ISSN: 2710-0588

An Animal Pose Estimation Method for Handling Keypoint Occlusion

Xiaoying Zhu[1, a], Yaning Jiang[2, *]

[1] School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China

[2] Faculty of Applied Sciences, Macao Polytechnic Institute, Macau, China

[a]18339617565@163.com, [*]18317694173@163.com

Abstract

Animal pose estimation plays a vital role in various fields such as animal behavior analysis and wildlife conservation. However, in real-world scenarios, keypoint occlusion frequently occurs, which limits the model's ability to accurately localize keypoints. To address this issue, this paper proposes an occlusion-aware animal pose estimation algorithm based on an improved HRNet. The proposed method incorporates a data augmentation strategy driven by strongly connected keypoints to enrich the diversity of occluded keypoints in the training set and enhance the model's inference capability for invisible keypoints. In addition, a dynamic upsampling module integrating both channel and spatial attention mechanisms is designed to improve the restoration quality of fine-grained features during the upsampling process. Furthermore, a progressive feature fusion strategy is introduced to reduce the information loss caused by large-scale upsampling in multi-scale feature integration, thereby further enhancing the fusion performance. Experimental results on the public AP-10K dataset demonstrate that the proposed method significantly outperforms the original HRNet and other comparison algorithms in terms of accuracy.

Keywords

Animal Pose Estimation, Keypoint Occlusion, Data Augmentation, Dynamic Upsampling.

1. Introduction

Animal pose estimation has emerged as an important application of computer vision in fields such as animal behavior analysis, biological research, and wildlife monitoring. In recent years, it has attracted increasing attention from the research community. Compared with human pose estimation, animal pose estimation faces greater challenges due to the significant variations in morphological structures, body proportions, and movement patterns across different species.

Most existing approaches draw inspiration from human pose estimation frameworks. Methods such as HRNet [1], OpenPose [2], and SimpleBaseline [3] have been widely adapted for animal keypoint detection tasks. Open-source toolkits like DeepLabCut [4] integrate deep learning with behavioral research, achieving promising results on small-scale animal datasets through transfer learning. However, these approaches typically rely on clean, unobstructed images and often struggle to generalize to complex environments where occlusions or visual ambiguities frequently occur.

To improve pose estimation performance under keypoint occlusion, several studies have introduced structural priors-such as skeletal graphs [5] or graph neural networks (GNNs) [6]-to enhance the model's understanding of semantic relationships among joints. However, most existing methods still rely primarily on conventional data augmentation techniques (e.g.,

ISSN: 2710-0588

occlusion patches [7], random erasing [8]) to simulate occlusions, which fail to fundamentally mitigate the prediction bias caused by severe keypoint occlusion.

Some works attempt to compensate for missing keypoints through structural reasoning modules or attention mechanisms. For instance, Masked Feature Completion combines spatial attention with feature restoration to directly generate substitute representations for occluded regions [9]. PoseRefiner-like approaches iteratively refine the predicted keypoints through secondary optimization to correct the initial predictions of occluded joints [10]. In addition, a few methods [11] model keypoint visibility during training, enabling the network to perceive occlusion states and adapt its predictions accordingly. Nevertheless, these methods typically rely on complex architectures or external priors, resulting in high computational and training resource requirements.

To address the limitations of existing animal pose estimation methods under occlusion scenarios, this paper proposes OCC-HRNet (Occlusion-aware HRNet), an enhanced framework that improves HRNet from three key perspectives: data augmentation, upsampling, and multi-scale feature fusion. Specifically, OCC-HRNet achieves three major innovations:

(1) A relation-driven data augmentation strategy is introduced to strengthen the model's structural awareness and feature completion capability for occluded keypoints;

(2) A hybrid attention-based dynamic upsampling module is designed to enhance fine-grained detail recovery during feature reconstruction;

(3) A progressive feature fusion strategy is proposed to mitigate the information loss caused by large-scale upsampling in traditional HRNet, thereby further improving the robustness and accuracy of pose estimation in complex environments.

2. The Proposed Method

To address the insufficient keypoint localization accuracy of HRNet under occlusion, we propose OCC-HRNet, a framework equipped with occlusion-aware capabilities. OCC-HRNet incorporates a triple optimization strategy-data augmentation, upsampling, and multi-scale feature fusion-to enhance keypoint detection in occluded scenarios. The overall architecture of OCC-HRNet is illustrated in Figure 1.

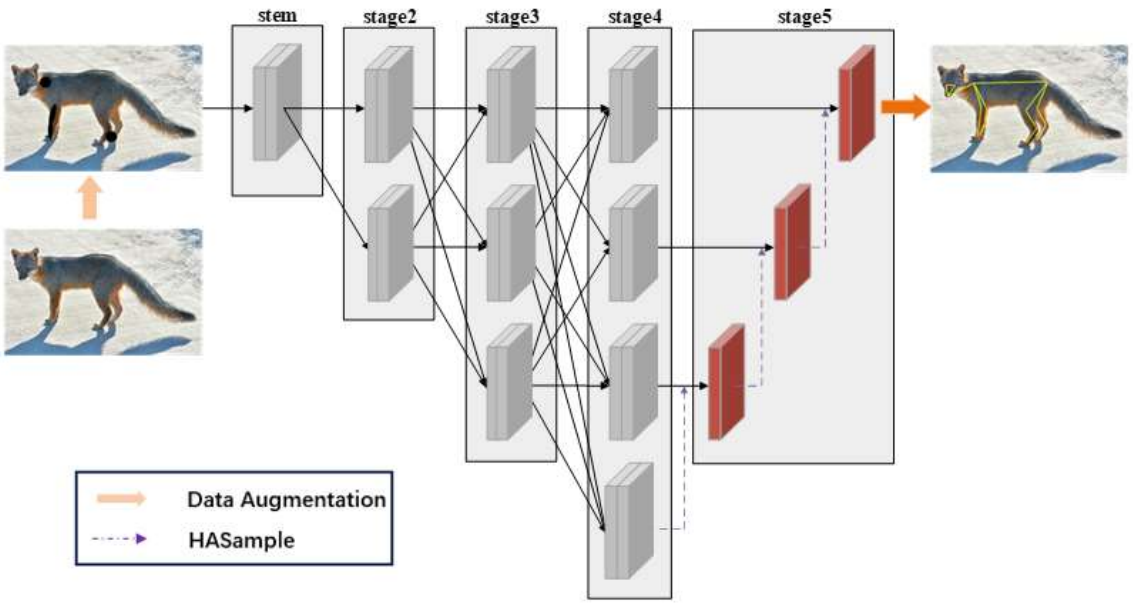

Figure 1. Architecture of OCC-HRNet

ISSN: 2710-0588

2.1. Strong Connectivity-Driven Data Augmentation Methods(SCDAM)

The strong-connection-driven data augmentation method is based on the strong connectivity relationships between keypoints, and selectively occludes certain regions of the image to enrich the dataset and improve model accuracy. In this work, we denote keypoints as K, with $K_i$representing the i-th keypoint. Figure 2 illustrates all keypoints of mammals.

In mammals, limb movements are achieved through the coordinated interaction of various body parts. The keypoints of the limbs and their interrelationships jointly determine the animal's morphology and motion posture. We hypothesize that the variation in connection strength among keypoints significantly affects the model's prediction performance-stronger connections lead to more accurate predictions for the associated keypoints. Based on this observation, we introduce the following definitions:

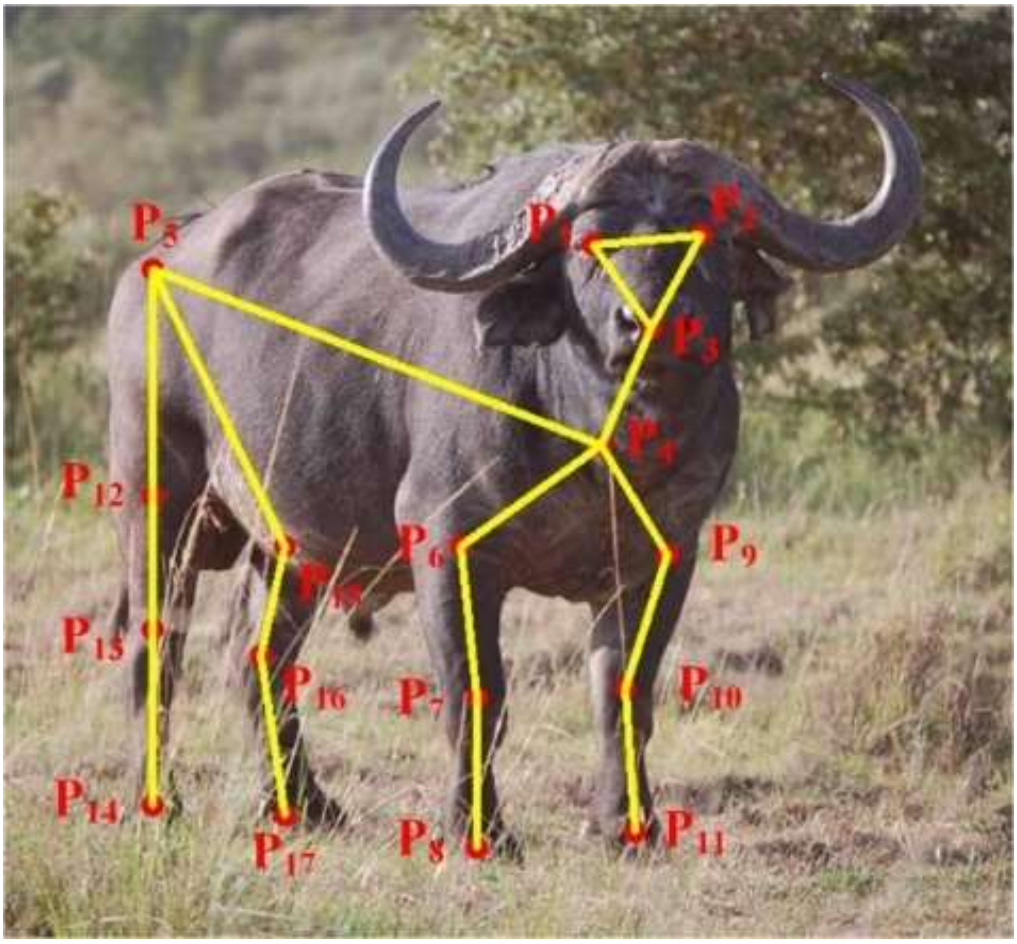

Figure 2. Anatomical body keypoints of mammals

(a) Strong Connection Between Keypoints: In the limbs of mammals, directly connected keypoints form strong connections. Formally, if $p_i$and $p_j$are directly connected, there exists a strong connection between them, denoted as (p_i, p_j). The strong connections present in Figure 2 include(p_6, p_7), (p_7, p_8), (p_9, p_{10}), (p_{10}, p_{11}), (p_{12}, p_{13}), (p_{13}, p_{14}), (p_{15}, p_{16}), (p_{16}, p_{17}).

Suppose a keypoint in Figure 2 is displaced in a certain direction or occluded; the model can leverage the strong connections(p_6, p_7) and $(p_7, p_8)$to infer the relative positions of $K_i$or K_j, thereby maintaining prediction accuracy under occlusion.

(b) Strong Connection Set: The set of all strong connections.

(c) Visible Keypoints: Keypoints in the input image that are not occluded.

ISSN: 2710-0588

(d) Visible Keypoint Set: The set of all visible keypoints in the input image, denoted as $P^v = \{p_1^v, p_2^v, \ldots, p_m^v\}$ where m is the number of visible keypoints.

To enhance the model's ability to predict occluded keypoints based on strong connections, we propose a strong-connection-driven data augmentation method built upon the above definitions. This method performs both connection-level occlusion and keypoint-level occlusion according to the strong connection set and the visible keypoint set, generating an augmented dataset.

In an image, if the cardinality of the visible keypoint set is large, partially occluding the image has little effect on model performance. Conversely, if the cardinality is too small, augmentation may hinder the model's ability to extract meaningful features. Therefore, we introduce an augmentation constraint, defined in Equation (1). In this equation, θ denotes the augmentation constraint threshold, and H is the total number of keypoints. Data augmentation is applied only when the image satisfies this constraint.

$$\left|P^v\right| / H > \theta \tag{1}$$

(a) Relationship Occlusion Strategy

In this study, relationship occlusion is performed according to Occlusion Strategies 1–3.

Occlusion Strategy 1:Let R_o denote the set of strong connections. $R_o \in \phi$

Occlusion Strategy 2: $(\forall p_i^v) p_i^v \in P^v$, $(\forall p_j^v) p_j^v \in P^v$, If a strong connection (p_i^v, p_j^v) exists,then$R_o = R_o \cup \{(p_i^v, p_j^v)\}$.

Occlusion Strategy 3: If $|P^v|/H > \theta$, then $\lceil |R_o| \mu_1 \rceil$ relationships are randomly selected from R_o for occlusion, where $\lceil \cdot \rceil$ denotes the ceiling operation, and μ_1 is a tunable parameter. The occlusion area is calculated according to Equation (2). In Equation (2), $h_0$and w_0 represent the height and width of the bounding box, respectively, and $\left\| p_i^v - p_j^v \right\|_2$ denotes the Euclidean distance between the two keypoints of the connection.

$$S_{occ1} = \Pi \left(\frac{\min(h_0, w_0)}{30} \right) \left\| p_i^v - p_j^v \right\|_2 \tag{2}$$

(b) Keypoint Occlusion Strategy

In this study, keypoint occlusion is performed according to *Occlusion Strategies 4–5*.

Occlusion Strategy 4: $(\forall p_i^v) p_i^v \in P^v$, If p_i^v has already been involved in a relationship occlusion, then$K^v \leftarrow P^v \setminus \{p_i^v\}$.

Occlusion Strategy 5: If $|P^v|/H > \theta$, then $\lceil |K^v| \mu_2 \rceil$ keypoints are randomly selected from K^v for occlusion, where μ_2 is a tunable parameter. The occlusion area is computed according to Equation (3).

$$S_{occ2} = \Pi \left(\frac{\min(h_0, w_0)}{30} \right)^2 \tag{3}$$

Figure 3 illustrates the occlusion results generated by the strong-connection-driven data augmentation method. In the figure, the black circular regions represent keypoint occlusions, while the black elliptical regions indicate relationship occlusions.

ISSN: 2710-0588

Figure 3. Comparison of data augmentation methods

2.2. Hybrid Attention-based Upsample Method(HASample)

During the fusion of feature maps at different resolutions, low-resolution feature maps are often upsampled using interpolation methods. Traditional interpolation techniques [12] generate new pixels by performing linear or nonlinear interpolation between known pixels, which may lead to a loss of image details. Under occlusion, such methods may fail to effectively capture information from the occluded regions. DySample [13] is a dynamic upsampling method that can generate an interpolation sampling set adaptively based on the input feature map and adjust the interpolation results using this set. This approach partially mitigates the detail loss caused by conventional interpolation.

In the process of generating the interpolation sampling set, DySample employs linear functions to produce offsets corresponding to the original sampling points. This offset generation can be regarded as extracting local contextual features. However, when occlusion is present, the model cannot infer the global spatial relationships of the occluded regions solely from local context. Moreover, this offset generation scheme struggles to effectively exploit the channel-wise relationships of the feature map, which provide important complementary contextual information. These limitations constrain the ability of the upsampled feature maps to preserve fine-grained details.

To address the above limitations, we propose an improved version of DySample, named HASample (Hybrid Attention-based Upsample Method). First, an original sampling set $\mathcal{G}$ is defined according to the size of X_{in}. Next, the O-Generator module generates offsets corresponding to the input feature map, which are then reshaped via a Pixel Shuffle operation to obtain the dynamic offsets $\mathcal{O}_1$. These offsets are fused with the original sampling set $\mathcal{G}$ to form the interpolation sampling set S_1. Finally, S_1 and the input feature map are fed into a bilinear interpolation grid function (GridSample) to produce the upsampled feature map. The overall architecture of HASample is illustrated in Figure 4.

In HASample, the O-Generator module plays a key role, and its detailed structure is illustrated in Figure 4. It consists of a channel attention module and a spatial attention module arranged in series, which weight the input features along the channel and spatial dimensions, respectively, to generate more accurate sampling offsets. Specifically, the channel attention module captures dependencies across different channels, enabling the model to focus on feature channels that contribute most to offset generation. Meanwhile, the spatial attention module employs convolutional kernels of three different sizes to extract multi-scale global spatial features, comprehensively evaluating the importance of features at different spatial locations within the feature map.

ISSN: 2710-0588

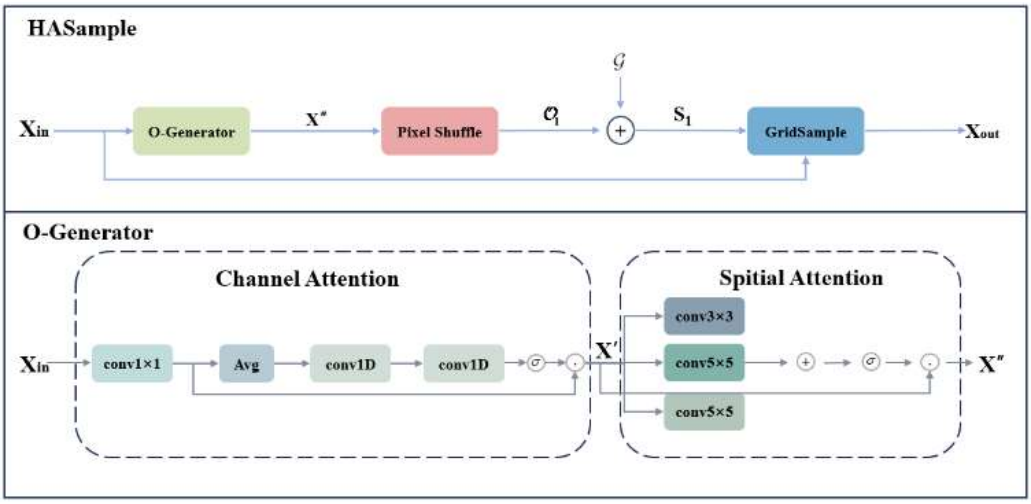

Figure 4. Overview of the HASample model architecture

Given an input feature map X_{in}, the O-Generator first applies the channel attention module to extract channel-wise features and generate X'. It then employs the spatial attention module to capture spatial features, producing X'', which serves as the offset map for the input feature map.

The channel attention module first compresses the channels of the feature map using a pointwise convolution. It then sequentially applies global average pooling to reduce the spatial dimensions and stacks two 1D convolutions to extract channel-wise features. The resulting feature map is passed through a sigmoid function to generate a weight matrix, which is then multiplied element-wise with the channel-compressed feature map to obtain X'. The detailed computation process of the channel attention module is formulated in Equation (4).

$$X' = \sigma(conv1D(conv1D(Avg(conv_{1\times1}(X_{in}))))) \odot conv_{1\times1}(X_{in}) \tag{4}$$

The spatial attention module first applies single-kernel convolutions of sizes 3×3, 5×5, 7×7 to compress the channel features while extracting spatial features at different scales. The resulting multi-scale spatial feature maps are then fused through element-wise summation followed by a sigmoid function to generate a weight matrix. Finally, the weight matrix is multiplied element-wise with 1 to obtain X''. The detailed computation of the spatial attention module is formulated in Equation (5).

$$X'' = \sigma(conv_{3\times3}(X') + conv_{5\times5}(X') + conv_{7\times7}(X')) \odot X' \tag{5}$$

2.3. Step-by-step Feature Fusion Module(SSFFM)

Fusing feature maps at different resolutions can yield richer and more informative representations. In HRNet, the fusion strategy involves upsampling the (1/8), (1/16), and (1/32) resolution feature maps by factors of 2, 4, and 8, respectively, to match the (1/4) resolution, followed by convolutional compression and concatenation. However, the large-scale upsampling (4× and 8×) performed during this process may introduce detail distortion in the fused feature maps.

To address this issue, we propose a progressive feature fusion module. The fusion process is illustrated in Figure 5. Starting from the (1/32) resolution feature map, it is first upsampled by a factor of 2 using HASample, and the resulting feature map is fused with the feature map of the same resolution. The fused feature map is then progressively upsampled using HASample and merged with the corresponding resolution feature map at each stage. This process is repeated until fusion with the original (1/4) resolution feature map is completed.

ISSN: 2710-0588

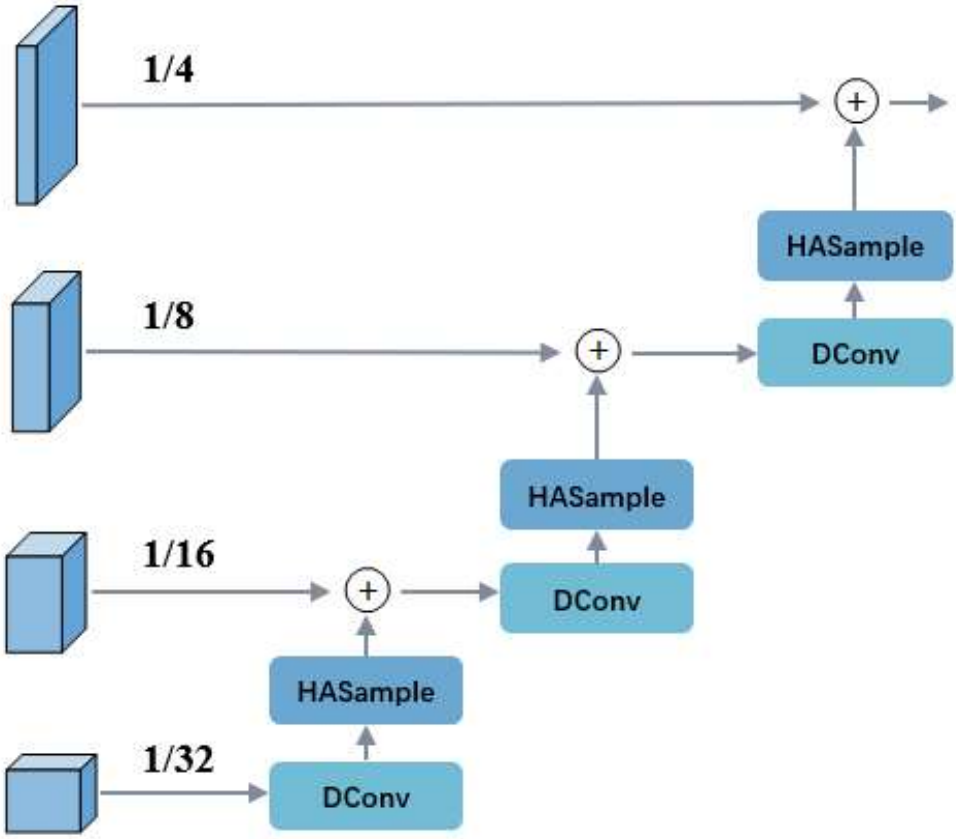

Figure 5. Step-by-step feature fusion Module

3. Experimental Results and Analysis

3.1. Experimental Setup

All experiments in this study were conducted on the same server. The server is equipped with an Intel(R) Xeon(R) Platinum 8260 @ 2.4GHz CPU, 376 GB of RAM, and an NVIDIA GeForce RTX 3090 (24 GB) GPU. The operating system is Ubuntu 20.04.1, and the deep learning framework is PyTorch 2.0.0, with Python 3.7 used for algorithm implementation.

To ensure fairness, all training procedures employ a uniform input image size of (256 \times 256), a batch size of 64, and 260 epochs. All other training parameters are set to their default values. All models are trained from scratch, without using pretrained weights from large-scale datasets such as ImageNet for initialization.

3.2. Dataset

The experiments in this study utilize the AP-10K dataset [14], which covers 23 families and 54 species. The dataset contains 10,015 images with a total of 13,028 instances. Each instance is annotated with 17 keypoints, representing 17 critical anatomical landmarks of mammals. Following the official AP-10K split, the dataset is divided into 7,023 training images, 995 validation images, and 1,997 test images.

3.3. Evaluation Metrics

This study adopts the widely recognized evaluation metric in the field of pose estimation-Object Keypoint Similarity (OKS) [15]. The calculation formula of OKS is shown in Equation (6).

$$OKS = \frac{\sum_i e^{\frac{-d_i^2}{2s^2k_i^2}} \delta(\upsilon_i > 0)}{\sum_i \delta(\upsilon_i > 0)} \tag{6}$$

ISSN: 2710-0588

In Equation (6), d_i represents the L2 distance between the predicted and ground truth keypoints, s denotes the object scale, k is a constant that controls the falloff for each keypoint, and v_i indicates whether the keypoint is visible.

This study uses the Average Precision (AP) based on OKS as the primary evaluation metric for the model.The calculation of AP is shown in Equation (7). In this equation, n represents the number of instances, T_k denotes the manually set OKS threshold, $T_k = 0.5 + 0.05k$,and k take values ranging from 0 to 9.

$$AP = \frac{\sum_{k=0}^{9} \frac{1}{n} \sum_{p=1}^{n} \delta(OKS > T_k)}{10} \tag{7}$$

3.4. Ablation Study

To evaluate the effectiveness of the proposed improvements, ablation experiments were conducted on the AP-10K dataset. The results are shown in Table 1.In the table, Model0 represents the baseline model (HRNet). Model1, Model2, and Model3 denote the models obtained by successively incorporating SCDAM, SSFFM, and HASample into the preceding model.

As shown in Table 1, Model1, Model2, and Model3 improve the performance of Model0, Model1, and Model2 by 0.5%, 0.5%, and 0.7%, respectively.These results clearly demonstrate the effectiveness of SCDAM, SSFFM, and HASample.

Table 1. Ablation results of all improvements in OCC-HRNet

Methods	improvements			AP	params(M)	FLOPs(G)
	SCDAM	SSFFM	HASample			
Model0				70.3	28.5	10.27
Model1	√			70.8	28.5	10.27
Model2	√	√		71.3	28.6	10.27
Model3	√	√	√	**72.0**	28.6	10.29

3.5. Comparative Experiments

This study evaluates the performance of OCC-HRNet by comparing its detection accuracy on the AP-10K dataset with seven different pose estimation methods.These seven methods are CSPNeXt-m [16], CSPNeXt-s, Hourglass [11], SimpleBaseline (ResNet_50) [3], SimpleBaseline (ResNet_101), HRNet, and DARK [17].Among them, DARK uses HRNet as its backbone network and applies the DARK technique for post-processing of the heatmaps.

Table 2 presents the results of the comparative experiments. As shown, the OCC-HRNet achieves an AP of 72.0% on the AP-10K dataset, representing a 1.7% improvement over HRNet.On the AP-10K dataset, OCC-HRNet also improves the AP50, AP75, AP^M, AP^L, and AR metrics by 0.3%, 0.6%, 1.1%, 1.8%, and 1.2%, respectively, compared to HRNet.These results indicate that OCC-HRNet demonstrates higher performance than HRNet, particularly in scenarios involving keypoint occlusion.When compared with other algorithms, OCC-HRNet also shows clear advantages. The AP of OCC-HRNet exceeds that of CSPNeXt-m, Hourglass, SimpleBaseline (ResNet_101), and DARK by 1.5%, 3.4%, 2.2%, and 0.8%, respectively.These findings further confirm the superior performance of OCC-HRNet under occluded keypoint conditions.

The model parameters of OCC-HRNet amount to 28.6M, and its computational complexity is 10.29 GFLOPs. Compared with HRNet, OCC-HRNet shows only a slight increase in model

ISSN: 2710-0588

parameters, indicating that the proposed method enhances detection accuracy without significantly increasing model or computational complexity.

Table 2. Comparative results of OCC-HRNet on AP-10K

Methods	Backbone	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	params(M)	FLOPs(G)
CSPNeXt [16]	CSPNeXt-m	70.5	93.6	77.6	52.4	71.0	73.7	13.6	2.56
CSPNeXt	CSPNeXt-s	60.5	88.7	63.2	47.3	60.7	64.8	6.02	1.91
Hourglass[11]	Hourglass	68.6	93.1	74.2	52.5	69.6	72.6	94.8	28.7
SimpleBaseline[3]	ResNet_50	68.0	92.6	73.8	55.2	68.7	71.8	34.0	7.26
SimpleBaseline	ResNet_101	69.8	84.1	73.9	66.2	69.0	71.8	53.1	12.1
HRNet[1]	HRNet_w32	70.3	93.7	77.2	53.5	70.5	73.8	28.5	10.27
DARK[17]	HRNet_w32	71.2	93.9	77.5	53.9	71.7	74.4	28.5	10.27
OCC-HRNet	OCC-HRNet	**72.0**	**94.0**	**77.8**	**54.6**	**72.3**	**75.0**	28.6	10.29

3.6. Comparison of Detection Results

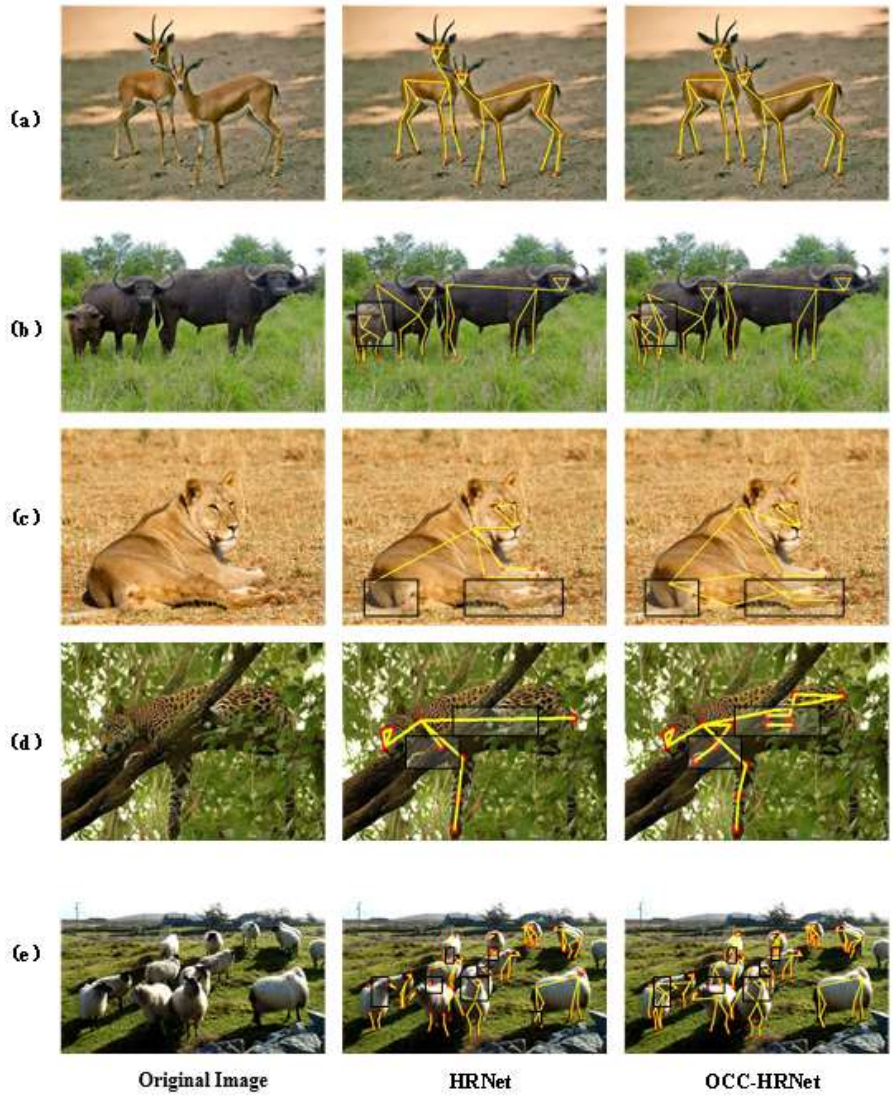

Figure 6. Comparison of results between HRNet and OCC-HRNet

ISSN: 2710-0588

Figure 6 presents several examples of keypoint detection results produced by OCC-HRNet and HRNet on the AP-10K dataset.Figure 6(a) shows the detection results under non-occluded scenarios. In this case, both OCC-HRNet and HRNet accurately detect the animal keypoints.Figure 6(b) illustrates the results under inter-animal occlusion. As shown, HRNet fails to detect several keypoints, including the unoccluded tail and the occluded hip and knee joints, whereas OCC-HRNet successfully detects all of them.Figure 6(c) depicts the results under self-occlusion scenarios. HRNet misses multiple keypoints on the legs, while OCC-HRNet accurately locates the occluded ones.

Figure 6(d) demonstrates occlusions caused by environmental factors. In this situation, HRNet misses a large number of keypoints obscured by branches, while OCC-HRNet successfully identifies all of them, indicating its strong generalization and robustness.This advantage is further confirmed in Figure 6(e), which presents a heavily occluded multi-target scenario. In the dense central region of the image, HRNet detects only a few keypoints and fails to establish correct skeletal connections. In contrast, OCC-HRNet accurately predicts most occluded keypoints, missing only a few on the legs.

4. Summary

To address the poor performance of existing algorithms in handling occluded keypoints in animal pose estimation, this paper proposes an improved OCC-HRNet model based on HRNet. The improvements in this work can be summarized as follows:

(1) A strong-connection-driven data augmentation method is proposed to adaptively perform occlusion augmentation on the dataset.

(2) A hybrid attention-based dynamic upsampling method is introduced to alleviate detail distortion during the upsampling process.

(3) A progressive feature fusion strategy is adopted to reduce information loss caused by large-scale upsampling.

Experimental results on the AP-10K dataset demonstrate the effectiveness and efficiency of the proposed method.

References

[1] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York, USA: IEEE, 2019: 5693-5703.

[2] Cao Z, Simon T, Wei S E, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017: 7291-7299.

[3] Xiao B, Wu H, Wei Y. Simple baselines for human pose estimation and tracking[C]// Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018: 466-481.

[4] Mathis A, Mamidanna P, Cury K M, et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning[J]. Nature Neuroscience, 2018, 21(9): 1281–1289.

[5] Zeng A, Sun X, Yang L, et al. Learning Skeletal Graph Neural Networks for Hard 3D Pose Estimation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021: 2282-2291.

[6] ZHAO Chenyang, WANG Yizhou, QIAO Yu, et al. Graph-PCNN: Two stage human pose estimation with graph pose refinement[C]// Proceedings of the 16th European Conference on Computer Vision (ECCV 2020). Glasgow: Springer, 2020: 492-508.

[7] DeVries T, Taylor G W. Improved regularization of convolutional neural networks with Cutout[C]// Proceedings of the 2017 NIPS Workshop. Long Beach: NIPS, 2017.

ISSN: 2710-0588

[8] Zhong Z, Zheng L, Kang G, et al. Random Erasing Data Augmentation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu: AAAI, 2020, 34(07): 13001-13008.

[9] LIU Y, ZHANG J, WANG X, et al. Masked feature completion for occlusion-aware human pose estimation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022: 12345-12355.

[10] Fieraru M, Khoreva A, Pishchulin L, et al. Learning to refine human pose estimation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2018: 205-214.

[11] NEWELL A, YANG K, DENG J. Stacked hourglass networks for human pose estimation[C]// European Conference on Computer Vision. Amsterdam: Springer, 2016: 483-499.

[12] DONG C, LOY C C, HE K, et al. Learning a deep convolutional network for image super-resolution[C]// Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer, 2014: 184-199.

[13] Xia Z H, Wang Y J, He S C, et al. DySample: Dynamic Sampling for Efficient Upsampling[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022: 12833–12843.

[14] YANG J, LI C, ZHANG P, et al. AP-10K: A benchmark for animal pose estimation in the wild[C]// Advances in Neural Information Processing Systems. 2021: 1-12.

[15] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common Objects in Context[C]// European Conference on Computer Vision (ECCV). Zurich: Springer, 2014: 740–755.

[16] Liu Yujian, Wang Kaipeng, Zhang Xiangxiang, et al. CSPNeXt: Hierarchical Split-and-Aggregate MLP for Lightweight Object Detection and Segmentation[J]. arXiv preprint arXiv:2207.09462, 2022.

[17] Zhang F, Zhu X, Dai Y, et al. Distribution-Aware Coordinate Representation for Human Pose Estimation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020: 7093–7102.