ZebraPoseNet: A Deep UNet-Based Framework with Attention Mechanisms for Animal Pose Estimation

Hongqian Yu

Penn state University, Pennsylvania, US

Hky5209@psu.edu

Abstract. Animal pose estimation plays a crucial role in wildlife monitoring, behavioral analysis, and conservation research. However, zebras present unique challenges due to their visually complex striped patterns, frequent occlusions in natural environments, and the scarcity of annotated datasets. Conventional deep learning models, such as UNet, struggle with feature extraction and keypoint localization in such scenarios, leading to reduced accuracy and generalization issues. In this study, we propose an improved UNet model that incorporates Squeeze-and-Excitation (SE) attention mechanisms and transfer learning to enhance keypoint detection accuracy. The SE blocks allow the model to emphasize important spatial features while suppressing background noise, and the transfer learning approach leverages knowledge from larger animal pose datasets to improve performance on limited zebra data. We evaluate our model on a custom-labeled zebra dataset and optimize it with a hybrid loss function. Experimental results demonstrate that our approach significantly reduces the mean per-keypoint error (MPKE) by 15% compared to the baseline UNet model, highlighting its effectiveness in real-world applications.

Keywords: Animal Pose Estimation, SE-Enhanced UNet, Transfer Learning, Intelligent Environmental Sensing.

1. Introduction

Animal pose estimation is a crucial research topic in computer vision with broad applications in wildlife monitoring, behavioral analysis, veterinary diagnostics, and conservation efforts. The ability to track the posture and movement of animals in their natural habitats provides invaluable insights into their behavior, social interactions, and overall health. Automated pose tracking enables researchers to study migration patterns, detect behavioral anomalies, and assess injuries in wildlife populations (Mathis et al., 2018). Non-invasive monitoring of animal movement also contributes to conservation strategies, aiding efforts to mitigate human-wildlife conflicts and ensure the sustainability of endangered species. In veterinary and agricultural applications, pose estimation is employed to detect locomotion disorders, identify early signs of illness in livestock, and improve animal welfare through precision farming technologies (Zuffi et al., 2018). Additionally, the study of animal biomechanics through pose estimation supports research in robotics and bio-inspired engineering, where insights from animal locomotion can influence the design of autonomous systems and artificial limbs.

Despite the significant progress in human pose estimation using deep learning, applying similar techniques to animals remains challenging due to species-specific anatomical variations, occlusions, and environmental complexities. Unlike human pose estimation, which benefits from extensive datasets such as COCO (Lin et al., 2014) and MPII (Andriluka et al., 2014), the availability of annotated datasets for non-human species is limited. Zebras present additional difficulties due to their unique striped coat patterns, which introduce texture-based ambiguities that can confuse feature extraction models. The visual similarity between different body regions complicates keypoint identification, leading to localization errors. Furthermore, zebras are often found in herds, where occlusions caused by overlapping individuals or environmental obstacles such as grass and trees create further challenges for deep learning models. These occlusions, combined with variations in lighting, body orientation, and motion blur, necessitate robust and adaptable models that can generalize effectively across diverse conditions.

To address these limitations, we propose an enhanced UNet model that integrates Squeeze-and-Excitation (SE) attention mechanisms to refine keypoint localization and suppress background noise. SE blocks dynamically recalibrate channel-wise feature representations, allowing the model to focus on more informative features while ignoring irrelevant details (Hu et al., 2018). By introducing SE blocks at multiple levels of the encoder, our model enhances its ability to differentiate keypoint-relevant regions from complex backgrounds, improving robustness to occlusions and intra-class variations. Additionally, we apply transfer learning from large-scale animal pose estimation datasets to improve model generalization on limited zebra data. Specifically, we employ a ResNet-based encoder pretrained on large-scale animal recognition datasets and fine-tune it for zebra keypoint detection. This approach allows the model to capture general animal structural features while adapting to species-specific pose variations.

Our key contributions include an attention-enhanced UNet architecture that improves keypoint detection by refining feature extraction, a transfer learning-based training strategy that leverages larger datasets to enhance model performance, and the construction of a custom zebra dataset with detailed keypoint annotations for benchmarking pose estimation models. The effectiveness of our approach is validated through a comprehensive experimental analysis, demonstrating significant performance improvements over baseline methods. This study highlights the importance of domain-specific model enhancements in animal pose estimation and sets a foundation for future work in multi-animal tracking, 3D keypoint estimation, and real-time deployment of pose estimation systems in wildlife monitoring applications.

2. Methodology

This section describes the dataset, preprocessing techniques, network architecture, training strategy, and evaluation methodology used in our study. Our dataset consists of zebra images annotated with keypoint locations corresponding to various anatomical landmarks, including the head, torso, and limbs. These annotations are manually curated and stored in the COCO-style format to ensure compatibility with established pose estimation frameworks. The dataset includes 1000 training images, 200 validation images, and 200 test images, with each image annotated with eight keypoints (e.g., snout, neck, front legs, hind legs, and tail). Since zebras appear in different lighting conditions, poses, and occlusion levels, we applied augmentation techniques such as random rotation, flipping, scaling, brightness and contrast variation, Gaussian noise addition, and synthetic occlusion to simulate real-world conditions.

Our proposed model is an improved UNet-based keypoint detection framework, incorporating Squeeze-and-Excitation (SE) attention mechanisms and a ResNet backbone to enhance feature extraction. The encoder consists of a ResNet-based feature extractor initialized with ImageNet-pretrained weights, replacing traditional convolutional layers in UNet. SE attention blocks (Hu et al., 2018) are added to each encoder layer to refine feature selection and suppress background noise. The SE module operates in three main steps: squeeze, excitation, and scaling. First, global average pooling is applied to the feature maps, summarizing spatial information into a channel descriptor. Next, a two-layer fully connected network learns the channel-wise dependencies using non-linearity (ReLU activation), followed by a sigmoid activation to generate attention weights. Finally, these weights rescale the original feature maps, enhancing the most important features while suppressing irrelevant ones. The SE operation is formulated as follows:

$$s_c = F_{\delta q}(U_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} U_c(i,j)$$

$$z = \sigma(W_2 \delta(W_{1\delta}))$$

$$ildeU_C = z_c U_c$$

Where Uc is the input feature map of channel c, sc is the squeezed channel descriptor W1, W2 and are the parameters of the fully connected layers δ, is the ReLU function, and σ is the sigmoid activation.

Additionally, a multi-scale attention module is incorporated at the bottleneck layer to improve the model's global contextual awareness. This module processes features at different receptive fields by applying dilated convolutions with varying dilation rates. Let F(x) represent the feature map at a given layer, the multi-scale attention response is computed as:

$$F_{m\delta}(\chi) = \sum_{d \in D} W_d * F(\chi)$$

Where *D* is a set of dilation rates and Wd are the convolutional kernels with different receptive fields. By aggregating features from multiple scales, the network improves robustness to spatial variations in zebra poses.

The decoder mirrors the encoder, using transposed convolutions for upsampling and skip connections to retain spatial information. The final output consists of heatmaps for each keypoint, where pixel intensities indicate keypoint presence probability, and a soft-argmax layer (Luvizon et al., 2019) extracts precise keypoint coordinates from the heatmaps.

To optimize the model for accurate keypoint detection, we employ a multi-task loss function that balances classification-based heatmap regression and coordinate-level refinement. The total loss function is defined as follows:

$$\Gamma = \lambda_1 \Gamma_{heatmap} + \lambda_2 \Gamma_{regre\delta\delta on}$$

Where:

$\Gamma_{heatmap}$ is the binary cross-entropy (BCE) loss applied to the predicted heatmaps, encouraging the model to output probability distributions that highlight keypoint locations.

$\Gamma_{regre\delta\delta ion}$ is the mean squared error (MSE) loss applied to the soft-argmax coordinates, refining precise localization.

λ_1 and λ_2 are weighting factors set empirically (0.8 and 0.2, respectively) to balance heatmap confidence and coordinate accuracy.

The BCE loss function is defined as:

$$\Gamma_{heatmap} = -\sum_i \left[yi \log \hat{y}i + (1 - yi) \log(1 - \hat{y}i) \right]$$

Where yi is the ground truth heatmap value and $\hat{y}i$ is the predicted probability. The MSE loss for keypoint refinement is computed as:

$$\Gamma regre\delta\delta on = \frac{1}{N} \sum_{i=1}^{N} \left\| \hat{p}i - pi \right\|^2$$

Where pi represents the ground truth keypoint coordinates, and $\hat{p}i$ represents the predicted coordinates extracted from the soft-argmax operation.

The model is trained using the AdamW optimizer, with an initial learning rate of 0.001 and a cosine annealing scheduler. The training process consists of 50 epochs, a batch size of 16, and early stopping based on validation loss to prevent overfitting. The model is trained using mixed precision training to optimize memory usage. For evaluation, we use mean per-keypoint error (MPKE) to measure the average Euclidean distance between predicted and ground-truth keypoints, percentage of correct keypoints (PCK@0.1) to assess keypoint detection accuracy, and precision-recall analysis to evaluate detection reliability. The next section presents the results and discusses the model's performance against baseline approaches.

3. Results and Discussion

To evaluate the performance of our proposed model, we compare it against baseline methods using quantitative and qualitative analyses. The model is assessed on our custom zebra dataset using standard pose estimation metrics, including mean per-keypoint error (MPKE), percentage of correct keypoints (PCK@0.1), and precision-recall evaluation. Our improved UNet model significantly outperforms the baseline UNet and other commonly used architectures, demonstrating enhanced accuracy, robustness to occlusions, and improved generalization to unseen zebra images. The table below presents a comparative analysis of different models evaluated on our dataset. Our improved UNet model achieves a 15% reduction in MPKE compared to the baseline UNet, leading to a more precise localization of zebra keypoints. The PCK@0.1 metric also indicates superior performance, demonstrating the model's ability to correctly identify keypoints within a 10% tolerance range of the ground-truth locations.

Model	MPKE(px)	PCK@0.1(%)	Precision(%)	Recall(%)
OpenPose	13.2	72.3	79.8	76.2
HRNet	11.7	76.5	82.1	78.8
Baseline UNet	12.5	75.0	80.2	78.5
Improved UNet	10.6	81.2	85.4	83.1

Figure 1. Performance Comparison of Pose Models

Our approach exhibits substantial improvement in PCK and precision-recall scores, indicating that the attention-enhanced UNet effectively localizes keypoints with higher confidence and fewer false positives. Figure 2 presents qualitative results comparing keypoint predictions from the baseline UNet and our improved model. The proposed model consistently localizes keypoints more accurately, even in cases of occlusion and complex body postures. Notably, while the baseline UNet struggles to correctly detect keypoints when the zebra's body overlaps with another zebra, our model successfully resolves such ambiguities using SE attention-enhanced feature extraction. We also evaluate failure cases where keypoint localization remains challenging, such as extreme occlusions where no visible cues exist for certain keypoints. In these instances, our model demonstrates higher reliability than traditional approaches but still has limitations in scenarios where keypoints are entirely obscured. Future work may incorporate temporal consistency models or 3D pose estimation techniques to further improve robustness in such cases.

Model Variant	MPKE(px)	PCK@0.1(%)
Baseline UNet	12.5	75.0
+SE Attention	11.1	79.5
+Multi-Scale Attention	10.9	80.3
Final Improved Model	10.6	81.2

Figure 2. Impact of Attention Modules on UNet Model Performance

To assess the contribution of each model component, we perform an ablation study analyzing how different architectural modifications impact performance. We evaluate the following configurations: (1) baseline UNet, (2) UNet with SE attention, (3) UNet with multi-scale attention, and (4) our final improved model incorporating both SE attention and transfer learning.

The ablation study confirms that both SE attention and multi-scale attention contribute significantly to performance improvements. The SE blocks enhance the model's ability to suppress irrelevant background features, while multi-scale attention improves robustness to pose variations. The combination of these components, along with transfer learning, yields the highest accuracy and lowest MPKE. Our experimental results demonstrate that the proposed improvements to the UNet architecture lead to superior pose estimation performance in zebras compared to existing methods.

The integration of SE attention enhances feature discrimination, allowing the model to focus on relevant anatomical structures while ignoring distractors. The use of transfer learning accelerates convergence and enables the model to leverage knowledge from large-scale animal pose datasets, reducing the need for extensive labeled zebra images. Despite these advancements, some limitations remain. The model struggles with extreme occlusions where keypoints are entirely hidden, and its performance degrades under significant motion blur. Future research could explore multi-frame temporal modeling, integrating recurrent architectures such as LSTMs or transformer-based sequence models to improve prediction consistency over time. Another potential direction is to extend this framework to 3D pose estimation, leveraging depth information to refine keypoint localization further. The deployment of real-time inference pipelines for field applications is another avenue for future research, facilitating automated wildlife tracking and behavioral analysis in ecological studies. Our findings establish a strong foundation for future research in zebra pose estimation and broader applications in animal keypoint detection. The proposed model not only enhances accuracy but also provides a generalizable framework that could be applied to other species with complex visual textures and occlusion challenges.

To better understand the practical performance of our improved UNet, we conducted a case study analyzing specific scenarios where the model excels and areas where it still faces challenges. The model performs particularly well in detecting keypoints under normal lighting conditions and minimal occlusions, achieving high accuracy in correctly localizing all keypoints. It also generalizes well to moderate occlusions, where a portion of the zebra's body is hidden by vegetation or another animal. The SE attention mechanism helps the network focus on distinguishing keypoint-relevant regions from background distractions, improving robustness.

However, certain failure cases were observed, particularly under extreme occlusions and motion blur. When more than 50% of the zebra’s body is occluded, the model struggles to infer keypoint locations accurately due to insufficient visual context. Similarly, in high-speed movement scenarios, motion blur distorts the texture patterns, leading to incorrect keypoint placements. Additionally, instances where zebras overlap heavily in herd formations pose additional challenges, as keypoints may be mistakenly assigned to the wrong individual.

Despite these limitations, our findings demonstrate that the integration of SE attention and multi-scale feature extraction significantly improves keypoint detection performance. Future improvements could involve incorporating temporal modeling techniques, such as LSTM-based sequence models or transformer architectures, to improve robustness in sequential image frames. Additionally, multi-view learning approaches could be explored to enhance pose estimation in densely populated environments.

4. Conclusion

In this study, we proposed an improved UNet model incorporating Squeeze-and-Excitation (SE) attention mechanisms and transfer learning to enhance zebra pose estimation. Our approach addresses key challenges in animal pose estimation, including occlusions, visual ambiguities caused by striped textures, and the lack of large-scale annotated datasets. By integrating SE attention, our model effectively enhances feature selection, reducing background noise and improving keypoint localization. Additionally, leveraging transfer learning enables our model to generalize better on limited zebra pose data, significantly reducing the mean per-keypoint error (MPKE) while improving overall precision and recall. Experimental results confirm that our enhanced UNet model outperforms baseline methods, including standard UNet and other widely used architectures, by achieving a 15% reduction in MPKE and superior performance in percentage of correct keypoints (PCK) evaluations. Our ablation study further demonstrates that SE attention and multi-scale feature extraction contribute significantly to the model’s improved accuracy and robustness.

Despite these advancements, certain limitations remain. Our model exhibits reduced accuracy when keypoints are entirely occluded or when images contain excessive motion blur. Addressing

these challenges requires further research into temporal modeling techniques, such as integrating recurrent architectures like LSTMs or transformer-based models for sequential frame processing. Additionally, extending our approach to 3D pose estimation could provide more accurate localization by leveraging depth information, particularly for applications requiring multi-animal tracking. Another important future direction is optimizing real-time inference capabilities to facilitate deployment in wildlife conservation efforts, allowing for automated tracking and behavioral analysis in real-world field environments.

In summary, this work contributes to advancing pose estimation techniques for zebras and provides a framework that can be extended to other animal species facing similar pose estimation challenges. By refining keypoint localization through attention-based mechanisms and leveraging domain-adaptive learning strategies, we establish a strong foundation for future developments in deep learning-based animal pose estimation. Our results demonstrate the potential for improving animal monitoring, conservation efforts, and veterinary diagnostics, ultimately enabling more efficient and scalable automated tracking systems for wildlife research and beyond.

References

[1] Andriluka, M., Pishchulin, L., Gehler, P., & Schiele, B. (2014). 2D human pose estimation: New benchmark and state of the art analysis. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3686-3693.

[2] Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2019). OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 43 (1), 172-186.

[3] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7132-7141.

[4] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. European Conference on Computer Vision (ECCV), 740-755.

[5] Luvizon, D. C., Picard, D., & Tabia, H. (2019). Human pose regression by combining indirect part detection and contextual information. Pattern Recognition, 94, 54-64.

[6] Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. Nature Neuroscience, 21 (9), 1281-1289.

[7] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 234-241.

[8] Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5693-5703.

[9] Zuffi, S., Kanazawa, A., Jacobs, D. W., & Black, M. J. (2018). 3D menagerie: Modeling the 3D shape and pose of animals. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5524-5532.