields such as biomechanics, psychology, and engineering [1, 2, 3, 4, 5, 6]. Multi-camera systems facilitate accurate motion capture without the need for intrusive markers; however, they require specialized tools and a cumbersome process of calibrating all cameras prior to on-site deployment. Furthermore, in the case of multiple objects in a scene, their correspondence across views should be estimated for 3D triangulation. These requirements limit the applicability of markerless motion capture, especially when studying subjects with unpredictable behavior in uncontrolled environments, such as animals in the wild or elderly people in their daily lives.

Can freely moving humans or animals themselves serve as calibration targets for multi-camera systems while simultaneously estimating their correspondences? If so, we can streamline motion capture from uncalibrated multi-view videos of freely moving animals such as livestock pigs (Fig. 1), eliminating the need for manual calibration and correspondence search.

This paper proposes a novel method for automatic extrinsic camera calibration and correspondence search which leverages the inherent articulated structure of moving subjects. By exploiting robust geometric constraints derived from these articulations, our method establishes reliable cross-view correspondences without requiring specialized tools and procedures. This enables markerless motion capture of freely moving subjects using only multi-view video sequences captured in unconstrained environments, thereby improving the flexibility and accessibility of motion capture.

Our key idea is inspired by *mental rotation* [7], a fundamental cognitive process in the human brain. When aligning an object in two images captured from different views, we can mentally rotate the object until it matches the reference view. Inversely, we can understand the relative pose between the views through this process. Furthermore, in the case of images with multiple objects, we can simultaneously understand their correspondences across views that are consistent with the estimated rotation.

Mimicking this psychological mechanism, we propose a neural model called *SteerPose*, which transforms a 2D pose from one view to another, given a hypothesized relative rotation between them. SteerPose realizes a rotation-covariant transform. Following Marcos *et al*. [8], a function $f(\cdot)$ is said to be rotation-covariant w.r.t. a transform $g(\cdot)$ if $f(g(\cdot)) = g'(f(\cdot))$ holds, where $g'(\cdot)$ is a second transform, which is itself a function of $g(\cdot)$. Considering a 3D-to-2D projection and a rotation in 3D as $f(\cdot)$ and $g(\cdot)$, SteerPose serves as $g'(\cdot)$ which *mentally rotates* the 2D projection of the latent 3D pose $f(\cdot)$ to be identical to that rotated in 3D by $g(\cdot)$ before projection.

By combining SteerPose with a differentiable bipartite matching module, we propose an end-to-end framework for simultaneous optimization of extrinsic camera calibration and correspondence search as shown in Fig. 1. Consider two sets of 2D poses of a target, such as humans or animals, extracted from images captured from different viewpoints using an off-the-shelf 2D pose estimator. Given a current estimate of the 3D rotation R between these views, SteerPose transforms the 2D poses of one side $p_i \in P$ to the other viewpoint as $q_i(R) \in Q(R)$. A matching module, such as Sinkhorn [9], finds the best match between $q_i(R) \in Q(R)$ and the 2D poses $p' \in P'$ observed on the other side based on the similarity between 2D poses. In addition to this 2D pose similarity loss, the estimated match is evaluated using our novel geometric loss, which guarantees that the estimated rotation R has a valid solution for translation. By optimizing these losses by refining R through back-propagation, we can estimate the relative pose between the views while simultaneously estimating the correspondences.

Our approach frames camera pose estimation and cross-view association as a single unified optimization problem. The key is to use body articulations as *structured* semantic keypoints. SteerPose learns how the 2D arrangement of body joint positions changes after rotations, and the predicted 2D poses can be used as *structured* correspondences in the matching module. We argue that body articulation can serve as general prior knowledge valid across a wide variety of targets as shown in Table 1.

We evaluated our method using diverse datasets covering both humans and several animal species, including dogs, pigs, pigeons, and cheetahs. The experimental results demonstrate that our approach outperforms both traditional epipolar geometry-based and state-of-the-art learning-based methods. We also quantitatively evaluated the contribution of our novel geometric loss by showing that the estimated rotation without this loss can be trapped by a false match that is valid in terms of 2D pose rotations but is inconsistent in terms of translation.

A notable application of the proposed method is to provide training datasets for image-based monocular 3D pose estimators of novel animals. As long as the target novel animal shares the same body articulation structure and their 2D poses can be estimated from captured images, SteerPose automatically finds corresponding instances from different views while calibrating the relative camera pose simultaneously. This result allows for triangulating the 3D poses, which can be used as a pseudo ground-truth for training monocular 3D pose estimators. We demonstrate that by leveraging class-agnostic SteerPose and simply capturing freely moving quadruped animals not included in the training data, we

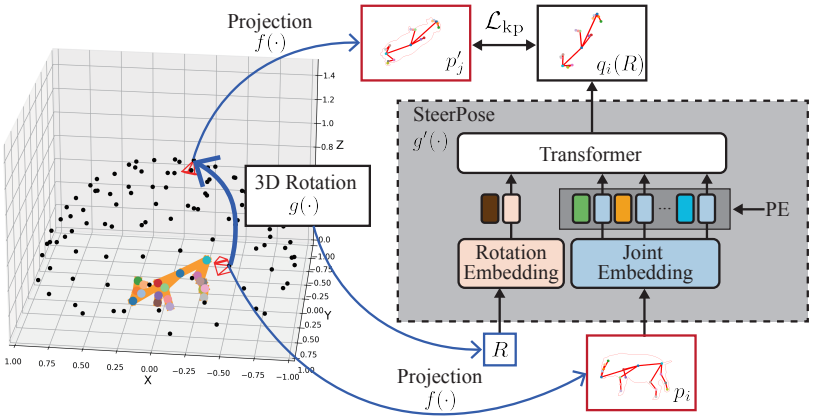


Figure 2: *SteerPose* learns to *rotate* an input 2D pose p_i to the 2D pose $q_i(R)$ observed from another view at relative rotation R in a fully-supervised manner. SteerPose serves as a function $g'(\cdot)$ that performs *mental rotation* to form a rotation-covariant transformation of 3D poses under projection $f(\cdot)$ satisfying $f(g(\cdot)) = g'(f(\cdot))$ where $g(\cdot)$ is the rotation in 3D by R. The training dataset was synthesized by rendering 3D poses from random viewpoints.

can achieve multi-view calibration and establish instance matches across views, resulting in the 3D reconstruction of these novel quadrupeds.

2 Related Work

Extrinsic calibration is the process of estimating the relative poses and positions of cameras. Conventional approaches require specialized reference objects with known geometries such as checkerboards, AR markers, or calibration wands [10, 11, 12]. Since procedures using these tools are usually performed on-site prior to deployment [13, 14, 4, 3], these calibration methods face challenges in “in-the-wild" scenarios.

Traditional SfM-based methods [15, 16] utilize static background features such as SIFT [17]. While this approach is known to be robust for moving cameras such as SLAM [18], it becomes challenging in wide-baseline setups such as motion capture due to appearance changes across views. In contrast, the use of dynamic foreground objects, in particular humans themselves, as calibration targets exploits semantic information such as their inherent motion and structure. This approach not only increases the robustness of the calibration but also eliminates the need for specialized calibration tools and tedious procedures [19, 20, 21, 22, 23, 24]. These methods, however, require correspondences of the targets between views, and/or a monocular 3D pose estimator.

Recent studies have integrated learnable components into camera pose estimation. Leveraging learned priors for keypoint detection, feature extraction, and matching [25, 26, 27] has demonstrated improved performance in camera pose estimation by establishing reliable correspondences across different views, particularly in wide-baseline scenarios. *Steerers* [28] facilitate keypoint matching in scenarios involving rotations by learning a rotation-equivariant linear mapping in the descriptor space. The Differentiable 8-point algorithm [29] unifies feature matching and camera pose estimation in an end-to-end framework, holistically addressing these challenges while eliminating the need for RANSAC [30]. More recent approaches, such as DUSt3R [31] and MASt3R [32], address correspondence estimation, and 3D reconstruction in a single pipeline by directly regressing dense 3D point maps from unposed images, enabling camera pose estimation without explicit correspondence matching.

Compared with these existing methods, the proposed method realizes simultaneous optimization of camera poses and correspondences using foreground objects as a calibration target, which provides a structured correspondence of semantic features, *i.e.*, joints, across views. A summary is provided in Appendix B

3 Two-view Extrinsic Calibration

3.1 SteerPose

The goal of SteerPose is to *mentally rotate* a 2D pose found in an image into a 2D pose in another view at rotation R. SteerPose realizes a rotation-covariant transform [8] of the latent 3D pose behind the observed 2D pose.

As shown in Fig. 2, we employ a Transformer [33] architecture that encodes articulated poses via multi-head self-attention and feed-forward networks to learn such a mapping. Relative camera rotation R is embedded into a 32-dim vector as a token to specify the target view into which SteerPose transforms the input 2D poses. For each pose, we construct 32-dim input tokens by combining the embeddings of the keypoints with learnable positional encodings that capture the hierarchical structure of the articulated body. The model learns to capture both local joint relationships and its global pose structure by processing these tokens through the self-attention layers, making it inherently robust to partial occlusions.

We train SteerPose in a fully-supervised manner for each class of animals such as humans, quadrupeds, birds, *etc*.. Suppose we have a 3D pose dataset of the target. For each 3D pose, we synthesize random pairs of two views on a unit sphere, and project the 3D pose onto such views. We can synthesize as many random pairs of 2D poses with known relative rotations for each 3D pose as possible, and we can repeat this for the 3D poses in the dataset.

Using these synthesized pairs, we train SteerPose to minimize the average 2D keypoint displacement $\mathcal{L}_{\mathrm{kp}}$ defined as

$$\mathcal{L}_{\mathrm{kp}}(R) = \frac{1}{M} \sum_{i=1}^{M} D(q_i(R), p'_i)\,, \tag{1}$$

where M denotes the number of synthesized training pairs. $q_i(R)$ and p'_i denote the 2D pose transformed by SteerPose given R and the target 2D pose respectively. $D(\cdot)$ computes the dissimilarity between 2D poses as the L2 distance between them. During training, we injected 2D random noise into the input 2D joint positions to mimic ambiguities in joint detection and randomly masked tokens to account for missing joints due to occlusions.

We assume that SteerPose is pretrained for the target class of animals hereafter. That is, given a 2D pose and the relative rotation to a target view, SteerPose can infer the 2D pose in the target view. This point is quantitatively evaluated in Sec. 5.2.

3.2 Joint Extrinsic Camera Calibration and Matching

We simultaneously solve the extrinsic camera calibration and 2D pose matching as an inference-time optimization utilizing a neural network with a pretrained SteerPose and differentiable Sinkhorn-based matching [9] as shown in Fig. 1. Suppose we have two sets of 2D poses $P = \{p_i\}$ $(i = 1, \ldots, N)$ and $P' = \{p'_{i'}\}$ $(i' = 1, \ldots, N')$ of the target estimated from images captured at different viewpoints C and C'. By specifying a relative rotation R from C, the pretrained SteerPose transforms each 2D pose $p_i \in P$ of C to those at another view at rotation R as $q_i(R) \in Q(R)$, and Sinkhorn algorithm can find the best correspondences between the transformed 2D poses $q_i(R) \in Q(R)$ and the observed 2D poses $p'_{i'} \in P'$ $(i' = 1, \ldots, N')$ at the other view C', based on the similarity between the 2D poses. If the specified rotation R is identical to the relative rotation between C and C', we can expect that each 2D pose $q_i(R)$ has a similar pose in P', and the Sinkhorn algorithm can find their correspondences correctly. Note that in the case of $N \neq N'$ due to occlusions, for example, we add dummy targets as described in [34].

Using this network, we realize simultaneous extrinsic calibration and 2D pose matching as optimization by back-propagation. Suppose we start the optimization with an initial estimate of the relative rotation R, and let the matching module find the best correspondences between the 2D poses under the current R. The key idea is to evaluate the geometric consistency of the current rotation R and the estimated correspondences, in addition to the matching cost used in the correspondence search, using the loss function defined later, and to refine R by back-propagating its gradient while keeping the pretrained SteerPose frozen. This is an inference-time optimization, and iteratively refining R simultaneously improves the correspondence search in the matching module.

Matching loss The correspondence search between the 2D pose sets $Q(R)$ and P' using Sinkhorn algorithm in the matching module assigns a weight to each of the possible combinations of the 2D poses between $Q(R)$ and P'. The similarity score for each possible pair of 2D poses $q_j(R)$ and p'_j is defined as $s(q_j(R), p'_j) = \frac{2}{1+\exp(\alpha \mathcal{L}_{\mathrm{kp}}(R))}$, where α is a scaling parameter. Denoting the weight and 2D poses of the jth pair by w_j, $q_j(R) \in Q(R)$, and $p'_j \in P'$, we evaluate the following matching loss to assess the confidence of the correspondence search:

$$\mathcal{L}_{\mathrm{match}} = \sum_{j=1}^{J} w_j (1 - s(q_j(R), p'_j))\,. \tag{2}$$

To enforce cycle consistency between views, we compute the matching loss bidirectionally.

Geometric consistency loss The geometric consistency loss $\mathcal{L}_{\mathrm{geom}}$ evaluates whether the combination of the current R and the estimated matches yields a valid solution of the relative translation, after enforcing collinearity and coplanarity constraints [35, 19]. For all N corresponding 2D pose pairs under current rotation R, we transform their positions into a normalized coordinate system. By stacking linear equations derived from collinearity and coplanarity constraints [19],

Target	# of subjects	5-point [43]			Class-specific SteerPose			Class-agnostic SteerPose		
		RRA@30↑	RTA@30↑	AUC@30↑	RRA@30↑	RTA@30↑	AUC@30↑	RRA@30↑	RTA@30↑	AUC@30↑
Cheetah [39]	1	0.5	0.61	0.43	0.82	0.86	0.58	0.81	0.85	0.57
Bama Pig [3]	4	0.64	0.77	0.56	0.88	0.89	0.76	0.92	0.97	0.81
Beagle Dog [3]	2	0.47	0.61	0.42	0.96	0.98	0.79	0.96	0.96	0.79
Pigeon [1]	5	0.77	0.81	0.58	0.94	0.94	0.75	N/A	N/A	N/A
Toddler [13]	4	0.4	0.54	0.35	0.68	0.76	0.52	N/A	N/A	N/A

Table 1: Comparison to a conventional geometry-based pose estimation method. Both the class-specific and class-agnostic models outperform the baseline method.

we can obtain a linear system of the form $A\,[x_1 \quad \dots \quad x_N \quad \boldsymbol{t}]^\top = \mathbf{0}$, where unknowns x_i are 3D poses recovered from the collinearity constraint and $\boldsymbol{t}$ is determined up to scale and sign ambiguity. This sign ambiguity is resolved via a chirality check [35].

Since the partial derivatives of the singular values $\sigma_2 \geq \sigma_1$ can be computed in closed-form [36, 29], we include the ratio σ_1/σ_2 as a loss $\mathcal{L}_{\text{geom}}$ in the optimization of R, to guarantee that the estimated R makes the smallest singular value as small as possible, and hence it has a solution of $\boldsymbol{t}$.

Consequently, we optimize $\mathcal{L}_{\text{match}} + \lambda\mathcal{L}_{\text{geom}}$ as the loss function, where λ is an empirically determined weighting parameter.

4 Multi-view Extrinsic Calibration

In the case of multiple cameras, we can simply calibrate them in a pair-wise manner as described in the previous section, and then integrate them into a unified coordinate system by employing motion averaging [37] and cycle consistency-based matching [38]. The integrated extrinsic parameters can be further optimized by a nonlinear optimization that minimizes the reprojection errors [35].

5 Experiments

5.1 Experimental Setups

Datasets We evaluate on five animal datasets: *Cheetah* from Acinoset [39], *Bama Pig* and *Beagle Dog* from MAMMAL [3], and *Pigeon* from 3D-POP [1]. For human subjects, we include *Toddler* from CMU Panoptic dataset [13] and *Volleyball* from EgoHumans dataset [40]. Except for *Cheetah*, all datasets contain multiple instances in the scene, as shown in Fig. 3. To evaluate our method, we employed both synthetic and real-world datasets. For the former, we project ground-truth 3D poses from these datasets into multi-view images and inject Gaussian noise with a standard deviation of 3 pixels to generate noisy 2D keypoints. For the latter, we utilize the 2D poses obtained from off-the-shelf pose estimators [41, 42] that are tailored to each animal class.

Metrics We assess the camera pose error using rotation error E_R and translation error $E_{\mathbf{t}}$, both expressed as angular differences in degrees, and report the median projection error E_{2D} in pixels between the reprojected 3D keypoints and their corresponding 2D observations. For the matching evaluation, we measure the precision P using the ground-truth instance correspondences across views. We further evaluate the precision of the estimated camera pose using three metrics: the relative rotation accuracy (RRA), relative translation accuracy (RTA), and area under curve (AUC) [31]. For a given threshold τ, AUC@τ is computed as the minimum of RRA@τ and RTA@τ.

5.2 Comparison with Existing Methods

Geometric Method Table 1 presents a comparison of SteerPose with the standard five-point method [43], which is a point-correspondence-based approach. We employ the RANSAC algorithm to define correspondences for the five-point method in multi-instance scenarios. This evaluation includes a diverse set of five species, including birds and humans. For quadruped species, we compared the results of class-specific and class-agnostic models.

The five-point algorithm struggles with pose estimation failures in challenging scenarios, particularly in degenerate configurations where the calibration targets are coplanar or when dealing with wide-baseline conditions. In contrast, SteerPose demonstrates a robust performance across these scenarios and effectively handles species with varying skeletal structures beyond quadrupeds. Both class-specific and class-agnostic SteerPose outperform the standard

Target	# of Subject	LightGlue [27]			MASt3R [32]			Ours		
		RRA@20↑	RTA@20↑	AUC@20↑	RRA@20↑	RTA@20↑	AUC@20↑	RRA@20↑	RTA@20↑	AUC@20↑
Cheetah [39]	1	0.30	0.25	0.30	0.44	0.59	0.42	0.62	0.89	0.48
		0.30	0.24	0.30	0.44	0.58	0.42	0.41	0.55	0.26
Bama Pig [3]	4	0.24	0.23	0.22	0.96	0.96	0.88	0.99	0.99	0.85
		0.38	0.38	0.35	0.89	0.90	0.81	0.77	0.73	0.45
Beagle Dog [3]	2	0.31	0.38	0.30	1.00	1.00	0.92	0.85	0.97	0.77
		0.30	0.41	0.29	1.00	1.00	0.94	0.76	0.76	0.47
Pigeon [1]	5	0.02	0.13	0.07	0.78	0.78	0.64	0.98	1.00	0.75
Toddler [13]	4	0.86	0.78	0.78	0.99	0.99	0.89	0.73	0.75	0.66
		0.83	0.77	0.76	1.00	1.00	0.90	0.75	0.75	0.66
Volleyball [40]	4	0.64	0.88	0.78	1.00	1.00	0.97	0.94	0.95	0.91
		0.64	0.88	0.78	1.00	1.00	0.97	0.77	0.86	0.72

Table 2: Evaluation of two-view camera pose estimation. Gray-shaded rows use the ground-truth 2D poses with Gaussian noise with $\sigma = 3$px, while white rows utilize 2D poses from off-the-shelf estimators. Our method achieves a performance comparable to that of SOTA methods.

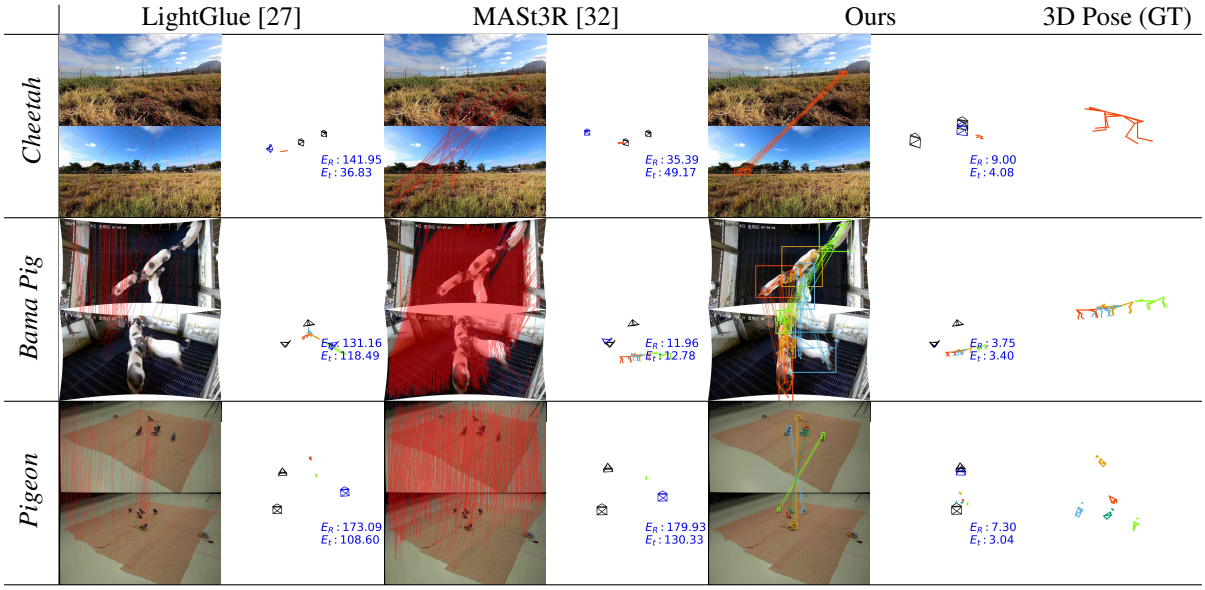


Figure 3: Two-view extrinsic calibration and matching results. The ground-truth and estimated camera poses are shown in red and green, respectively. Our method reliably establishes correspondences and accurately estimates camera poses for 3D reconstruction.

correspondence-based methods, demonstrating their versatility and reliability in pose estimation and matching. For detailed information about the class-agnostic SteerPose, please see Appendix A.

Learning-based Methods Table 2 and Fig. 3 compare our method with the state-of-the-art learning-based methods. SuperPoint [25] and LightGlue [27] leverage learned priors for keypoint detection, description, and matching in 2D images. The relative poses are then estimated from these matches using RANSAC [30] and the essential matrix decomposition [35]. Then, a bundle adjustment [35] is employed to refine the poses. MASt3R [32] simultaneously performs 3D scene reconstruction and matching based on monocular 3D point-map estimation [31].

The results show that SteerPose outperforms Lightglue, both quantitatively and qualitatively. Our method, which leverages structured keypoints obtained from articulated poses, achieves accuracy comparable to that of MASt3R. As shown in Fig 3, our method is more robust in challenging scenarios, such as those with poor background textures or wide baselines between cameras (*e.g.*, *Cheetah* and *Pigeon*), where MASt3R often struggles.

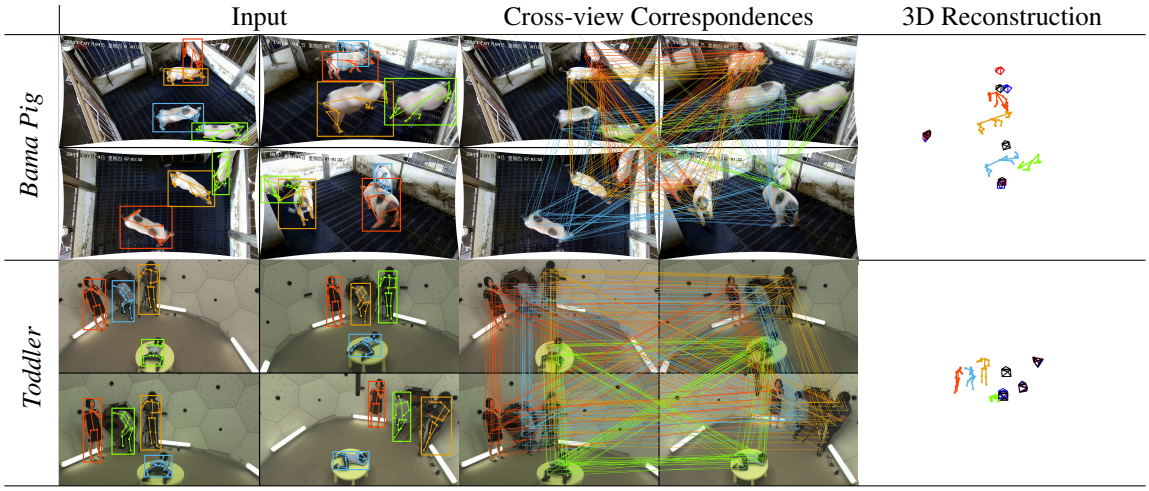


Figure 4: Visualization for multi-view calibration results. The ground-truth cameras are in black, the initial camera poses after registration are in blue, and the final bundle-adjustment results are shown in red. Our method establishes reliable correspondences and accurately estimates camera poses for 3D reconstruction. For quantitative evaluation, please refer to Table 6.

	Bama Pig [3]			*Pigeon* [1]		
	$E_R \downarrow$	$E_\mathbf{t} \downarrow$	$P \uparrow$	$E_R \downarrow$	$E_\mathbf{t} \downarrow$	$P \uparrow$
w/o $\mathcal{L}_{\text{geom}}$	11.35	7.78	0.83	20.70	12.02	0.99
w $\mathcal{L}_{\text{geom}}$	8.30	4.71	0.90	9.20	5.70	0.99

Table 3: Contribution of the geometric consistency loss $\mathcal{L}_{\text{geom}}$.

5.3 Multi-view Integration

Fig. 4 illustrates the multi-view camera calibration results. We utilize RANSAC [30] to identify the best camera poses for each pair of views by evaluating the geometric consistency based on the reprojection error. For class-specific models, we use the results in Table 2. By integrating the pairwise camera poses and cross-view correspondences into a unified coordinate system, we further refine them through a nonlinear optimization of the reprojection errors. This highlights the capability of our method to automatically achieve extrinsic camera calibration, correspondence finding, and 3D reconstruction as a markerless motion capture within a single framework by simply capturing novel articulated animals in uncontrolled environments.

5.4 Ablation Study

Table 3 reports the contribution of our geometric consistency loss $\mathcal{L}_{\text{geom}}$. We use two class-specific models trained on quadrupeds (*Bama Pig*) and bipeds (*Pigeon*) to demonstrate the effectiveness of this loss term. We observed that incorporating $\mathcal{L}_{\text{geom}}$ significantly improves extrinsic calibration and matching accuracy by enforcing keypoint geometric consistency in both scenarios.

6 Conclusion

We proposed SteerPose which learns to *rotate* an input 2D pose to match one in another view at a relative rotation R. SteerPose leverages the articulation of 3D poses as prior knowledge, and functions as $g'(\cdot)$ which performs *mental rotation* to form a rotation-covariant transform of 3D poses under projection $f(\cdot)$ satisfying $f(g(\cdot)) = g'(f(\cdot))$ where $g(\cdot)$ represents the 3D rotation by R.

Our method combines pretrained SteerPose with the Sinkhorn algorithm to perform simultaneous extrinsic calibration and correspondence search as an optimization problem during inference. The novel geometric consistency loss explicitly guarantees that the optimized rotation and match have a valid solution for relative translation. Experimental results on diverse in-the-wild datasets of humans and animals demonstrated the effectiveness and robustness of the proposed

method. In particular, by leveraging the articulation of human or animal poses, our calibration approach effectively addresses challenging wide-baseline configurations and textureless backgrounds, which are common in uncontrolled environments such as those involving animals in the wild. Furthermore, we showed that our class-agnostic model enables the 3D reconstruction of novel animals in multi-view setups.

Our proposed calibration requires only 2D pose estimation or annotation, which is becoming widely available in many contexts. We believe that the proposed method can be used as a practical calibration tool for studies involving on-site multi-view 3D captures, and can provide triangulated 3D poses for training monocular 3D pose estimators of novel targets.

References

[1] Hemal Naik, Alex Hoi Hang Chan, Junran Yang, Mathilde Delacoux, Iain D Couzin, Fumihiro Kano, and Máté Nagy. 3d-pop-an automated annotation approach to facilitate markerless 2d-3d tracking of freely moving birds with marker-based motion capture. In *CVPR*, pages 21274–21284, 2023.

[2] Urs Waldmann, Alex Hoi Hang Chan, Hemal Naik, Máté Nagy, Iain D Couzin, Oliver Deussen, Bastian Goldluecke, and Fumihiro Kano. 3d-muppet: 3d multi-pigeon pose estimation and tracking. *IJCV*, pages 1–18, 2024.

[3] Liang An, Jilong Ren, Tao Yu, Tang Hai, Yichang Jia, and Yebin Liu. Three-dimensional surface motion capture of multiple freely moving pigs using mammal. *Nature Communications*, 14(1):7727, 2023.

[4] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87:4–27, 2010.

[5] Tanmay Nath, Alexander Mathis, An Chi Chen, Amir Patel, Matthias Bethge, and Mackenzie Weygandt Mathis. Using deeplabcut for 3d markerless pose estimation across species and behaviors. *Nature protocols*, 14(7): 2152–2176, 2019.

[6] Jessy Lauer, Mu Zhou, Shaokai Ye, William Menegas, Steffen Schneider, Tanmay Nath, Mohammed Mostafizur Rahman, Valentina Di Santo, Daniel Soberanes, Guoping Feng, et al. Multi-animal pose estimation, identification and tracking with deeplabcut. *Nature Methods*, 19(4):496–504, 2022.

[7] Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972): 701–703, 1971.

[8] Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *ICCV*, Oct 2017.

[9] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *NeurIPS*, 26, 2013.

[10] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE TPAMI*, 22(11):1330–1334, 2000.

[11] S. Garrido-Jurado, R. Muñoz-Salinas, F.J. Madrid-Cuevas, and M.J. Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280–2292, 2014.

[12] Joel Mitchelson and Adrian Hilton. Wand-based multiple camera studio calibration. *CVSSP Technical Report*, 2003.

[13] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE TPAMI*, 2017.

[14] Nils Hasler, Bodo Rosenhahn, Thorsten Thormahlen, Michael Wand, Jürgen Gall, and Hans-Peter Seidel. Markerless motion capture with unsynchronized moving cameras. In *CVPR*, 2009.

[15] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L Schönberger. Global structure-from-motion revisited. In *ECCV*, 2024.

[16] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016.

[17] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, November 2004.

[18] Takafumi Taketomi, Hideaki Uchiyama, and Sei Ikeda. Visual slam algorithms: A survey from 2010 to 2016. *IPSJ Transactions on Computer Vision and Applications*, 9(1):16, 2017.

[19] Sang-Eun Lee, Keisuke Shibata, Soma Nonaka, Shohei Nobuhara, and Ko Nishino. Extrinsic camera calibration from a moving person. *IEEE RA-L*, 7(4):10344–10351, 2022. doi:10.1109/LRA.2022.3192629.

[20] Kosuke Takahashi, Dan Mikami, Mariko Isogawa, and Hideaki Kimata. Human pose as calibration pattern; 3d human pose estimation with multiple unsynchronized and uncalibrated cameras. In *Proc. CVPRW*, 2018.

[21] Yan Xu, Yu-Jhe Li, Xinshuo Weng, and Kris Kitani. Wide-baseline multi-camera calibration using person re-identification. In *CVPR*, 2021.

[22] Gaku Nakano. Camera calibration using parallel line segments. In *Proc. ICPR*, pages 1505–1512, 2021. doi:10.1109/ICPR48806.2021.9412183.

[23] Nicola Garau, Francesco G. B. De Natale, and Nicola Conci. Fast automatic camera network calibration through human mesh recovery. *Journal of Real-Time Image Processing*, 2020.

[24] Sang-Eun Lee, Ko Nishino, and Shohei Nobuhara. Spatiotemporal multi-camera calibration using freely moving people. *IEEE Robotics and Automation Letters*, 10(5):4818–4825, 2025.

[25] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPRW*, pages 224–236, 2018.

[26] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, pages 4938–4947, 2020.

[27] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *ICCV*, pages 17627–17638, 2023.

[28] Georg Bökman, Johan Edstedt, Michael Felsberg, and Fredrik Kahl. Steerers: A framework for rotation equivariant keypoint descriptors. In *CVPR*, pages 4885–4895, 2024.

[29] Barbara Roessle and Matthias Nießner. End2end multi-view feature matching with differentiable pose optimization. In *ICCV*, pages 477–487, 2023.

[30] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *CACM*, 24(6):381–395, 1981.

[31] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024.

[32] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r. *arXiv:2406.09756*, 2024.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[34] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

[35] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[36] Catalin Ionescu, Orestis Vantzos, and Cristian Sminchisescu. Matrix backpropagation for deep networks with structured layers. In *ICCV*, 2015.

[37] Venu Madhav Govindu. Combining two-view constraints for motion estimation. In *CVPR*, 2001.

[38] Junting Dong, Qi Fang, Wen Jiang, Yurou Yang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation and tracking from multiple views. *IEEE TPAMI*, 44(10):6981–6992, 2022.

[39] Daniel Joska, Liam Clark, Naoya Muramatsu, Ricardo Jericevich, Fred Nicolls, Alexander Mathis, Mackenzie W Mathis, and Amir Patel. Acinoset: A 3d pose estimation dataset and baseline models for cheetahs in the wild. In *ICRA*, pages 13901–13908, 2021.

[40] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Egohumans: An egocentric 3d multi-human benchmark. *arXiv preprint arXiv:2305.16487*, 2023.

[41] Shaokai Ye, Anastasiia Filippova, Jessy Lauer, Steffen Schneider, Maxime Vidal, Tian Qiu, Alexander Mathis, and Mackenzie Weygandt Mathis. Superanimal pretrained pose estimation models for behavioral analysis. *Nature communications*, 15(1):5165, 2024.

[42] Peng Lu, Tao Jiang, Yining Li, Xiangtai Li, Kai Chen, and Wenming Yang. RTMO: Towards high-performance one-stage real-time multi-person pose estimation, 2023.

[43] David Nistér. An efficient solution to the five-point relative pose problem. *IEEE TPAMI*, 26(6):756–770, 2004.

[44] Jiacong Xu, Yi Zhang, Jiawei Peng, Wufei Ma, Artur Jesslen, Pengliang Ji, Qixin Hu, Jiehua Zhang, Qihao Liu, Jiahao Wang, et al. Animal3d: A comprehensive dataset of 3d animal pose and shape. In *ICCV*, pages 9099–9109, 2023.

[45] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. In *3DV*, pages 106–115, 2024.

Appendices

Appendix A Generalizability of SteerPose

Table 4 reports the test loss of SteerPose trained under different conditions. The "class-specific" column represents SteerPose trained on the training set of the target dataset itself. The "class-agnostic" columns represent SteerPose trained on Animal3D dataset [44], which contains 40 quadruped species. The results indicate that SteerPose, when trained for quadrupeds as a class-agnostic model, performs comparably well on the same animals from a different dataset, *Beagle Dog*, and generalizes effectively to unseen species, *i.e.*, species not included in Animal3D, such as *Cheetah* and *Bama Pig*.

Note that there were small discrepancies between the skeleton structures of *Beagle Dog* and Animal3D. For example, Animal3D has a neck, whereas *Beagle Dog* does not. To evaluate the class-agnostic SteerPose trained with Animal3D using the test set from *Beagle Dog*, we used SteerPose itself to align the skeleton structure by applying identity rotation to the 2D poses. That is, by masking the missing 2D joints in the input, rotating 2D poses to the same viewpoint by SteerPose can estimate the 2D keypoint positions of the masked joints. This can also fill in missing keypoints due to occlusions. This "keypoint completion" step can enhance the generalizability of the class-agnostic model as shown in the two "Class-agnostic" columns in Table 4.

Table 6 also demonstrates that the class-agnostic model achieves comparable results in multi-view calibration and cross-view correspondence matching tasks against the class-specific models, as detailed in Sec. C.

Appendix B Comparison with Other Methods

Table 5 summarizes the advantages of leveraging articulation in our approach for extrinsic camera calibration. As shown in Fig. 3 of the main text, the 2D poses as structured correspondences can realize robust calibrations, in particular for wide-baseline scenarios.

Appendix C Additional Calibration Results

Fig. 5 shows a qualitative comparison of our two-view calibration and matching results, demonstrating a consistent performance across different scenarios. Table 6 presents quantitative multi-camera calibration results for both the class-specific and agnostic models. This nearly identical performance highlights the effectiveness and generalizability of our approach to unseen animal species during training. The qualitative multi-view calibration results are shown in Fig. 6. Our method refines camera poses through the nonlinear optimization of reprojection errors within a unified coordinate system, leading to accurate camera poses and 3D reconstructions.

Appendix D Training Details and Implementation

D.1 Training Data Generation

Training data were generated by sampling diverse camera viewpoints relative to the 3D poses of the target. Specifically, we placed 100 cameras uniformly over the unit hemisphere using the Fibonacci sphere method, with their orientations directed towards the origin. To model the camera roll variations, we uniformly sampled 20 rotations around the optical axis of the camera from each viewpoint. We randomly sampled 3,000 camera pairs and applied orthographic projections to obtain 2D poses for both views. Each training pair consists of 2D-2D pose correspondences and the associated relative camera rotation. The synthesized pairs are divided into training, validation, and test sets in a ratio of 7:2:1. To enhance the robustness against partial occlusions, we augmented the training data by randomly masking 10% to 30% of the joints in the 2D poses. Table 7 summarizes the details of the datasets used in the experiments.

D.2 Network architecture of SteerPose

Fig. 7 illustrates the network architecture of the proposed SteerPose. The input consists of 2D positions of N joints in the image and three parameters representing the rotation by the Rodrigues vector. Each of the N 2D joint positions and the rotation vector are projected onto the 32-dimensional vectors by MLP layers.

Each of the 32-dimensional vectors is combined with a positional encoding by adding them element-wise. Here we used a look-up embedding with $N + 1$ tokens for positional encoding. The embeddings with the positional encoding are used

Target	Class-specific	Class-agnostic*	Class-agnostic
Cheetah [39]	0.098	0.217	0.202
Bama Pig [3]	0.121	0.21	0.173
Beagle Dog [3]	0.166	0.208	0.16

Table 4: Generalizability of SteerPose. Both "class-specific" and "class-agnostic" columns list the average keypoint errors for the target test set. The class-agnostic SteerPose generalizes well to novel animals. * indicates results obtained without applying keypoint completion.

Method	Correspondence Search	Joint Optimization	Structured Correspondence
RelPose++ [45]	×	×	×
SuperGlue [26]	✓	×	×
Diff. 8-Pts [29]	✓	✓	×
MASt3R [32]	✓	✓	×
Ours	✓	✓	✓

Table 5: Comparison of key properties of different camera pose estimation methods. Use of the target 2D poses as structured correspondences can contribute to both calibration and matching.

as input tokens of the transformer encoder with five layers. The output tokens of the transformer encoder are averaged together to form a single 32-dimensional vector by mean pooling and then transformed to N 2D joint positions.

D.3 Implementation Details

We used the Adam optimizer with a learning rate of 0.01 for 1000 iterations. We schedule the learning rate using a linear scheduler with a start factor of 1.0 and an end factor of 0.01. The geometric consistency loss is activated after 40% of the total optimization iterations and once a reliable set of multiple bi-directional matches has been established. The scale parameter $\alpha = 3$ was used to calculate the similarity score.

For each scenario, we sampled 100-120 image pairs from multi-view sequences. The calibration success was determined by a reprojection error of less than 10 pixels for the reconstructed 3D poses. Otherwise, we retried the calibration with a new random initial rotation, up to a maximum of 5 attempts.

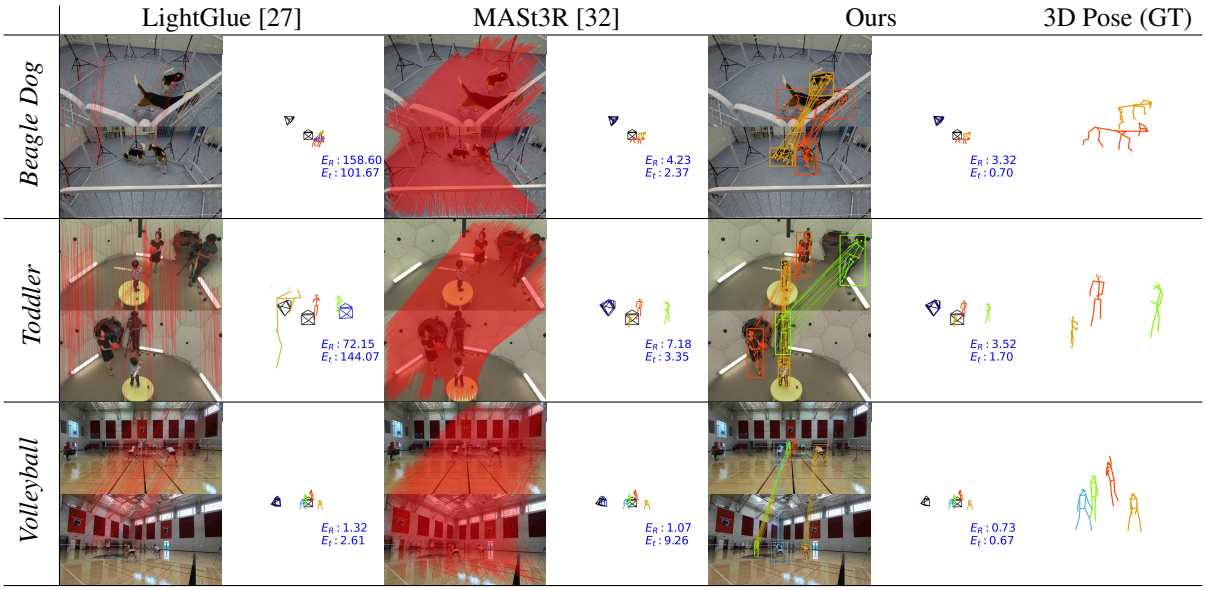


Figure 5: Two-view extrinsic calibration and matching results. Colored lines across the views represent the estimated correspondences. The ground truth and estimated camera poses are shown in red and green, respectively. The 3D poses were triangulated using the estimated camera poses.

Scenario	Camera IDs	Motion Averaging [37]			Bundle Adjustment [35]		
		E_R	$E_\mathbf{t}$	E_{2D}	E_R	$E_\mathbf{t}$	E_{2D}
Bama Pig [3]	[0,1,5,9]	2.20	4.16	5.86	2.42	3.73	3.23
		4.39	4.91	4.86	2.43	3.75	3.23
Beagle Dog [3]	[4,6,7,9]	7.50	5.24	10.10	5.93	4.10	3.09
		7.39	4.90	9.47	5.97	4.11	3.09
Toddler [13]	[7,12,22,28]	0.85	1.07	2.62	0.29	0.38	1.76

Table 6: Evaluation of extrinsic calibration and matching in multi-view setups. Gray-shaded cells denote the class-agnostic results.

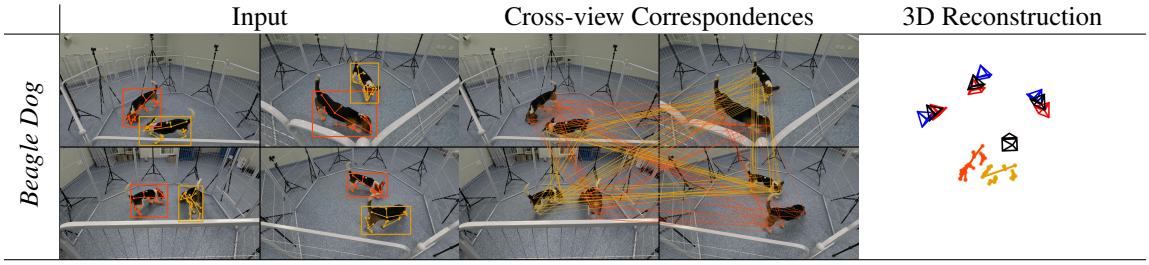


Figure 6: Visualization for multi-view calibration results. The inputs (left) show detected 2D poses of multiple instances from off-the-shelf pose estimators, with the colored bounding boxes indicating the instance IDs. Cross-view correspondences are shown as colored lines (Middle). Multi-view integration (Right) refines the camera poses and reconstructs the 3D poses. The ground-truth cameras are in black, the initial camera poses after motion averaging are in blue, and the final bundle-adjustment results are shown in red.

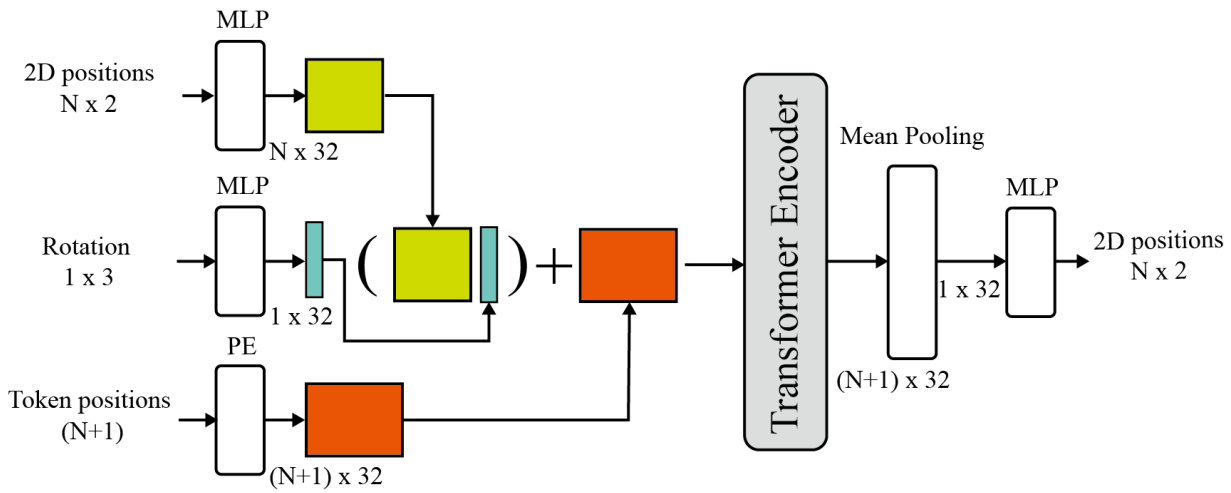


Figure 7: Network architecture of SteerPose

	Train	Calibration (Synthetic)	Calibration (Real)
Cheetah [39]	Jules flick1 (20190309)	Romeo flick (20190227)	Romeo flick (20190227)
Bama Pig [3]	0–1400	1400–1750	0–1750 (HRNet)
Beagle Dog [3]	0–80	80–112	0–112 (SuperAnimal [41])
Pigeon [1]	Sequence8_n01_01072022	Sequence5_n05_01072022	Sequence5_n05_01072022
Toddler [13]	171204_pose1–6	170915_toddler5	170915_toddler5 (RTMO [42])
Volleyball [40]	171204_pose1–6	001_volleyball	001_volleyball (RTMO [42])

Table 7: Dataset for training and calibration in synthetic and real scenarios, where the numbers indicate the frame indices.