applied sciences

MDPI

Article

Animal Pose Estimation Based on 3D Priors

Xiaowei Dai [1], Shuiwang Li [2], Qijun Zhao [1,3,*] and Hongyu Yang [1,3,*]

[1] National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu 610065, China
[2] College of Information Science and Engineering, Guilin University of Technology, Guilin 541006, China
[3] College of Computer Science, Sichuan University, Chengdu 610065, China
* Correspondence: qjzhao@scu.edu.cn (Q.Z.); yanghongyu@scu.edu.cn (H.Y.)

Abstract: Animal pose estimation is very useful in analyzing animal behavior, monitoring animal health and moving trajectories, etc. However, occlusions, complex backgrounds, and unconstrained illumination conditions in wild-animal images often lead to large errors in pose estimation, i.e., the detected key points have large deviations from their true positions in 2D images. In this paper, we propose a method to improve animal pose estimation accuracy by exploiting 3D prior constraints. Firstly, we learn the 3D animal pose dictionary, in which each atom provides prior knowledge about 3D animal poses. Secondly, given the initially estimated 2D animal pose in the image, we represent its latent 3D pose with the learned dictionary. Finally, the representation coefficients are optimized to minimize the difference between the initially estimated 2D pose and the 2D projection of the latent 3D pose. Furthermore, we construct 2D and 3D animal pose datasets, which are used to evaluate the algorithm's performance and learn the 3D pose dictionary, respectively. Our experimental results demonstrate that the proposed method makes good use of the 3D pose knowledge and can effectively improve 2D animal pose estimation.

Keywords: animal pose estimation; dictionary learning; 3D constraints

check for updates

Citation: Dai, X.; Li, S.; Zhao, Q.; Yang, H. Animal Pose Estimation Based on 3D Priors. *Appl. Sci.* **2023**, *13*, 1466. https://doi.org/10.3390/app13031466

Academic Editor: Byung-Gyu Kim

Received: 8 November 2022
Revised: 16 January 2023
Accepted: 19 January 2023
Published: 22 January 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, animal pose estimation has attracted more and more attention owing to its potential applications in biology, neuroscience, ecology, agriculture, and entertainment [1–4]. For instance, animal pose estimation [5–7] could be employed in markerless motion capture systems to capture the movements of wild animals. Animal pose estimation [8,9] would also support advances in entertainment, where animal animation is still largely performed manually. In neuroscience [3,10], tracking animals is fundamental for understanding behavior or relating motion to brain activity. In bio-inspired robotics [11,12], understanding how animals move can help to design more efficient robots.

Despite the promising applications in various fields, there are few works on animal pose estimation to date. This has much to do with the challenges posed by the following factors. Firstly, research on animal pose estimation is largely limited by data scarcity. Typical human pose estimation benchmarks, such as Human3.6M [13], capture data from 15 sensors and rely on small markers attached to the subject's body to obtain motion. However, such data acquisition methods are hard to be applied to animals because it is impractical to ask animals to cooperate during data acquisition, especially in wild conditions. Secondly, due to occlusions, complex backgrounds, and unconstrained illumination conditions in the wild, the quality of the collected images is often uncontrollable, resulting in unreliable animal pose estimation. Thirdly, the variations in animal poses are far more complicated than in humans. The high deformations of animal pose (or body) make annotating and predicting animal pose quite challenging.

Most existing methods for animal pose estimation focus on addressing data scarcity by transferring knowledge from other more accessible domains such as synthetic animal

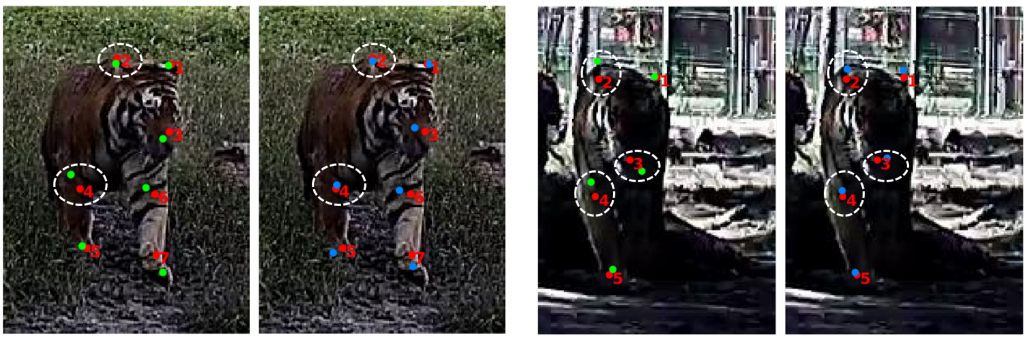
data [6,14–16] or human data [17]. However, there are domain gaps [18,19] between synthetic images and real images, which prevent models trained on synthetic data from generalizing well to real-world images. Furthermore, 2D poses are the projection of 3D body configurations, during which the relationship between the joints in 3D poses is distorted. Consequently, the models that are trained with 2D annotations alone may not conform to the real configurations of animal poses. This problem can become severe, especially in wild-animal images with uncontrolled quality, for which large errors in animal pose estimation would be obtained using existing methods.

In general, the performance can be improved by preprocessing (such as image segmentation [20,21], image enhancement [22], etc.) or postprocessing. However, the preprocessing of wildlife images, such as denoising or enhancement, is a non-trivial task in itself. In this paper, we propose a method for estimating 2D animal poses with 3D priors, in which 3D priors are constructed with synthetic 3D poses and encoded in a 3D pose dictionary using sparse dictionary learning. As can be seen in Figure 1, taking tigers as an example, the proposed method can estimate more accurate 2D poses in poor-quality images.

A preliminary version of this work has been published in the 33rd British Machine Vision Conference (BMVC) [23]. This paper extends the BMVC paper in the following aspects: (i) We extensively compare the proposed algorithm with more algorithms. (ii) The proposed method is also evaluated on more datasets, which demonstrates its good generalization. (iii) This paper considers more widely distributed noise for more comprehensive and detailed performance evaluation and analysis. (iv) We implement the proposed method through both traditional optimization and deep learning. The contributions of this paper can be summarized as follows:

- We propose a novel method for refining 2D animal pose estimation using 3D priors, which can be easily incorporated into existing 2D pose estimation methods;
- We present both conventional optimization and learning-based neural networks to implement the proposed method;
- We build a 3D animal pose dataset and manually annotate a 2D pose dataset for animal pose estimation;
- Extensive experiments are conducted to evaluate the proposed method. The experimental results show that the proposed method is effective in improving 2D animal pose estimation accuracy.

The remainder of this paper is organized as follows: We discuss related work and motivation in Section 2. Section 3 introduces the establishment of the relevant datasets. Section 4 provides details of the proposed method. Our experimental evaluation results are presented in Section 5. Finally, we finish the paper with concluding remarks in Section 6.



(**a**) (**b**)

Figure 1. (**a**,**b**) are examples from the Amur dataset. The red point represents ground truth, which is annotated by us according to the ATRW [24] definition. The number represents the index of the joint [24]. The green point represents the initial pose estimated by HRNet [25]. The blue point represents the refined pose, which is obtained via the postprocessing of the initial estimation using the proposed method. The dashed circles indicate significant refinement.

2. Related Work

2.1. The 3D Human Pose Estimation Methods

2.1.1. Deep-Learning-Based Human Pose Estimation

Thanks to deep learning and large-scale datasets, significant progress has been made in 3D human pose estimation. The existing deep-learning-based human pose estimation methods can be divided into the following categories:

Full Supervision. Supervised approaches rely on large datasets that contain millions of images with corresponding 3D pose annotations. Martinez et al. [26] trained a neural network on 2D poses and corresponding 3D ground truth. Due to its simplicity, it can be quickly trained for a large number of epochs leading to high accuracy and serves as a baseline for many following approaches. Recently, Transformer [27] equipped with a global self-attention mechanism has also become popular in 3D pose estimation [28,29]. However, the major downside of all supervised methods is that they do not generalize well to images with unseen poses.

Weak Supervision. Weakly supervised approaches require only a small set of annotated 3D poses or unpaired 2D and 3D poses [30–32]. Wandt et al. [30] used a discriminator to represent the prior distribution of 3D poses. Meanwhile, in order to avoid the over-fitting problem that may be caused by using only the projection error, a camera network was also introduced to estimate the camera parameters. Based on [30], Li et al. [31] introduced a random variable to explicitly generate multiple 3D poses from a 2D input and select the optimal 3D pose from them. Wandt et al. [33] decomposed the observed 2D pose into a 3D pose and camera rotation using multi-view consistency constraints. Compared with completely supervised approaches, these weakly supervised methods generalize and transfer better to new domains. However, they still struggle with poses that are very different from the annotated training set.

Unsupervised. Drover et al. [34] proposed an unsupervised learning approach to monocular human pose estimation. They randomly projected an estimated 3D pose back to 2D. Chen et al. [35] extended [34] with a cycle consistency loss that was computed by lifting the randomly projected 2D pose to 3D and inversing the previously defined random projection. Yu et al. [36] further introduced a learnable scaling factor for the input 2D poses. Wandt et al. [37] improved their results by learning the camera distribution of the training set. The unsupervised methods do not generalize better than supervised ones, but they allow training models without 3D priors.

2.1.2. Dictionary-Based Human Pose Estimation

Dictionary-based methods are also widely used in human pose estimation [38,39]. In these methods, a 3D pose is defined by a set of joints and is assumed to be represented by a linear combination of predefined pose bases and sparse coefficients. Given the 2D correspondence of the joints in a single image, the calculation problem is to simultaneously estimate the coefficients of the sparse representation as well as the viewpoint of the camera. For example, the authors of [40] proposed a sparse representation-based approach to estimate a 3D human pose from 2D annotations in a single image. They presented a projected matching pursuit algorithm for reconstructing 3D poses and camera settings by minimizing the reprojection error. Wang et al. [41] proposed to estimate the 3D pose by minimizing an L1-norm penalty between the projection of 3D joints and 2D detections to reduce the impact of inaccurate 2D pose estimations. Zhou et al. [38] adopted an augmented 3D shape model to achieve a linear representation of shape variability in 2D and proposed to use spectral-norm regularization to penalize invalid cases caused by augmentation. Akhter et al. [42] integrated joint-angle limits into the sparse representation to reduce the possibility of invalid reconstruction. Such methods have achieved promising results in 3D human pose estimation, inspiring us to exploit a pose dictionary in animal pose estimation.

2.2. Animal Pose Estimation

Recently, deep-learning-based approaches [25,43–45] have made significant progress in human pose estimation. Attributed to these methods, the studies of [1] and [46] are the first to use convolutional neural networks [25,43,47] for animal pose estimation. Mathis et al. [1] developed a pose estimation model, which they called DeepLabCut, by modifying a previously published human pose model called DeeperCut [48]. The DeepLabCut model, like the DeeperCut model, was built on the popular ResNet architecture [47], which is one of the most advanced deep-learning models for image classification. Pereira et al. [46] implemented a modified version of a model called SegNet [49], which they called LEAP. Graving et al. [2] presented a multi-scale deep-learning model, called DeepPoseKit, which is based on the DenseNet [50].

It is appealing to directly estimate animal pose via extending the deep-learning algorithms designed for human subjects, such as DeepLabCut [1], LEAP [46], and DeepPoseKit [2]. These models have been widely used for laboratory animals (e.g., fruit flies or mice). However, they are hardly applicable for non-laboratory animals due to the higher cost of data acquisition and annotations. To solve data limitations, the authors of [17] proposed a cross-domain adaptation scheme to learn a shared feature space between human and animal images, so that their network can learn from existing human pose datasets. Furthermore, they also designed a "progressive pseudo-label-based optimization" (PPLO) to improve model performance by bringing the target domain data into training with pseudo-labels. Mu et al. [15] used the synthetic animal data generated from CAD models to train their model, which was then used to generate pseudo-labels for unlabeled real animal images. To handle noisy pseudo-labels, they designed three consistency constraints to evaluate the quality of the predicted labels. Li et al. [16] designed a multi-scale domain adaptation module (MDAM) to reduce the domain gap between the synthetic and real data. Furthermore, a coarse-to-fine pseudo-label updating strategy was further introduced to gradually replace noisy pseudo-labels with more accurate ones during training.

In addition, there are also several works focusing on shape reconstruction. Kanazawa et al. [51] learned the deformation and stiffness of 3D animal meshes from manually annotated 2D images. Zuffi et al. [14] introduced the skinned multi-animal linear model (SMAL), which is similar to the popular skinned multi-person linear model (SMPL) [52] for humans. Zuffi et al. [6] fit the SMAL model to several images of the same animal and then refined the shape to better fit the image data, so as to obtain shape and texture (SMAL with Refinement, SMALR). These algorithms [6,14,51] also require ground-truth pose to train the models.

Despite the improvements made by these approaches [1,2,15–17,46], animal pose estimation is still non-trivial. These methods are trained on images with 2D pose annotations, which may not follow the real distribution of animal poses and very likely degrade their performance, especially for poor-quality images. In contrast to these methods, this paper aims to explore the 3D knowledge of animals for refining estimated animal poses such that the misaligned key points caused by image noise can be amended toward their true positions. Inspired by [38–42], we used a 3D pose dictionary to encode 3D prior constraints, which is simple and effective. Note that previous studies [38–42] require large-scale real 3D human poses as the training data for the dictionary, but collecting real 3D animal pose data is very difficult if not impractical. To address the lack of 3D animal pose data, we collected and synthesized data for 3D pose dictionary learning.

3. Dataset Collection

3.1. Cat

We built a cat dataset by collecting images on the Internet, which contained more than 400 images. This dataset, called Cat, was used to learn a 3D pose dictionary. The details are as follows: First, we defined the key points with reference to [51] and annotated them manually (as illustrated in Figure 2 and Table 1). Second, we used [51] to synthesize their deformations. Lastly, the joints on the 3D deformed shapes could form a 3D pose, which is

explained below. These joints were defined with reference to ATRW [24]. As illustrated in Figure 3, there were 15 joints related to ears, nose, shoulders, paws, hips, knees, root of tail, and center (i.e., the midpoint of nose and root of tail).

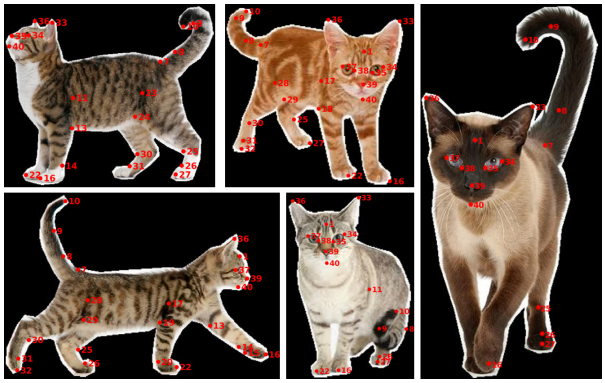
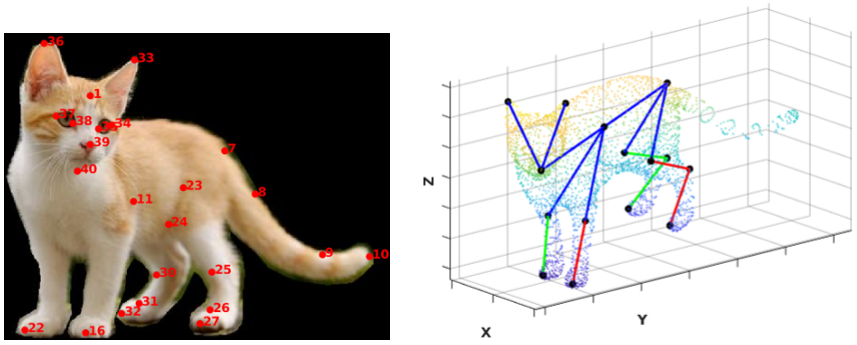


Figure 2. Examples in the proposed dataset Cat. Each red point marks a key point. The number represents the index of the key point, as shown in Table 1.

Table 1. Definition of key points in the proposed dataset, Cat.

Index	Definition	Index	Definition	Index	Definition	Index	Definition
1	forehead	11	left shoulder	21	right front ankle	31	right ankle
2	spine 0	12	left front thigh	22	right front toe	32	right toe
3	spine 1	13	left front shin	23	left thigh	33	left ear
4	spine 2	14	left front foot	24	left shin	34	left eye outer corner
5	spine 3	15	left front ankle	25	left foot	35	left eye inner corner
6	spine 4	16	left front toe	26	left ankle	36	right ear
7	root of tail	17	right shoulder	27	left toe	37	right eye outer corner
8	tail 1	18	right front thigh	28	right thigh	38	right eye inner corner
9	tail 2	19	right front shin	29	right shin	39	nose
10	end of tail	20	right front foot	30	right foot	40	chin



(**a**) (**b**)

Figure 3. Pose definition: (**a**) key points defined in [51]. The number represents the index of the key point, as shown in Table 1; (**b**) joints defined in [24] are represented by black points. The blue, red and green lines indicate the connections between the joints on the trunk, left limb and right limb, respectively.

3.2. Amur

To evaluate the effectiveness of our proposed method for low-quality images, we constructed another dataset by recycling the low-quality frames in the public ATRW dataset [24]. The ATRW was proposed for tiger detection, pose estimation, and reidentification. However, it provides ground-truth annotations only for good-quality images. Hence, we manually marked the key points for the low-quality Amur images in the dataset

and called the obtained dataset Amur. This dataset was used for evaluating the performance of our proposed method in refining the key points detected in low-quality images.

4. Proposed Method

It is well known that 2D poses are projections of 3D configurations. From this 2D projection alone, human observers are able to effortlessly imagine its possible 3D pose and camera position. Therefore, we assumed that prior knowledge about 3D poses can be encoded by atoms in a 3D pose dictionary. In particular, when the image quality is poor, and the textural information is missing, the 3D pose dictionary can provide 3D structure information, which could effectively help 2D pose estimation. The overview of the proposed method is shown in Figure 4, which can be roughly divided into three modules:

1. Pose dictionary learning: First of all, 3D poses were generated, as described in Section 3.1. These 3D poses were then used as the training data for dictionary learning to obtain the 3D pose dictionary (see Section 4.1);
2. Initial pose estimation: The initial pose of the sample was estimated using existing pose estimation algorithms such as HRNet [25];
3. Pose refinement: The initial pose p was used together with the 3D pose dictionary B to obtain the latent 3D pose P, which was then reprojected to obtain a more accurate 2D pose p' (see Section 4.3).

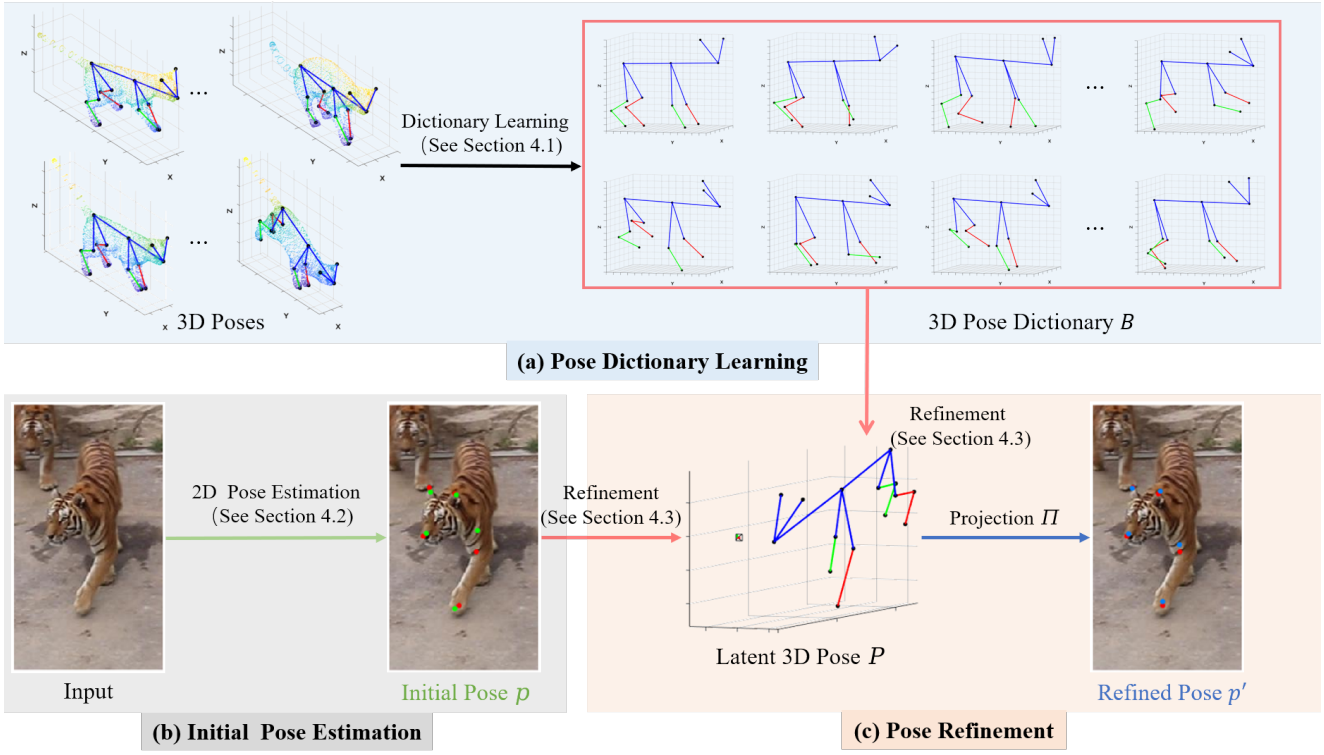


Figure 4. Overview: (**a**) we constructed a 3D pose dataset (as described in Section 3.1) to learn a 3D pose dictionary B, which was used to provide 3D prior constraints; (**b**) in the pose estimation stage, the 2D pose p corresponding to the image was initially estimated using existing algorithms; (**c**) refinement was performed by combining the initial pose p and the 3D constraints provided by the 3D pose dictionary B, resulting in a more accurate 2D pose p'. In (**b**,**c**), the blue, red and green lines indicate the connections between the joints on the trunk, left limb and right limb, respectively.

4.1. Pose Dictionary Learning

We used dictionary learning to find a good set of bases for 3D poses, and the dataset established in Section 3.1 was used as training data. We believe that a reasonable 3D

configuration space of 3D poses can be generated on these bases, so as to make the sparse representation of any 3D pose possible. Our task of dictionary learning can be formulated as follows:

$$\begin{aligned} \min_{B_1,\ldots,B_K,C} \sum_{n=1}^{N} \frac{1}{2} \left\| P_n - \sum_{k=1}^{K} C_{k,n} B_k \right\|_F^2 + \lambda \|C\|_1 \\ s.t. \quad C_{k,n} \geq 0, \quad \|B_k\|_F \leq 1 \\ k \in [1, K], \quad n \in [1, N] \end{aligned} \tag{1}$$

where λ is a non-negative parameter, K is the size of the dictionary, N represents the number of training samples, P_n denotes a 3D pose in the collected dataset, B_k is the base pose to be learned, and $C_{k,n}$ represents the n-th coefficient of the representation of P_n. The two terms in the cost function corresponding to the reconstruction error and the sparsity of representation, respectively. Examples of learned dictionary atoms are shown in Figure 4.

The problem in Equation (1) is locally solved by alternately updating C and B_ks via the projected gradient descent, an algorithm widely used in dictionary learning that converges to a local optimum [53]. The cost function can be rewritten as

$$f(\widetilde{B}, C) = \frac{1}{2} \left\| P_n - \sum_{k=1}^{K} C_{k,n} B_k \right\|_F^2 + \lambda \sum_{k,n} C_{k,n} \tag{2}$$

where $\widetilde{B}$ is the concatenation of B_ks. These steps are summarized in Algorithm 1.

Algorithm 1 Pose Dictionary Learning

Input: $P_1, \cdots, P_N$
Output: $B_1, \cdots, B_K$

initialize $\widetilde{B}, C$, step size δ_1 and δ_2;
while not converged **do**
    while not converged **do**
        $C \leftarrow C - \delta_1 \nabla_c f(\widetilde{B}, C)$
        for $k = 1$ to K **do**
            for $n = 1$ to N **do**
                if $C_{k,n} < 0$ **then**
                $C_{k,n} \leftarrow 0$
    $\widetilde{B} \leftarrow \widetilde{B} - \delta_2 \nabla_{\widetilde{B}} f(\widetilde{B}, C)$
for $k = 1$ to K **do**
    if $\|B_k\|_F > 1$ **then**
        $B_k \leftarrow B_k / \|B_k\|_F$

4.2. Initial Pose Estimation

It is worth emphasizing that the proposed method works as a plug-in postprocessing module and can be attached to existing animal pose estimation methods. Existing animal pose estimation methods can be divided into two categories: fully supervised learning and domain adaptation. In this paper, we chose state-of-the-art algorithms from the two categories for initial pose estimation. Mu et al. [15] proposed a novel consistency-constrained semi-supervised learning method (CC-SSL) to bridge the domain gap between real and synthetic images. Li et al. [16] designed a multi-scale domain adaptation module for unsupervised domain adaptation (UDA) on animal pose estimation. In addition, fully supervised methods such as Hourglass [43], ResNet [47], and HRNet [25], which perform well in human pose estimation, were also used for initial animal pose estimation.

4.3. Pose Refinement

4.3.1. Optimization-Based Pose Refinement

In this paper, we simplify the relationship between 3D and 2D poses as follows:

$$p \approx \Pi P \tag{3}$$

where $p \in \mathbb{R}^{2\times J}$ and $P \in \mathbb{R}^{3\times J}$ denote 2D and 3D poses, respectively. J is the number of joints or key points. Π is usually defined based on the weak perspective camera model as

$$\Pi = \begin{bmatrix} a & 0 & 0 \\ 0 & a & 0 \end{bmatrix} \tag{4}$$

where a is a scalar depending on the focal length and the distance to the object [38].

According to Equation (3), 2D pose p can be obtained if the latent 3D pose P is accessible. However, estimating a 3D pose from a 2D pose is an ill-posed problem. Fortunately, this pursuit is actually not in vain, as it seems, and attempts in the field of 3D human pose estimation have justified that lifting 2D poses to 3D poses is very fruitful [26,38]. We assume that the latent 3D pose has a combined representation:

$$P = \sum_{k=1}^{K} c_k R_k B_k \tag{5}$$

where $B_k \in \mathbb{R}^{3\times J}$ for $k \in [1, K]$ represents a base pose in the learned dictionary. While c_k denotes the weight of each base pose, R_k is a rotation matrix. Then, we substitute Equation (5) into Equation (3) to obtain

$$\begin{aligned} p \approx \Pi P = \Pi \sum_{k=1}^{K} c_k R_k B_k = \sum_{k=1}^{K} M_k B_k \\ s.t. \quad M_k M_k^T = c_k^2 I_2, \quad k \in [1, K] \end{aligned} \tag{6}$$

where $M_k = c_k \bar{R}_k$, $\bar{R}_k$ is the first two row vectors of R_k, and I_2 is the unit matrix of size 2×2.

Further, we use the following objective function to estimate the latent 3D pose:

$$\begin{aligned} \min_{M_1,\ldots,M_K} \frac{1}{2} \left\| p - \sum_{k=1}^{K} M_k B_k \right\|_F^2 + \alpha \sum_{k=1}^{K} \|M_k\|_2 \\ s.t. \quad M_k M_k^T = c_k^2 I_2, \quad k \in [1, K] \end{aligned} \tag{7}$$

where α is a predefined coefficient of the regularization. An auxiliary variable Z is introduced, and Equation (7) is rewritten as

$$\begin{aligned} \min_{\widetilde{M},Z} \frac{1}{2} \left\| p - Z\widetilde{B} \right\|_F^2 + \alpha \sum_{k=1}^{K} \|M_k\|_2 \\ s.t. \widetilde{M} = Z, \widetilde{M} = \begin{bmatrix} M_1 \cdots M_K \end{bmatrix}, \widetilde{B} = \begin{bmatrix} B_1 \\ \vdots \\ B_K \end{bmatrix} \end{aligned} \tag{8}$$

The steps are summarized in Algorithm 2. With $\{M_k\}_1^K$ we can finally obtain our refined 2D pose p', which is the projection of the estimated latent 3D pose P, i.e.,

$$p' = \sum_{k=1}^{K} M_k B_k \tag{9}$$

Algorithm 2 Pose Refinement

Input: p, α

Output: $M_1, \cdots, M_K$

initialize $Z = Y = 0, \mu > 0$;

while not converged **do**

  for $k = 1$ to K **do**

    $Q_k^t = Z_k^t - \frac{1}{\mu} Y_k^t$

    $M_k^t = \mathcal{D}_{\frac{\alpha}{\mu}}(Q_k^t)$

  $Z^{t+1} = (p\widetilde{B}^T + \mu\widetilde{M}^{t+1} + Y^t)(\widetilde{B}\widetilde{B}^T + \mu I)^{-1}$

  $Y^{t+1} = Y^t + \mu(\widetilde{M}^{t+1} - Z^{t+1})$

where $\mathcal{D}_{\frac{\alpha}{\mu}}$ is proximal operator [38].

4.3.2. Deep-Learning-Based Pose Refinement

We provide a deep-learning-based implementation, in which both the coefficients c_k and the rotation matrix R_k in Equation (6) are estimated using a neural network. Figure 5 shows the pipelines. The proposed network consists of the coefficient network and camera network. Given the initial pose p as network input, the coefficient network and camera network estimate coefficients c_k and rotation matrix R_k in Equation (6), respectively. The coefficients c_k and the rotation matrix R_k estimated via the network are combined with the learned 3D pose dictionary B to generate the 3D pose P corresponding to the initial pose p. Finally, the 3D pose P is reprojected to obtain the refined pose p'.

Network Architecture. The 2D input vectors are connected to a fully connected layer to expand the dimensionality to 1024 and then fed into subsequent residual blocks. Afterward, the network splits into two paths that predict the coefficients and the rotation matrices. Each path has two consecutive residual blocks followed by a fully connected layer that downscales the features to the required scale.

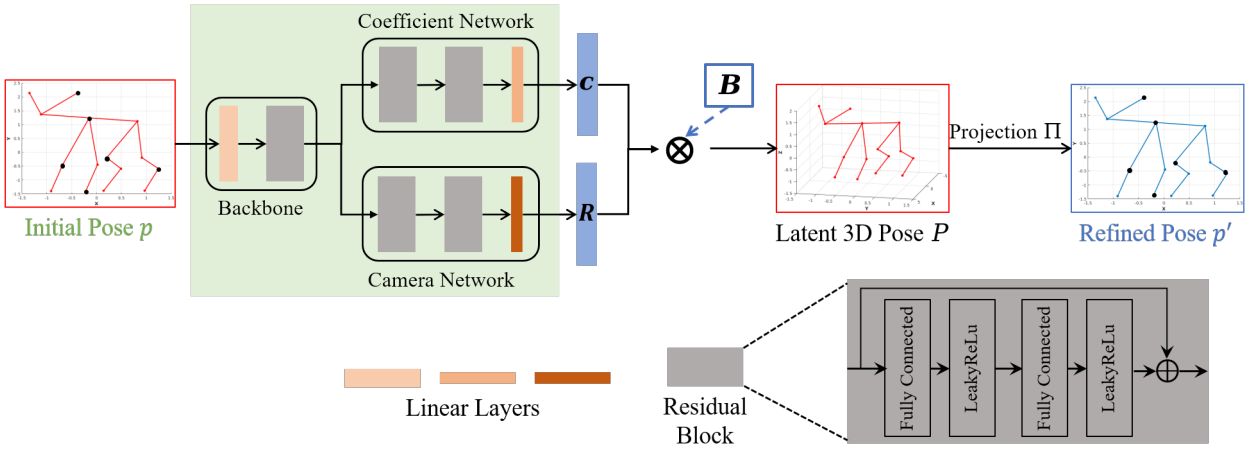


Figure 5. Overview. Given the initial pose p, the coefficient network and camera network estimate coefficients c_k and rotation matrix R_k in Equation (6), respectively. The coefficients c_k and the rotation matrix R_k are combined with the 3D pose dictionary B to generate a 3D pose corresponding to the initial pose p. The 3D pose P is reprojected to obtain the refined pose p'.

5. Experiments

We evaluate the effectiveness of the proposed method in this section. As a pose refinement module, the proposed method can be regarded as the postprocessing of existing pose estimation algorithms. Therefore, we applied the proposed method to state-of-the-art pose estimation methods and evaluated its effectiveness on different datasets.

5.1. Datasets

Synthetic animal. The synthetic animal dataset in [15] consists of images of elephants, horses, hounds, sheep, and tigers. In this dataset, animal textures and backgrounds are randomly synthesized using the COCO dataset [54]. For each animal species, 5000 images were generated with random textures and 5000 images with textures derived from the original CAD model. We only used the tiger subset, denoted by SA-Tiger, in our experiments (see Figure 6a).

TigDog. The TigDog dataset [55] provided key-point annotations for horses and tigers, where the images were taken from YouTube (for horses) and National Geographic documentaries (for tigers). Each image was annotated with 19 key points. These key points were distributed as follows: two key points on the eyes, one key point on the chin, two key points on the shoulders, twelve key points on the legs, one key point on the hip, and one key point on the neck. Among them, 11 key points overlapped with those defined by the ATRW [24]. These overlapping key points were used for evaluation in this paper. For convenience, we call the subset of tiger images in this dataset TD-Tiger (see Figure 6b).

Amur. As described in Section 3.2, Amur is a tiger image dataset. The image quality in this dataset is poor due to factors such as complex backgrounds, unconstrained illumination, and four-limbed movement. We manually annotated these samples for experimental evaluation (see Figure 6c).



(**a**) (**b**) (**c**)

Figure 6. Example images in (**a**) SA-Tiger, (**b**) TD-Tiger, and (**c**) Amur.

5.2. Experimental Setup

As in [56], we used the percentage of correctly localized key points (PCK) as the metric for evaluation. For the i-th sample in the test set, PCK defines the predicted position of the j-th key point, $\tilde{y}_i^j$, which would be correct if it falls within a threshold of its ground-truth position y_i^j.

$$\left\| y_i^j - \tilde{y}_i^j \right\|_2 \leq \beta \mathcal{D} \tag{10}$$

where $\mathcal{D}$ is the reference normalizer, namely the maximum side length of the image bounding box for animals, and β controls the threshold for correctness. β was set to 0.05 as in [56].

5.3. Results on the SA-Tiger Dataset

In order to analyze the robustness of the proposed method to noise, we added noise to ground truth, just as [57]. The corrupted dataset was then used to evaluate the performance of the proposed 3D constraints in correcting the corrupted 2D poses. As in [57], we estimate the scale s by finding the maximal length of the bounding box along $x, y - axis$. Then, we sampled zero-mean Gaussian noise with standard variance $\sigma = \delta\% \times s$, where δ were set to 3, 4, 5, and 6.

The SA-Tiger dataset had 10,000 samples, which were randomly divided into 6000 training samples and 4000 testing samples. We trained the neural networks in both fully supervised and self-supervised ways. Fully Supervised: The initial pose p was used as the network input. c and R were estimated using the coefficient network and camera network,

respectively, and then combined with the learned dictionary B to obtain the refined pose p' ($\Pi \sum_{k=1}^{K} c_k R_k B_k$). The network was trained by calculating the error between the refined pose p' and the ground-truth pose p_{gt}, as in Equation (11). Self-supervised: The loss function was the error between the input pose (i.e., the initial pose p) and the refined pose p' (see Equation (12)).

$$\mathcal{L}_{reproj}^{full} = \|p' - p_{gt}\|_F^2 \tag{11}$$

$$\mathcal{L}_{reproj}^{self} = \|p' - p\|_F^2 \tag{12}$$

The different scales of noise indirectly reflect the difficulty levels of the pose estimation tasks. GT+$\mathcal{N}(0, \delta)$ simulates different initial pose p and then evaluates the refinement performance of different coefficient solutions. It can be seen from Table 2 that full supervision achieves the greatest improvement. However, in practical applications, it is difficult to obtain ground-truth data with the same distribution as the pose to be refined for training. In addition, there is no significant improvement in self-supervision. These indicate that the network heavily relies on training data. In contrast, optimization-based pose refinement has better practicability and generalization. Specifically, the improvement is 1.9%, 3.8%, 4.5%, and 4.8% respectively, and the top two gains are achieved on GT+$\mathcal{N}(0, 5)$ and GT+$\mathcal{N}(0, 6)$. Thus, when we mention the proposed method in the following sections, the optimization-based pose refinement is indicated unless specified.

Table 2. PCK@0.05 accuracy on SA-Tiger with different levels of Gaussian noise added to the ground truth(GT) pose (%). The best results are marked in bold.

Method	Ear	Nose	Shoulder	Front Paw	Hip	Knee	Back Paw	Tail	Center	Mean
GT+$\mathcal{N}(0, 3)$	75.9	75.2	75.5	75.9	75.0	74.4	74.1	75.6	75.8	75.3
Ours (Optimization)	**75.6**	**77.4**	**78.3**	**76.5**	**78.3**	**76.9**	**75.1**	**80.7**	**81.1**	**77.2**
Ours (Fully supervised)	85.7	86.2	89.4	83.5	86.9	87.9	87.3	86.2	95.1	87.3
Ours (Self-supervised)	67.2	65.0	77.7	69.2	71.9	76.3	69.4	71.3	38.7	67.3
GT+$\mathcal{N}(0, 4)$	55.0	54.1	53.8	53.5	55.2	54.5	54.0	54.8	53.0	54.2
Ours (Optimization)	**55.2**	**57.4**	**61.4**	**54.3**	**60.3**	**59.9**	**56.0**	**59.8**	**64.5**	**58.0**
Ours (Fully supervised)	73.8	71.7	83.0	71.0	79.0	78.9	77.3	73.3	90.6	77.3
Ours (Self-supervised)	49.0	52.8	61.2	50.4	61.5	58.3	49.9	55.2	36.7	51.9
GT+$\mathcal{N}(0, 5)$	40.0	38.5	38.7	40.6	39.3	39.3	39.7	40.3	39.7	39.7
Ours (Optimization)	**41.0**	**42.8**	**46.0**	**40.6**	**48.3**	**44.8**	**41.3**	**45.5**	**54.3**	**44.2**
Ours (Fully supervised)	62.0	60.8	75.1	59.5	69.7	69.0	67.4	63.7	82.8	67.3
Ours (Self-supervised)	37.0	38.9	48.3	37.4	49.0	40.9	36.4	43.0	34.4	39.7
GT+$\mathcal{N}(0, 6)$	29.2	29.1	29.5	29.0	29.5	30.3	29.0	28.5	30.5	29.4
Ours (Optimization)	**29.8**	**32.5**	**38.5**	**30.2**	**40.3**	**35.7**	**30.3**	**34.8**	**43.6**	**34.2**
Ours (Fully supervised)	51.0	50.0	67.6	50.8	61.3	58.4	58.0	54.1	74.2	57.8
Ours (Self-supervised)	26.8	31.3	36.7	27.3	41.5	31.5	25.8	31.0	27.8	30.1

5.4. Results on the TD-Tiger Dataset

The quality of animal data varies in a wide range, especially those collected in the wild. To investigate the performance of the proposed method on images with different qualities, we divided the images into different subsets according to the PCK values of the initially estimated poses. Specifically, we set three intervals of PCK values, $(0, 45]$, $(45, 65]$, and $(65, 100]$. We evaluated the effectiveness of the proposed method for images in different intervals, respectively.

It is now known that in CC-SSL [15] and UDA [16], the trained model can be used directly. However, Hourglass [43], ResNet [47], and HRNet [25] are not currently designed for animal pose estimation on the TD-Tiger. Therefore, we trained these fully supervised networks according to the data settings of CC-SSL [15] and UDA [16]. Images

in TD-Tiger were split into training and verification sets, with 6523 images for training and 1765 images for verification. The PCK@0.05 accuracies obtained using the proposed method and different initial pose estimation methods on TD-Tiger are shown in Table 3. As can be seen from the results, while the proposed method maintained comparable accuracy for high-quality images, it obviously improved the accuracy for poor-quality images (i.e., in $(0, 45]$). It should be noted that the initial pose for high-quality images was already fairly accurate (i.e., in $(65, 100]$), and refinement may not be necessary. The success of the proposed method proves that the deviated key points can be effectively pulled back toward their true positions by exploiting 3D prior constraints on the animal poses.

Table 3. PCK@0.05 accuracy on TD-Tiger (%). The best results are marked in bold.

Method	$(0, 45]$	$(45, 65]$	$(65, 100]$
Hourglass [43]	28.3	55.7	87.6
Ours	**29.4**	**57.0**	**87.6**
ResNet [47]	29.3	54.3	88.1
Ours	**29.4**	**54.8**	**88.1**
HRNet [25]	29.5	55.4	87.8
Ours	**30.5**	**56.1**	**87.8**
CC-SSL [15]	28.8	54.7	73.6
Ours	**32.6**	**55.3**	**73.6**
UDA [16]	30.0	54.7	75.0
Ours	**34.7**	**56.4**	**75.0**

5.5. Results on the Amur Dataset

HRNet [25] and ResNet [47] trained for tiger pose estimation have been provided by MMPose [58]. Following the MMPose [58], we trained the Hourglass [43] using 2193 images from the training set in the ATRW [24]. For a fair comparison, we retrained CC-SSL [15] and UDA [16]. The training strategy was as follows: The source domain was the Synthetic Animal [15], which was the same as CC-SSL [15] and UDA [16]; the target domain training data were the training set in the ATRW [24]. As can be seen, with the proposed method, all the initial models had improvements in performance or were comparable to initial results. For example, UDA [16] achieved a 4.4% improvement in the interval $(0, 45]$; Meanwhile, in the interval $(65, 100]$, the proposed method was also comparable. On the whole, the proposed method had noticeably superior performance in challenging samples. This justifies the effectiveness of the proposed 3D constraints for 2D tiger pose estimation, particularly in challenging cases. The results of different methods on Amur are shown in Table 4 and Figure 7.

Table 4. PCK@0.05 accuracy on Amur (%). The best results are marked in bold.

Method	$(0, 45]$	$(45, 65]$	$(65, 100]$
Hourglass [43]	34.1	56.4	84.0
Ours	**34.4**	**57.1**	**84.1**
ResNet [47]	35.0	56.8	84.3
Ours	**36.8**	**57.4**	**84.3**
HRNet [25]	35.9	57.0	84.7
Ours	**39.2**	**57.0**	**84.8**
CC-SSL [15]	32.1	55.4	75.4
Ours	**35.1**	**55.9**	**76.0**
UDA [16]	34.4	55.2	76.5
Ours	**38.8**	**56.1**	**77.2**

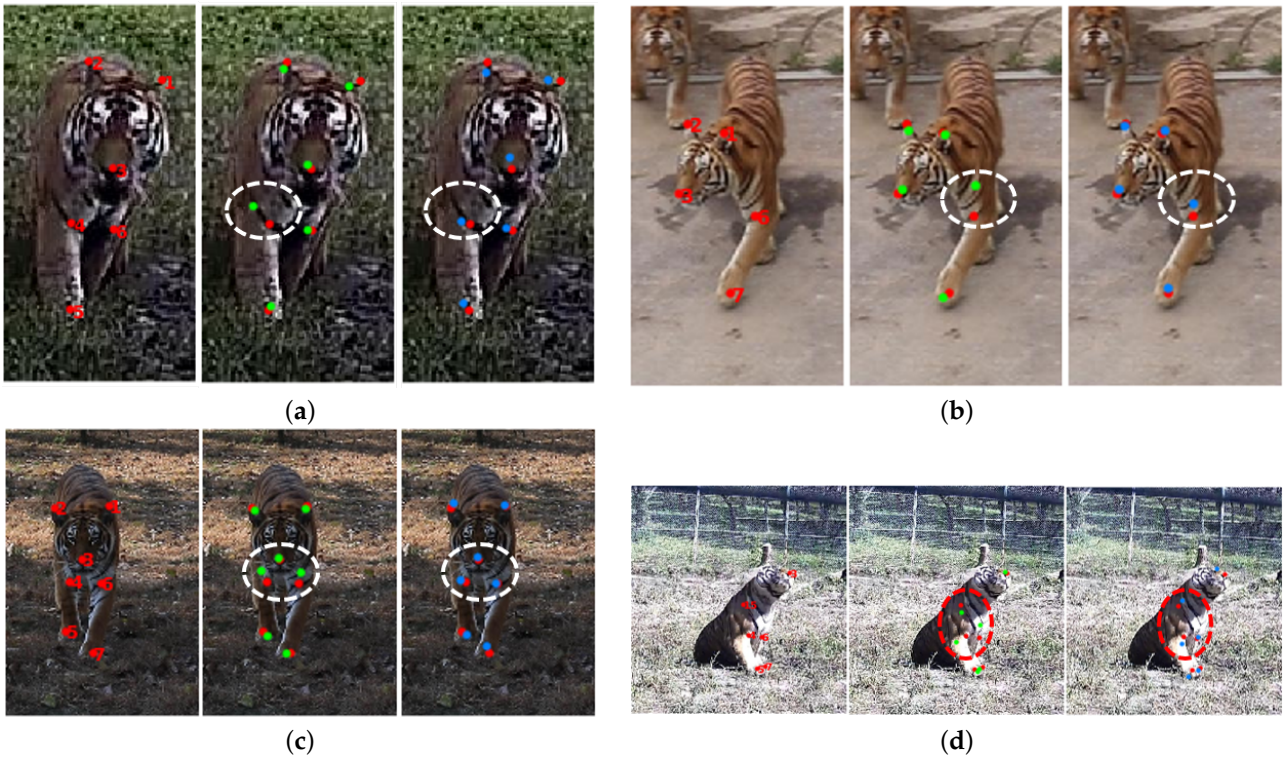
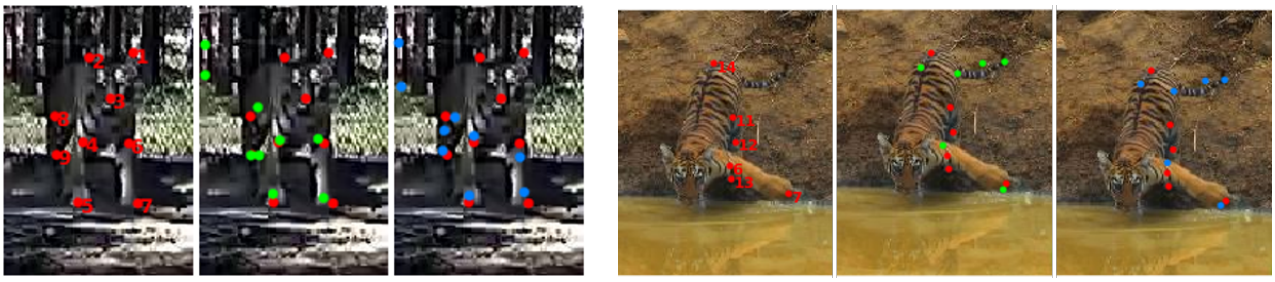


Figure 7. Qualitative Results: (**a**) underexposed; (**b**) motion blur; (**c**) complex background (similar to tiger body texture); (**d**) overexposure. The red point, green point, and blue point represent the ground truth, the initially estimated pose, and the refined pose, respectively. The number represents the index of the joint [24]. Both the red and white dashed circles indicate significant refinement.

6. Discussions and Conclusions

6.1. Discussions

As shown in Figure 7, the proposed method could effectively improve the animal pose estimation accuracy for poor-quality images due to occlusion, complex background, and uncontrollable illumination. In addition, we show two samples in which the animals were subject to the initial pose or the pose dictionary. In the first sample (Figure 8a), the initial poses completely violated the reasonable distribution of 2D poses, leading to irreparable errors. The second sample (Figure 8b) was not only limited by the initial pose but also by the pose dictionary. The training data for 3D pose dictionary learning were synthetic. If more real data can be used for training, a pose dictionary that is more robust to noise can be obtained. We will consider this in future work. Moreover, the dictionary encoded prior knowledge for specific classes of animals. It can be generalized between animal species with similar shapes, e.g., cats and tigers, but not between arbitrary animals. Therefore, obtaining a more diverse dictionary might be a potential way to further improve the estimation performance.



(**a**) (**b**)

Figure 8. Failure Cases. Example in (**a**) Amur, and (**b**) TD-Tiger. The red point, green point, and blue point represent the ground truth, the initially estimated pose, and the refined pose, respectively. The number represents the index of the joint [24].

6.2. Conclusions

In this paper, we presented a method to estimate 2D animal poses with 3D constraints. The 3D constraints were constructed from synthetic 3D poses and encoded in the 3D pose dictionary using sparse dictionary learning. Extensive experiments were conducted to evaluate the proposed method. Experimental results showed that the constraints provided by the 3D pose dictionary produced more reasonable animal pose estimation. In addition, we constructed two datasets: Cat for 3D pose dictionary learning and Amur for algorithm evaluation. The findings revealed in this paper can promote the further development of animal pose estimation.

Author Contributions: Conceptualization, X.D. and Q.Z.; methodology, X.D. and Q.Z.; validation, X.D.; writing—original draft preparation, X.D.; writing—review and editing, X.D., S.L. and Q.Z.; supervision, Q.Z. and H.Y.; funding acquisition, Q.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (Grant No. 62176170), and the Science and Technology Department of Tibet (Grant No. XZ202102YD0018C), and 2021 Guidance Special Provincial Supporting-2021 Pilot Program (Information Software)-Yang Hongyu, College of Computer Science (Grant No. 0082604151226), and CAAI-Huawei MindSpore Open Found.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Three open-access datasets, ATRW, TigDog, and Synthetic Animal, were used in this work. Their links are as follows: ATRW: https://cvwc2019.github.io/challenge.html; TigDog: http://calvin-vision.net/datasets/tigdog/; Synthetic Animal: https://www.cs.jhu.edu/~qiuwch/animal/; (all of the above datasets accessed on 28 July 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mathis, A.; Mamidanna, P.; Cury, K.M.; Abe, T.; Murthy, V.N.; Mathis, M.W.; Bethge, M. DeepLabCut: Markerless Pose Estimation of User-Defined Body Parts with Deep Learning. *Nat. Neurosci.* **2018**, *21*, 1281–1289. [CrossRef] [PubMed]
2. Graving, J.M.; Chae, D.; Naik, H.; Li, L.; Koger, B.; Costelloe, B.R.; Couzin, I.D. DeepPoseKit, A Software Toolkit for Fast and Robust Animal Pose Estimation Using Deep Learning. *Elife* **2019**, *8*, e47994. [CrossRef] [PubMed]
3. Mathis, M.W.; Mathis, A. Deep Learning Tools for the Measurement of Animal Behavior in Neuroscience. *Curr. Opin. Neurobiol.* **2020**, *60*, 1–11. [CrossRef] [PubMed]
4. Mathis, A.; Schneider, S.; Lauer, J.; Mathis, M.W. A Primer on Motion Capture with Deep Learning: Principles, Pitfalls, and Perspectives. *Neuron* **2020**, *108*, 44–65. [CrossRef] [PubMed]
5. Biggs, B.; Roddick, T.; Fitzgibbon, A.; Cipolla, R. Creatures Great and SMAL: Recovering the Shape and Motion of Animals From Video. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; pp. 3–19.

6. Zuffi, S.; Kanazawa, A.; Black, M.J. Lions and Tigers and Bears: Capturing Non-Rigid, 3D, Articulated Shape From Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3955–3963.
7. Zuffi, S.; Kanazawa, A.; Berger-Wolf, T.; Black, M.J. Three-D Safari: Learning to Estimate Zebra Pose, Shape, and Texture From Images "In the Wild". In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5359–5368.
8. Shih, L.Y.; Chen, B.Y.; Wu, J.L. Video-Based Motion Capturing for Skeleton-Based 3D Models. In Proceedings of the Pacific-Rim Symposium on Image and Video Technology, Tokyo, Japan, 13–16 January 2009; pp. 748–758.
9. Pantuwong, N.; Sugimoto, M.; A Novel Template-Based Automatic Rigging Algorithm for Articulated-Character Animation. *Comput. Animat. Virtual Worlds* **2012**, *23*, 125–141. [CrossRef]
10. Pereira, T.D.; Shaevitz, J.W.; Murthy, M. Quantifying Behavior to Understand the Brain. *Nat. Neurosci.* **2020**, *23*, 1537–1549. [CrossRef] [PubMed]
11. Seok, S.; Wang, A.; Chuah, M.Y.; Otten, D.; Lang, J.; Kim, S. Design Principles for Highly Efficient Quadrupeds and Implementation on the MIT Cheetah Robot. In Proceedings of the IEEE International Conference on Robotics and Automation, Karlsruhe, Germany, 6–10 May 2013; pp. 3307–3312.
12. Zhao, D.; Song, S.; Su, J.; Jiang, Z.; Zhang, J. Learning Bionic Motions by Imitating Animals. In Proceedings of the IEEE International Conference on Mechatronics and Automation, Beijing, China, 13–16 October 2020; pp. 872–879.
13. Ionescu, C.; Papava, D.; and Olaru, V.; Sminchisescu, C. Human3. 6m: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1325–1339. [CrossRef] [PubMed]
14. Zuffi, S.; Kanazawa, A.; Jacobs, D.W.; Black, M.J. 3D Menagerie: Modeling the 3D Shape and Pose of Animals. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6365–6373.
15. Mu, J.; Qiu, W.; Hager, G.D.; Yuille, A.L. Learning From Synthetic Animals. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 12386–12395.
16. Li, C.; Lee, G.H. From Synthetic to Real: Unsupervised Domain Adaptation for Animal Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 1482–1491.
17. Cao, J.; Tang, H.; Fang, H.S.; Shen, X.; Lu, C.; Tai, Y.W. Cross-Domain Adaptation for Animal Pose Estimation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9498–9507.
18. Chen, W.; Wang, H.; Li, Y.; Su, H.; Wang, Z.; Tu, C.; Lischinski, D.; Cohen-Or, D.; Chen, B. Synthesizing Training Images for Boosting Human 3D Pose Estimation. In In Proceedings of the Fourth International Conference on 3D Vision, Stanford, CA, USA, 25–28 October 2016; pp. 479–488.
19. Varol, G.; Romero, J.; Martin, X.; Mahmood, N.; Black, M.J.; Laptev, I.; Schmid, C. Learning From Synthetic Humans. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 109–117.
20. Singh, P.; Bose, S.S. A Quantum-clustering Optimization Method for COVID-19 CT Scan Image Segmentation. *Expert Syst. Appl.* **2021**, *185*, 115637. [CrossRef] [PubMed]
21. Mittal, H.; Pandey, A.C.; Saraswat, M; Kumar, S.; Pal, R.; Modwel, G. A Comprehensive Survey of Image Segmentation: Clustering Methods, Performance Parameters, and Benchmark Datasets. *Multimed. Tools Appl.* **2022**, *81*, 35001–35026. [CrossRef]
22. Singh, P.; Bose, S.S. Ambiguous D-means Fusion Clustering Algorithm Based on Ambiguous Set Theory: Special Application in Clustering of CT Scan Images of COVID-19. *Knowl.-Based Syst.* **2021**, *231*, 107432. [CrossRef] [PubMed]
23. Dai, X.; Li, S.; Zhao, Q.; Yang, H. Animal Pose Refinement in 2D Images with 3D Constraints. In In Proceedings of the 2022-33rd British Machine Vision Conference, London, UK, 21–24 November 2022;
24. Li, S.; Li, J.; Tang, H.; Qian, R.; Lin, W. ATRW: A Benchmark for Amur Tiger Re-Identification in the Wild. In Proceedings of the ACM International Conference on Multimedia, Seattle, WA USA, 12–16 October 2020; pp. 2590–2598.
25. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5693–5703.
26. Martinez, J.; Hossain, R.; Romero, J.; Little, J.J. A Simple Yet Effective Baseline for 3D Human Pose Estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2640–2649.
27. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; and Kaiser, Ł.; Polosukhin, I. Attention is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
28. Zhao, W.; Wang, W.; Tian, Y. GraFormer: Graph-Oriented Transformer for 3D Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–23 June 2022; pp. 20438–20447.
29. Li, W.; Liu, H.; Tang, H.; Wang, P.; Van G.L. Mhformer: Multi-Hypothesis Transformer for 3D Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–23 June 2022; pp. 13147–13156.
30. Wandt, B.; Rosenhahn, B. Repnet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 7782–7791.
31. Li, C.; Lee, G.H. Weakly Supervised Generative Network for Multiple 3D Human Pose Hypotheses. In Proceedings of the 2020—31st British Machine Vision Conference, Virtual Event, UK, 7–10 September 2020.

32. Usman, B.; Tagliasacchi, A.; Saenko, K.; Sud, A. MetaPose: Fast 3D Pose from Multiple Views without 3D Supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–23 June 2022; pp. 6759–6770.
33. Wandt, B.; Rudolph, M.; Zell, P.; Rhodin, H.; Rosenhahn, B. Canonpose: Self-supervised Monocular 3D Human Pose Estimation in the Wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13294–13304.
34. Drover, D.; MV, R.; Chen, C.H.; Agrawal, A.; Tyagi, A.; Phuoc H.C. Can 3D Pose be Learned from 2D Projections Alone? In Proceedings of the European Conference on Computer Vision Workshops, Munich, Germany, 8–14 September 2018.
35. Chen, C.H.; Tyagi, A.; Agrawal, A.; Drover, D.; Mv, R.; Stojanov, S.; Rehg, J.M. Unsupervised 3D Pose Estimation with Geometric Self-supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5714–5724.
36. Yu, Z.; Ni, B.; Xu, J.; Wang, J.; Zhao, C.; Zhang, W. Towards Alleviating the Modeling Ambiguity of Unsupervised Monocular 3D Human Pose Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 10–17 October 2021; pp. 8651–8660.
37. Wandt, B.; Little, J.J.; Rhodin, H. ElePose: Unsupervised 3D Human Pose Estimation by Predicting Camera Elevation and Learning Normalizing Flows on 2D Poses. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–23 June 2022; pp. 6635–6645.
38. Zhou, X.; Zhu, M.; Leonardos, .; Daniilidis, K. Sparse Representation for 3D Shape Estimation: A Convex Relaxation Approach. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1648–1661. [CrossRef] [PubMed]
39. Wang, C.; Qiu, H.; Yuille, A.L.; Zeng, W. Learning Basis Representation to Refine 3D Human Pose Estimations. In Proceedings of the AAAI Conference on Artificial intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 8925–8932.
40. Ramakrishna, V.; Kanade, T.; Sheikh, Y. Reconstructing 3D Human Pose from 2D Image Landmarks. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 573–586.
41. Wang, C.; Wang, Y.; Lin, Z.; Yuille, A.L.; Gao, W. Robust Estimation of 3D Human Poses from A Single Image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2361–2368.
42. Akhter, I.; Black, M.J. Pose-Conditioned Joint Angle Limits for 3D Human Pose Reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1446–1455.
43. Newell, A.; Yang, K.; Deng, J. Stacked Hourglass Networks for Human Pose Estimation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 483–499.
44. Wei, S.E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional Pose Machines. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4724–4732.
45. Xiao, B.; Wu, H.; Wei, Y. Simple Baselines for Human Pose Estimation and Tracking. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 466–481.
46. Pereira, T.D.; Aldarondo, D.E.; Willmore, L.; Kislin, M.; Wang, S.S.-H.; Murthy, M.; Shaevitz, J.W. Fast Animal Pose Estimation Using Deep Neural Networks. *Nat. Methods* **2019**, *16*, 117–125. [CrossRef] [PubMed]
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
48. Insafutdinov, E.; Pishchulin, L.; Andres, B.; Andriluka, M.; Schiele, B. DeeperCut: A Deeper, Stronger, and Faster Multi-person Pose Estimation Model. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 34–50.
49. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
50. Jégou, S.; Drozdzal, M.; Vazquez, D.; Romero, A.; Bengio, Y. The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 11–19.
51. Kanazawa, A.; Kovalsky, S.; Basri, R.; Jacobs, D. Learning 3D Deformation of Animals From 2d Images. *Comput. Graph. Forum* **2016**, *35*, 365–374. [CrossRef]
52. Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; Black, M.J. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graph.* **2015**, *34*, 1–16. [CrossRef]
53. Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G. Online Learning for Matrix Factorization and Sparse Coding. *J. Mach. Learn. Res.* **2010**, *11*, 19–60.
54. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
55. Del Pero, L.; Ricco, S.; Sukthankar, R.; Ferrari, V. Articulated Motion Discovery Using Pairs of Trajectories. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2151–2160.
56. Yu, X.; Zhou, F.; Chandraker, M. Deep Deformation Network for Object Landmark Localization. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 52–70.

57. Mei, J.; Chen, X.; Wang, C.; Yuille, A.; Lan, X.; Zeng, W. Learning to Refine 3D Human Pose Sequences. In Proceedings of the International Conference on 3D Vision, Québec City, QC, Canada, 16–19 September 2019; pp. 358–366.
58. OpenMMLab. Available online: https://github.com/open-mmlab/mmpose (accessed on 28 July 2022).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.