Animal Pose Refinement in 2D Images with 3D Constraints

Xiaowei Dai[1]
xwdai@foxmail.com
Shuiwang Li[2]
lishuiwang0721@163.com
Qijun Zhao[1,3]
qjzhao@scu.edu.cn
Hongyu Yang[1,3]
yanghongyu@scu.edu.cn

[1] National Key Laboratory of Fundamental Science on Synthetic Vision
Sichuan University
Chengdu, China

[2] College of Information Science and Engineering
Guilin University of Technology
Guilin, China

[3] College of Computer Science
Sichuan University
Chengdu, China

Abstract

Animal pose has many potential applications in various fields. However, uncontrollable illumination, complex backgrounds and random occlusions in in-the-wild animal images often lead to large errors in pose estimation. To address this problem, we propose a method for refining the initial animal pose with 3D prior constraints. First, we learn a 3D pose dictionary from synthetic data with each atom providing 3D pose prior knowledge. Then, the 3D pose dictionary is used to linearly represent the potential 3D pose corresponding to the 2D pose that has been initially estimated for the animal in 2D image. Finally, the representation coefficients are optimized to minimize the difference between the initially-estimated 2D pose and the 2D-projection of the potential 3D pose. Moreover, to deal with the data scarcity, we construct 2D and 3D animal pose datasets, which are used to evaluate algorithm performance and learn 3D pose dictionary, respectively. Experimental results show that the proposed method is capable to utilize 3D pose knowledge well and is effective in improving 2D animal pose estimation.

1 Introduction

As a challenging task in computer vision, animal pose estimation has a wide range of practical applications. For instance, animal pose estimation could be employed in markerless motion capture systems to remove intrusive markers. Animal pose estimation would also support advances in entertainment, where most animal animations are still performed manually. In neuroscience, tracking animals is fundamental for understanding the relationship between behavior (or movement) and brain activity. In bio-inspired robotics, understanding how animals move can help to design more efficient robots. Despite its promise, there is

(a) (b)

Figure 1: Examples with poses successfully refined by the proposed method. Ground truth, initial and refined poses are shown in **red**, **green** and **yellow** colors, respectively. The circles indicate significant deviations.

little work related to animal pose estimation. This can be attributed to the following factors. First, animals have far more variation in shape and pose than the human. Second, there is a lack of well-annotated datasets for animal pose estimation. Typical human pose estimation benchmarks, such as Human3.6M [10], capture data from 15 sensors and obtain motion relying on small markers attached to the subjects body. However, it is impractical to bring wild animals into a laboratory environment for scanning in specific poses.

Most existing approaches address data limitations by transferring knowledge from other more accessible domains such as synthetic animal data [2, 18, 27, 28, 29] or human datasets [3]. However, there are some domain gaps [4, 9, 22] between synthetic images and real images, which prevent models trained on synthetic images from generalizing well to real-world images. In addition, the existing animal pose models are trained on images only with 2D annotations, because 3D pose annotations are very hard to obtain or even to define. 2D poses are projection of the 3D body configuration, and 3D structural information is distorted in this process. Therefore, models trained only with 2D annotations may not conform to the true prior configurations of animal poses.

This problem can become severe, especially in wild animal images with uncontrollable illuminations, complicated backgrounds and random occlusions, as illustrated in Fig. 1. These factors often lead to large errors in animal pose estimation, which are not considered by existing methods. Given this, we propose a method with 3D constraints to refine 2D animal pose, and encode 3D prior constraints in 3D pose dictionary. As can be seen in Fig. 1, taking tigers as example, the proposed method can estimate more accurate 2D poses. The contributions of this paper can be summarized as follows:

- We propose a novel method for 2D animal pose refinement using 3D constraints for the first time.
- We construct a 3D animal pose dataset using synthetic methods for 3D pose dictionary learning. Furthermore, we collect and manually annotate images to build a 2D animal pose dataset for algorithm evaluation.
- Extensive experiments are conducted to evaluate the proposed method. Experimental results show that the proposed method is effective in improving accuracy of 2D animal pose estimation.

The remainder of this paper is organized as follows. We discuss related work and motivation in Section 2. Sections 3 and 4 provide details of the established datasets and the proposed method, respectively. Our experimental results are presented in Section 5. Finally, we conclude the paper in Section 6.

2 Related Work

2.1 Dictionary-based Human Pose Estimation

Dictionary-based methods are widely used in human pose estimation [24, 26]. In these methods, 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, Ramakrishna et al. [20] propose a sparse representation based approach to estimate 3D human pose from 2D annotations in a single image. They present a projected matching pursuit algorithm for reconstructing 3D poses and camera settings by minimizing the re-projection error. Wang et al. [23] propose to estimate the 3D pose by minimizing an L1-norm penalty between the projection of the 3D joints and the 2D detections to reduce the impact of inaccurate 2D pose estimations. Zhou et al. [26] adopt an augmented 3D shape model to achieve a linear representation of shape variability in 2D and propose to use the spectral-norm regularization to penalize invalid cases caused by the augmentation. Akhter and Black [1] integrate 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 pose dictionary in animal pose estimation.

2.2 Animal Pose Estimation

Thanks to the great success of deep learning, these neural networks have been applied to pose estimation in laboratory animals such as fruit flies, mice, and locusts [6, 7, 15, 19]. These laboratory animals are usually in a controlled environment, and researchers can easily collect and annotate data for neural network training. For wild animals, however, available datasets are very few due to the difficulty and cost of collection and annotation. To solve this problem, Cao et al. [3] believe that there is similarity between human and quadruped mammals. Thus, they propose a cross-domain adaptation scheme to learn a shared feature space between human and animal poses, so that their network can learn from existing human pose datasets. Mu et al. [18] use synthetic animal data generated from CAD models to train their model, and then generate pseudo-labels for unlabeled real animal images. Subsequently, the generated pseudo-labels are gradually incorporated into training based on three consistency check criteria. Li and Lee [12] design a multi-scale domain adaptation module to reduce the gaps between synthetic and real data. Meanwhile, a coarse-to-fine pseudo-label update strategy is introduced, and more accurate pseudo-labels are gradually replaced in the training process, so as to improve the accuracy of the model.

Despite the improvements made by these approaches [3, 6, 12, 15, 18, 19], animal pose estimation is still non-trivial. These methods are trained on images with 2D pose annotations, which may not follow the actual distribution of animal poses and degrade their performance very likely, especially for poor quality images. In this paper, we aim to explore 3D prior

constraints to refine the initial animal pose, so such that the misaligned keypoints caused by image noise can be corrected towards their true positions. Inspired by [1, 20, 23, 24, 26], we use 3D pose dictionary to encode 3D prior constraints, which is simple and effective. Note that [1, 20, 23, 24, 26] 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 collect and synthesize data for 3D pose dictionary learning.

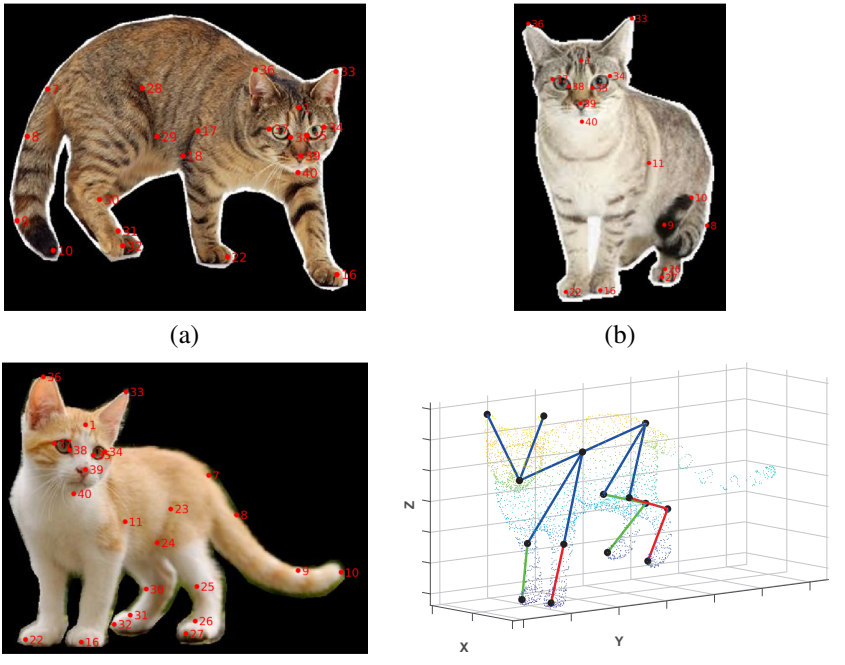


Figure 2: (a-c) are examples of keypoints in the dataset **Cat**. Red dots indicate the keypoints that are defined in Table 1. (d) The joints defined in ATRW [13].

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

Table 1: Keypoint definition for the dataset **Cat**.

3 Dataset Collection

In order to learn 3D pose configurations and impose 3D prior constraints in 2D pose refinement, we built a 3D cat pose dataset called **Cat**. This dataset contains more than 400 images from the internet. The details are as follows. First, we define the keypoints (as illustrated in Fig. 2(a)–2(c) and Table 1) with reference to Kanazawa et al. [11]. Second, we use Kanazawa et al. [11] to synthesize corresponding deformable shapes. Third, we pick the joints defined by ATRW [13] on the 3D deformable shapes to generate 3D poses (as illustrated in Fig. 2(d)). The synthesized 3D pose dataset can be used for 3D pose dictionary learning, and the learned 3D pose dictionary encodes 3D prior constraints that facilitate animal pose refinement.

4 Proposed Method

In this section, we first describe the proposed method for constructing a dictionary of 3D animal pose and then introduce the proposed method for 2D pose refinement with 3D constraints based on the learned dictionary. The pipelines of constructing the dictionary of 3D animal pose and 2D pose refinement with 3D constraints is illustrated in Fig. 3.

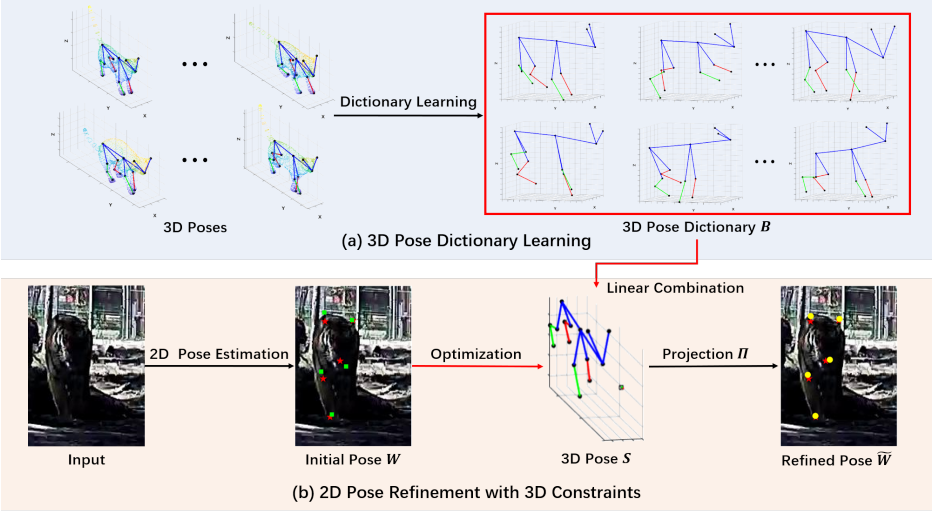


Figure 3: Overview. (a) We construct a 3D pose dataset (as described in Section 3) to learn a 3D pose dictionary B, which is used to provide 3D prior constraints. (b) In the 2D pose refinement stage, the 2D pose W corresponding to the image is initially estimated by the existing algorithms. The optimization is performed via combining the initial pose W and the 3D constraints provided by the 3D pose dictionary B, a more accurate 2D pose $\widetilde{W}$ is obtained.

4.1 3D Pose Dictionary Learning

With the proposed 3D pose dataset descibed in Section 3, we use dictionary learning to find a good basis for 3D poses. It is hoped that this basis generates a complete 3D pose space, so

that any sparse representation of 3D poses seems reasonable under this basis. Our dictionary learning task can be formulated as follows:

$$\min_{B,C} \sum_{n=1}^{N} \frac{1}{2} \left\| S_n - \sum_{k=1}^{K} C_{n,k} B_k \right\|_F^2 + \lambda \|C\|_1, \quad s.t. \quad C_{n,k} \geq 0, \quad \|B_k\|_F \leq 1, \quad \forall_k \in [1,K], \quad n \in [1,N], \tag{1}$$

where λ is a non-negative parameter; K is the number of atoms in the dictionary; N represents the number of training samples; S_n denotes a 3D pose in the collected dataset; B_k is the basis pose to be learned, and $C_{n,k}$ represents the kth coefficient of the representation of S_n. The two term in the cost function corresponding to the reconstruction error and the sparsity of representation, respectively. K is set to 128 by default in this paper.

4.2 2D Pose Refinement with 3D Constraints

As illustrated in the orange background of Fig. 3. Given an input image, an initial 2D pose W is first estimated using the existing methods such as HRNet [21]. For this initial 2D pose, it actually corresponds to some latent 3D pose S. It is generally assumed that they are related to projection (camera calibration matrix) Π. Specifically,

$$W \approx \Pi S, \tag{2}$$

where $W \in \mathbb{R}^{2\times p}$, $S \in \mathbb{R}^{3\times p}$, and p represents the number of joints. W and S represent 2D pose and 3D pose, respectively. Π is usually defined based on the weak perspective camera model as:

$$\Pi = \begin{bmatrix} a & 0 & 0 \\ 0 & a & 0 \end{bmatrix}, \tag{3}$$

where a is a scalar depending on the focal length and the distance to the object [26].

According to Eq. (2), 2D pose W can be obtained if the latent 3D pose S is accessible. Conversely, the given 2D pose W, the latent 3D pose S is almost unknown. Although estimating 3D pose from a 2D image is an ill-posed problem, this pursuit does not actually appear to be in vain. Similar attempts such as face recognition have proven to be very fruitful in recovering 3D information from 2D images. Therefore, with the estimated 2D pose W we hope to recover the latent 3D pose using the learned 3D pose dictionary. Similar to the active shape model [5], we assume the latent 3D pose has a sparse representation:

$$S = \sum_{k=1}^{K} c_k R_k B_k, \tag{4}$$

where $B_k \in \mathbb{R}^{3\times p}$ for $k \in [1,K]$ represents a basis pose in the learned dictionary. While c_k denotes the weight of each basis pose, R_k rotation matrix. Then, we use the following objective funtion to estimate a latent 3D pose:

$$\min_{M_i,\ldots,M_K} \frac{1}{2} \left\| W - \sum_{k=1}^{K} M_k B_k \right\|_F^2 + \alpha \sum_{k=1}^{K} \|M_k\|_2, \tag{5}$$

where $M_k = c_k \Pi R_k$ with $M_k M_k^T = c_k^2 I_2$, and I_2 is the unit matrix of size 2×2. α is a predefined coefficient of the regularization. With $\{M_k\}_1^K$ we can finally obtain our refined 2D pose $\widetilde{W}$,

which is the projection of the estimated latent 3D pose S, i.e.,

$$\widetilde{W} = \Pi S = \Pi \sum_{k=1}^{K} c_k R_k B_k = \sum_{k=1}^{K} M_k B_k. \tag{6}$$

The 3D constraints lie in that the latent 3D pose S is a sparse representation of the learned basis poses, and our refined estimated 2D pose is a projection of this representation.

5 Experiments

5.1 Datasets

Amur. ATRW [13] is a relatively complete dataset that can be used for Amur tiger detection, pose estimation and re-identification. Some poor frames are discarded due to occlusion, motion artifacts, illumination or other noise. To validate the advantages of the proposed method, we manually annotate the poor samples discarded by ATRW (excluding some extreme cases) to build a more challenging dataset, which is called **Amur**.

Synthetic Animal. The synthetic dataset in [18] consists of images of elephant, horse, hound, sheep and tiger. In this dataset, animal textures and backgrounds are randomly synthesized using COCO dataset [14]. For each animal species, $5,000$ images are generated with random texture and $5,000$ images with the texture coming with the original CAD model. We only use the tiger subset, denoted by **SA-Tiger**, in our experiments.

5.2 Evaluation Metrics

As in [25], we use the percentage of correctly localized keypoints (PCK) as the metric for evaluation. For the j th sample in the test set of size N, PCK defines the predicted position of the i th landmark $\tilde{y}_j^i$ to be correct if it falls within a threshold of the ground-truth position y_j^i, that is, if

$$\left\| y_j^i - \tilde{y}_j^i \right\|_2 \leq \beta \mathcal{D}, \tag{7}$$

where $\mathcal{D}$ is the reference normalizer, namely, maximum side length of the image bounding box for animals. The parameter β controls the threshold for correctness. β is set to 0.05 as in [25].

The quality of animal data varies greatly, especially data collected in the wild. To investigate the performance of the proposed method on different quality images, we divide images into different quality levels. We use the PCK of initial pose as the quality threshold to divide the test data. We empirically set three intervals: $(0,45]$, $(45,65]$ and $(65,100]$.

5.3 Comparison with State of the Art

Existing animal pose estimation algorithms can be divided into two categories: full supervision and domain adaptation. In the experiments, we select the state-of-the-art (SOTA) algorithms from these two categories for comparison. Fully supervised methods such as ResNet [8] and HRNet [21], which perform well in human pose estimation, are also used in animal pose estimation. ResNet [8] and HRNet [21] trained for tiger pose estimation have been provided by MMPose [17]. In addition, [18] propose a novel Consistency Constrained Semi-Supervised Learning method (CC-SSL) to bridge the domain gap between

Method	$(0,45]$	$(45,65]$	$(65,100]$
ResNet [8]	35.0	56.8	84.3
Ours	**37.2**	**57.4**	**84.3**
HRNet [21]	35.9	57.0	84.7
Ours	**38.8**	**57.8**	**84.7**
CC-SSL [18]	32.1	55.4	75.4
Ours	**35.1**	**55.4**	**75.4**
UDA [12]	34.4	55.2	76.5
Ours	**38.5**	**55.7**	**76.5**

Table 2: Results on the Amur (%).

real and synthetic images. [12] design a multi-scale domain adaptation module for Unsupervised Domain Adaptation (UDA) on animal pose estimation. For fair comparison, we retrain CC-SSL [18] and UDA [12]. The training strategy is as follows: the source domain is the Synthetic Animal [18], which is the same as CC-SSL [18] and UDA [12]; The training data for the target domain is the training set in ATRW [13]. We evaluate the above algorithms and the proposed method on the Amur dataset.

The results are shown in Table 2. As can be seen, with the proposed method incorporated all the initial models have improvements on performance or are comparable to initial results. For example, in the most challenging case (i.e., $(0,45]$), the proposed method outperforms UDA [12] on average with gains of 4.1%, thanks to that the 3D constraints can provide effective prior knowledge to help the pose estimation. Meanwhile, for the cases when the initial models work well (e.g., for images of high quality), the proposed method maintains the accuracy of the models. This is because the proposed method aims to improve the initial pose, and if the initial pose is already close to the true positions (i.e., in $(65,100]$), the proposed method will most probably keep the initial estimation unchanged. Generally, this demonstrates the effectiveness of the proposed 3D constraints for 2D pose refinement, especially in challenging cases.

5.4 Adding Noise to Ground Truth 2D Poses

Section 5.3 evaluates the effectiveness of the proposed method in refining the initial poses obtained by SOTA pose estimation methods. The results show that better improvement is achieved on the samples with larger initial errors (e.g., $(0,45]$). To further investigate its effectiveness for samples in $(0,45]$, we use the same method of adding noise as in [16] to corrupt the ground truth poses. These corrupted data are used to simulate the noisy initial 2D poses. The magnitude of the noise indicates the different degrees of deviation between the initial 2D pose and the ground truth (GT). As in [16], we estimate the scale s by finding the maximal length of bounding box along $x, y-axis$. Then we sample zero mean Gaussian noise with standard variance $\sigma = \delta\% \times s$, where δ are set to 5, 10 and 15. The PCK@0.05 of the proposed method at different Gaussian noises are shown in Table 3. It can be seen that the proposed 3D constraints have a large impact on the pose estimation. Specifically, gains of 3.2%, 4.6% and 2.1% are achieved over GT + $\mathcal{N}(0,5)$, GT + $\mathcal{N}(0,10)$ and GT + $\mathcal{N}(0,15)$, respectively. The reason why the proposed method is less effective for samples with very high noise levels might be that the initial poses do not lie in the reasonable distribution of 2D poses and cannot be reparable (See Fig. 5).

Method	Ear	Nose	Shoulder	Front Paw	Hip	Knee	Back Paw	Tail	Center	Mean
GT + $\mathcal{N}(0,5)$	39.7	38.0	**38.5**	**39.7**	**39.5**	38.6	39.3	39.7	39.4	39.2
Ours	**41.5**	**46.3**	35.6	39.6	32.7	**45.3**	**42.3**	**47.1**	**56.3**	**42.4**
GT + $\mathcal{N}(0,10)$	11.7	12.1	12.4	11.3	12.0	11.4	11.6	11.0	12.2	11.7
Ours	**13.9**	**16.4**	**18.8**	**12.7**	**19.5**	**19.8**	**13.7**	**15.6**	**22.0**	**16.3**
GT + $\mathcal{N}(0,15)$	5.8	5.3	5.4	5.2	5.4	5.3	5.3	5.7	5.3	5.4
Ours	**6.3**	**6.3**	**9.3**	**5.7**	**10.9**	**8.2**	**6.2**	**7.4**	**9.8**	**7.5**

Table 3: Results on the SA-Tiger adding Gaussian noise to ground truth 2D pose (%).

5.5 Different Scales of 3D Pose Dictionary

Since the number of basis is a very important factor in dictionary learning, we also evaluate the impact of the number of 3D basis poses (i.e., K), on the proposed method for 2D pose refinement. The proposed method with K of 64, 128 and 256 are evaluated on the $(0,45]$ of Amur, respectively. The results are summarized in Table 4. As can be seen, in average, when K is 128 the PCK@0.05 achieves the highest values, namely 38.8% on Amur. So we set K to 128 in the proposed method by default.

Method	Ear	Nose	Shoulder	Front Paw	Hip	Knee	Back Paw	Tail	Center	Mean
Ours ($K = 64$)	**56.8**	**61.1**	19.8	34.2	29.2	25.7	**28.5**	30.0	54.5	38.1
Ours ($K = 128$)	55.8	59.3	**22.8**	**36.3**	**29.2**	**28.4**	26.3	**30.0**	**54.5**	**38.8**
Ours ($K = 256$)	55.8	61.1	19.9	36.3	25.0	23.1	26.3	30.0	54.5	37.7

Table 4: Results of the proposed method with different K on the $(0,45]$ of Amur (%).

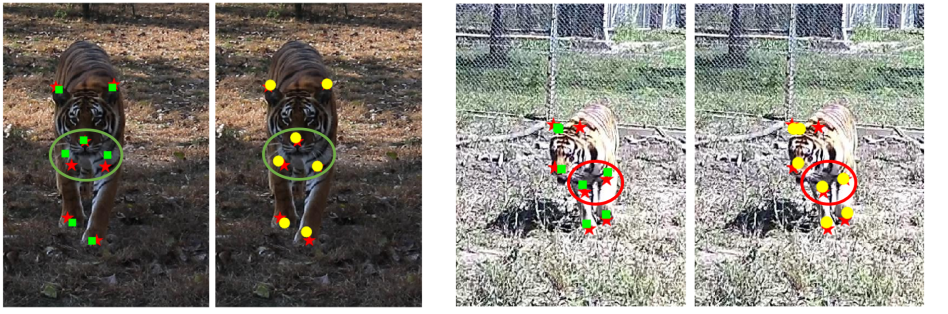
5.6 Qualitative Results

As shown in Fig. 1 and Fig. 4, in the poor quality in-the-wild animal images caused by occlusion, complex background and uncontrollable illumination, the proposed method can effectively improve the pose estimation accuracy. In addition, we also show the failure samples (See Fig. 5). In these cases, the initial poses completely violate the reasonable distribution of 2D poses, leading to irreparable errors. Moreover, the training data for 3D pose dictionary learning is synthetic. If more real data can be used for training, the pose dictionary that is more robust to noise can be obtained. We will consider these in future work.

6 Conclusion

In this paper, we present a method to refine 2D animal pose with 3D constraints. The 3D constraints are constructed with synthetic 3D poses and encoded in the 3D pose dictionary using sparse dictionary learning. Extensive experiments are conducted to evaluate the proposed method. Experimental results show that the proposed method is effective in improving 2D animal pose estimation, especially in challenging cases. In addition, we build a 3D animal pose dataset using synthetic methods, and collect and manually annotate a 2D pose dataset. We believe that the data collected and the proposed method will help advance the field.

Limitations and Future Work. It is worth emphasizing that the proposed method works as a plug-in post-processing module and can be attached to existing animal pose estimation



(a) (b)

Figure 4: Examples with poses successfully refined by the proposed method. Ground truth, initial and refined poses are shown in **red**, **green** and **yellow** colors, respectively.

(a) (b)

Figure 5: Examples with poses unsuccessfully refined by the proposed method. Ground truth, initial and refined poses are shown in **red**, **green** and **yellow** colors, respectively.

methods. Inspired by [1, 20, 23, 24, 26], we employ a dictionary to encode 3D prior constraints to refine the initial animal pose. To address the scarcity of 3D animal data, we use synthetic data for dictionary learning. However, if there is enough real data to learn the 3D pose dictionary, this is a way to further improve the refinement accuracy. Moreover, the dictionary is 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 refinement performance.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (Nos. 62176170 and 61773270) and CAAI-Huawei MindSpore Open Found.

References

[1] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *Proc. CVPR*, pages 1446–1455, 2015.

[2] B. Biggs, T. Roddick, A. Fitzgibbon, and R. Cipolla. Creatures great and smal: Re-

covering the shape and motion of animals from video. In *Proc. ACCV*, pages 3–19, 2018.

[3] J. Cao, H. Tang, H. S. Fang, X. Shen, C. Lu, and Y. W. Tai. Cross-domain adaptation for animal pose estimation. In *Proc. ICCV*, pages 9498–9507, 2019.

[4] W. Chen, H. Wang, Y. Li, Z. Su, H.and Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing training images for boosting human 3d pose estimation. In *Proc. 3DV*, pages 479–488, 2016.

[5] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[6] J. M. Graving, D. Chae, H. Naik, L. Li, B. Koger, B. R. Costelloe, and I. D. Couzin. Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning. *Elife*, 8:e47994, 2019.

[7] S. Günel, H. Rhodin, D. Morales, J. Campagnolo, P. Ramdya, and P. Fua. Deepfly3d, a deep learning-based approach for 3d limb and appendage tracking in tethered, adult drosophila. *Elife*, 8:e48571, 2019.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016.

[9] J. Hoffman, E. Tzeng, T. Park, J. Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *Proc. ICML*, pages 1989–1998, 2018.

[10] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2013.

[11] A. Kanazawa, S. Kovalsky, R. Basri, and D. Jacobs. Learning 3d deformation of animals from 2d images. *Computer Graphics Forum*, 35(2):365–374, 2016.

[12] C. Li and G. H. Lee. From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In *Proc. CVPR*, pages 1482–1491, 2021.

[13] S. Li, J. Li, H. Tang, R. Qian, and W. Lin. Atrw: A benchmark for amur tiger re-identification in the wild. In *Proc. Multimedia*, pages 2590–2598, 2020.

[14] T. Y. Lin, M. Maire, J. Belongie, S.and Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proc. ECCV*, pages 740–755, 2014.

[15] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9):1281–1289, 2018.

[16] J. Mei, X. Chen, C. Wang, A. Yuille, X. Lan, and W. Zeng. Learning to refine 3d human pose sequences. In *Proc. 3DV*, pages 358–366, 2019.

[17] MMPose. Openmmlab pose estimation toolbox and benchmark, 2021. https://mmpose.readthedocs.io/en/latest/index.html.

[18] J. Mu, W. Qiu, G. D. Hager, and A. L. Yuille. Learning from synthetic animals. In *Proc. CVPR*, pages 12386–12395, 2020.

[19] T. D. Pereira, D. E. Aldarondo, L. Willmore, M. Kislin, S. S. H. Wang, M. Murthy, and J. W. Shaevitz. Fast animal pose estimation using deep neural networks. *Nature Methods*, 16(1):117–125, 2019.

[20] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *Proc. ECCV*, pages 573–586, 2012.

[21] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *Proc. CVPR*, pages 5693–5703, 2019.

[22] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *Proc. ICCV*, pages 109–117, 2017.

[23] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao. Robust estimation of 3d human poses from a single image. In *Proc. CVPR*, pages 2361–2368, 2014.

[24] C. Wang, H. Qiu, A. L. Yuille, and W. Zeng. Learning basis representation to refine 3d human pose estimations. In *Proc. AAAI*, pages 8925–8932, 2019.

[25] X. Yu, F. Zhou, and M. Chandraker. Deep deformation network for object landmark localization. In *Proc. ECCV*, pages 52–70, 2016.

[26] X. Zhou, M. Zhu, S. Leonardos, and K. Daniilidis. Sparse representation for 3d shape estimation: A convex relaxation approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(8):1648–1661, 2016.

[27] S. Zuffi, A. Kanazawa, D. W. David Jacobs, and M. J. Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *Proc. CVPR*, pages 6365–6373, 2017.

[28] S. Zuffi, A. Kanazawa, and M. J. Black. Lions and tigers and bears: Capturing non-rigid, 3d, articulated shape from images. In *Proc. CVPR*, pages 3955–3963, 2018.

[29] S. Zuffi, A. Kanazawa, T. Berger-Wolf, and M. J. Black. Three-d safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In *Proc. ICCV*, pages 5359–5368, 2019.