Primate Spatio-Temporal Relational Network for Unified Behavior Quantification in Cage Environments

Zhiyuan Chen[1,2], Shuangqingyue Zhang[2,3], Xueyin Li[4], Zongli Jiang[1], Xibo Ma[2,5*]
[1]School of Computer Science, Beijing University of Technology, Beijing, China
[2]MAIS, Institute of Automation Chinese Academy of Sciences, Beijing, China
[3]College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China
[4]State Grid Gansu Electric Power Company, Gansu, China
[5]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
*Corresponding author: xibo.ma@ia.ac.cn

***Abstract*—Quantifying macaque position, identity, and action is essential for neuroscience and pharmacology studies, yet remains challenging in cage environments due to frequent occlusions and rapid limb movements. This paper introduces the Primate Spatio-Temporal Relational Network (PSTRN), a framework designed to extract robust behavioral information from short video clips. PSTRN combines a hierarchical spatial encoder with a cross-frame relational attention module that applies temporal attention across all frames, adaptively aggregating complementary visual information to resolve ambiguities caused by occlusion or motion blur. A hierarchical 3D convolutional fusion then integrates multi-scale spatio-temporal features to jointly capture fine-grained limb motions and global body dynamics, enabling unified prediction of bounding boxes, identities, and actions. Experiments show that PSTRN achieves state-of-the-art performance, with 99.5% mAP_{50} and 85.1% F1 score on the single-macaque benchmark, and 92.6% mAP_{50} and 82.0% F1 score on the multi-macaque benchmark. These results demonstrate that PSTRN effectively captures macaque dynamics in realistic cage environments and offers a reliable solution for comprehensive behavior quantification.**

Keywords*—*Primate behavior analysis, Spatio-temporal feature learning, Cross-frame relational attention, Multi-scale temporal fusion

I. Introduction

Macaques serve as valuable experimental models for functional state assessment and drug safety evaluation due to their close genetic and physiological similarity to humans [1], [2]. Quantifying natural behaviors, including actions, spatial location, and individual identity, is essential for assessing functional status and understanding changes induced by experimental conditions [3], [4]. However, behavioral monitoring in laboratory-housed macaques still relies predominantly on manual observation or task-driven training protocols [5], [6]. Manual monitoring is labor-intensive and susceptible to observer influence [7], while training macaques to perform predefined actions often alters spontaneous activity patterns and limits experimental scalability [8].

Recently, video-based behavioral analysis has gained increasing attention as a non-invasive and scalable alternative [9], [10]. Yet accurately detecting macaque position, identity, and actions in cage-housed environments remains challenging [11].

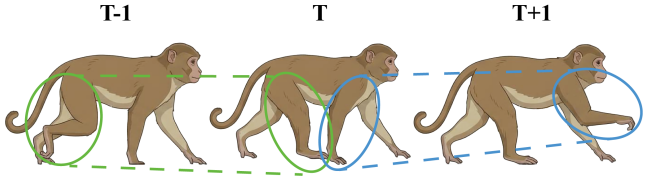

Fig. 1: Example of temporally ordered motion cues in macaque walking.

Critical body parts are frequently occluded by cage structures, while rapid limb movements introduce motion blur, complicating visual analysis. These challenges are further amplified in multi-individual settings, where overlapping interactions introduce additional ambiguity. Under such conditions, visual evidence available in a single frame is often incomplete or ambiguous, necessitating the incorporation of temporal context to resolve occlusions and motion uncertainty.

More fundamentally, many macaque actions cannot be reliably characterized by static poses alone. Instead, they are defined by temporally ordered motion patterns that span multiple frames [12]. As illustrated in Fig. 1, the discriminative cues of a walking action are distributed across successive frames, where coordinated limb movements unfold over time and form structured body-part transitions that reflect the underlying action dynamics. Capturing such temporally ordered motion patterns therefore poses inherent challenges for frame-based approaches, which lack access to the full temporal structure of the action.

To address these challenges, recent pipeline-based approaches have demonstrated encouraging progress for primate behavior analysis [13]. However, multi-stage pipelines and separate models for different tasks not only makes them susceptible to cascading errors but also increases training complexity and deployment overhead. These limitations highlight the need for an end-to-end framework that can jointly estimate macaque location, identity, and actions directly from RGB video in typical cage-housed environments.

In this paper, we introduce the Primate Spatio-Temporal Relational Network (PSTRN), an end-to-end framework for

macaque behavior analysis in standard laboratory cages. Specifically, PSTRN employs a Hierarchical Spatial Encoder to extract multi-scale frame-wise features. These features are then processed by a Cross-frame Relational Attention Module, which aggregates complementary information across frames and hierarchically integrates fine-grained limb movements with global body-level transitions. By explicitly modeling the temporal evolution of body movements as illustrated in Fig. 1, the network is able to model structured motion patterns that are difficult to infer from isolated frames, producing spatio-temporal representations for the joint prediction of bounding boxes, identities, and action categories.

The main contributions of this paper can be summarized as follows:

1) We propose PSTRN, an end-to-end framework that jointly performs macaque behavior analysis from RGB video, avoiding cascading errors and reducing training and deployment complexity inherent.
2) We introduce a cross-frame relational attention mechanism that explicitly models temporal dependencies at each scale, enabling robust capture of structured motion dynamics even under severe occlusion or motion blur.
3) Experiments on single- and multi-macaque benchmarks demonstrate consistent improvements over single-frame and multi-stage pipeline baselines, achieving F1 scores of 85.1% and 82.0%, respectively, while maintaining competitive localization accuracy in challenging cage-housed environments.

Open Science Statement: We will release the source code and model checkpoints upon acceptance via GitHub.

II. Related Work

Animal behavior analysis in videos generally involves two closely related tasks, including action recognition for identifying behavior categories from video sequences and position detection for localizing objects in individual frames. This section is organized around these two perspectives, with a particular focus on complex environments.

A. Animal action recognition in videos

With advances in deep learning and the availability of large-scale video datasets, video-based action recognition has achieved significant progress. DeepVideo [14] was among the first to apply convolutional neural networks (CNNs) to video learning and explored strategies for integrating temporal information across frames. Subsequent works, including Two-Stream Networks [15], 3D CNNs [16], and Temporal Segment Networks [17], further refined spatio-temporal feature modeling and enabled effective action recognition in videos. These architectures have also been extended to animal behavior analysis, yielding promising results on species such as flies [18], mice [19], and shrimp [20].

However, action recognition in macaques is more challenging than in other model organisms [11], [21]. Dense fur obscures visual cues, actions exhibit high variability in three-dimensional space, and frequent interactions with cage structures introduce severe occlusion and motion blur. In such challenging scenarios, robust action recognition requires both accurate spatial localization and temporal reasoning across frames [22]. Most existing action recognition methods assume localized subjects or rely on separate localization components [23], limiting their applicability to realistic cage-housed environments.

B. Animal position detection and frame-based behavior analysis

Animal position detection has been significantly advanced by the introduction of the YOLO architecture [24], which enables efficient end-to-end object localization. Motivated by this paradigm, several studies formulate animal motion analysis as a frame-wise detection task, achieving real-time performance in raw videos. Wang et al. [25] optimized YOLOv5 for monitoring cannibalistic behavior in juvenile groupers, while Li et al. [26] employed an improved SOLOv2 model for dairy cow segmentation to assist lameness detection.

Despite their efficiency, these YOLO-based approaches primarily rely on static appearance cues from individual frames and lack explicit temporal modeling [27]. As a result, they struggle to recognize actions whose semantics emerge from temporally ordered motion patterns, such as walking or climbing, particularly when single-frame observations are ambiguous due to occlusion.

These limitations highlight the need for a unified framework that jointly performs spatial localization and temporally aware action recognition, while preserving the efficiency of frame-wise detection and enabling explicit cross-frame temporal modeling.

III. Primate Spatio-Temporal Relational Network

This paper presents the Primate Spatio-Temporal Relational Network (PSTRN), an end-to-end framework for joint macaque detection, identification, and action recognition from RGB videos in cage-housed environments. PSTRN formulates macaque behavior analysis as a spatio-temporal detection problem, where object localization provides the basis for temporal reasoning, and action semantics are inferred from cross-frame relational interactions rather than isolated frame-wise cues. As illustrated in Fig. 2, PSTRN is composed of four main components, including Temporal Sampling Module, Hierarchical Spatial Encoder, Cross-frame Relational Attention Module, and Multi-task Prediction Head. Each component is described in detail below.

A. Temporal Sampling Module

Temporal Sampling Module constructs a fixed-length video clip centered at a target frame to provide short-term temporal context for subsequent processing. Specifically, given the target frame $I_t \in \mathbb{R}^{H\times W\times 3}$ at time t and a clip length of T frames, the video clip $\mathcal{X}_t \in \mathbb{R}^{T\times H\times W\times 3}$ can be constructed by uniformly sampling frames from the preceding and succeeding time steps:

$$\mathcal{X}_t = \left\{ I_{t-\lfloor T/2 \rfloor}, \ldots, I_t, \ldots, I_{t+\lfloor T/2 \rfloor} \right\} \tag{1}$$

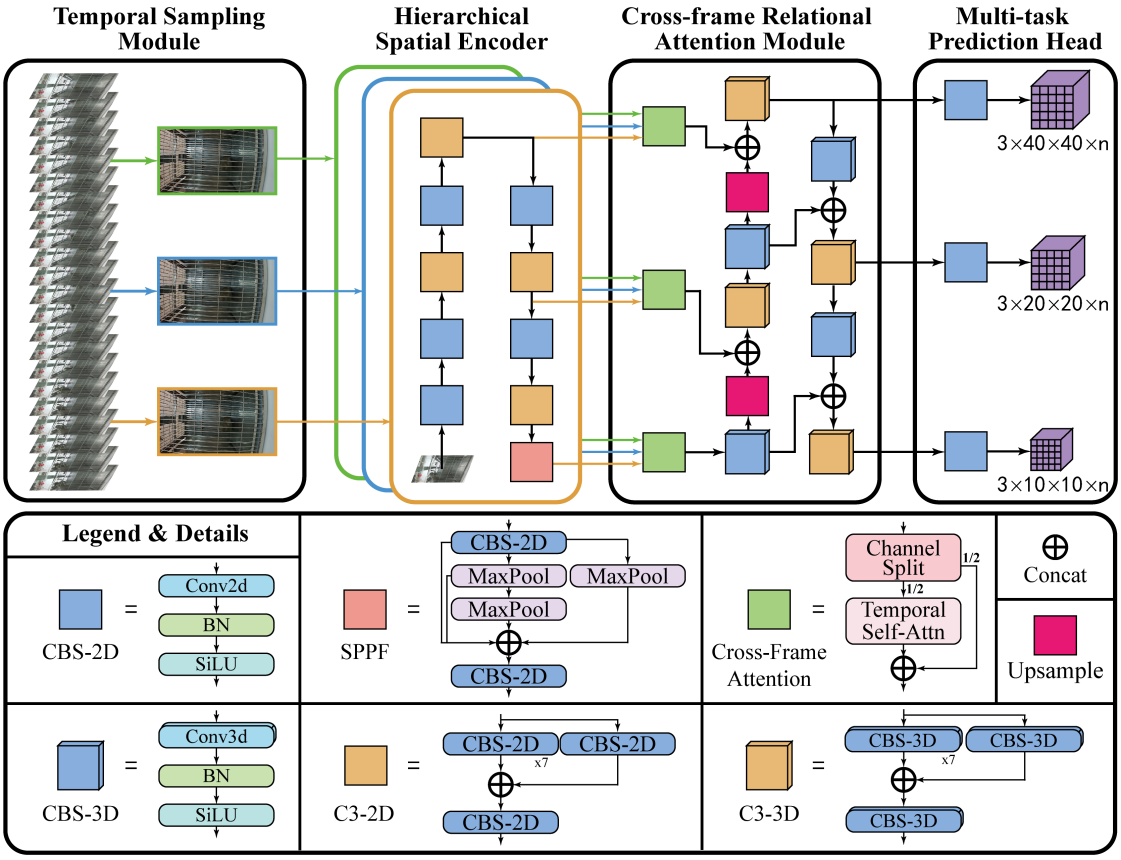

Fig. 2: Overview of proposed PSTRN.

When frame indices fall outside the video boundaries, the closest valid frame is replicated to maintain a consistent clip length.

B. Hierarchical Spatial Encoder

Hierarchical Spatial Encoder extracts multi-scale spatial representations from each frame independently using spatial encoding blocks Φ_{enc} with identical architectures as shown in Fig. 2. The encoder produces multi-scale feature maps $\mathbf{F}_t^{(s)} \in \mathbb{R}^{C_s \times \frac{H}{s} \times \frac{W}{s}}$ for downsampling ratios s:

$$\mathbf{F}_t^{(s)} = \Phi_{\text{enc}}\left(I_t\right), \quad s \in \{8, 16, 32\} \tag{2}$$

These features capture both fine-grained local details and global body-level structures for subsequent temporal modeling.

C. Cross-frame Relational Attention Module

Cross-frame Relational Attention Module models temporal interactions across frames to enhance feature robustness under occlusion and motion uncertainty. Specifically, given temporally aligned feature maps at scale s, the input clip $\mathbf{X}^{(s)} \in \mathbb{R}^{C_s \times T \times \frac{H}{s} \times \frac{W}{s}}$ can be defined as the stacked feature sequence across T frames centered at time t::

$$\mathbf{X}^{(s)} = \left\{ \mathbf{F}_{t+k}^{(s)} \right\}_{k=-\lfloor \frac{T}{2} \rfloor}^{\lfloor \frac{T}{2} \rfloor} \tag{3}$$

The module applies cross-frame attention along the temporal dimension at each spatial location:

$$\text{Attn}\left(\mathbf{X}^{(s)}\right) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d}}\right)\mathbf{V} \tag{4}$$

Where $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ are linear projections of $\mathbf{X}^{(s)}$. This operation adaptively weights information from different time steps, enabling reliable visual cues to be propagated across frames.

The temporally enhanced features are then integrated through a hierarchical fusion process Φ_{3D} based on 3D convolutions, producing feature maps $\mathbf{G}^{(s)}$ at multiple spatial scales:

$$\mathbf{G}^{(s)} = \Phi_{\text{3D}}\left(\text{Attn}\left(\mathbf{X}^{(s)}\right)\right), \quad s \in \{8, 16, 32\} \tag{5}$$

This process produces spatio-temporal representations at multiple spatial resolutions for downstream prediction.

D. Multi-task Prediction Head

Multi-task Prediction Head takes the fused feature maps at three spatial resolutions and performs joint prediction of macaque localization, identity, and action categories. Specifically, for each scale $s \in \{8, 16, 32\}$, the corresponding prediction branch $\mathcal{H}^{(s)}$ generates an output tensor:

$$\mathbf{Y}^{(s)} = \mathcal{H}^{(s)}\left(\mathbf{G}^{(s)}\right) \tag{6}$$

Where $\mathbf{Y}^{(s)}$ includes bounding box confidence, bounding box regression parameters, identity labels, and action classification scores at the corresponding scale. Predictions from all scales are combined to form the final detection results. During inference, non-maximum suppression [28] is applied to remove overlapping predictions and retain the most confident macaque localization, identity, and action outputs.

E. Loss function

PSTRN is trained using a multi-task loss defined as:

$$\mathcal{L} = \lambda_{\text{conf}}\mathcal{L}_{\text{conf}} + \lambda_{\text{box}}\mathcal{L}_{\text{box}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{act}}\mathcal{L}_{\text{act}} \tag{7}$$

Where $\mathcal{L}_{\text{conf}}$, $\mathcal{L}_{\text{box}}$, $\mathcal{L}_{\text{cls}}$, and $\mathcal{L}_{\text{act}}$ correspond to object confidence loss, bounding box regression loss, identity classification loss, and action recognition loss, respectively. The coefficients λ balance the contribution of each task during joint training.

The confidence loss is defined as the binary cross-entropy between predicted objectness scores and ground-truth labels:

$$\begin{aligned}\mathcal{L}_{\text{conf}} = -\frac{1}{N}\sum_{i=1}^{.N}\big[&y^i_{\text{conf}}\log(x^i_{\text{conf}}) \\ &+ (1-y^i_{\text{conf}})\log(1-x^i_{\text{conf}})\big]\end{aligned} \tag{8}$$

Where N denotes the total number of predicted bounding boxes, y^i_{conf} indicates whether the i-th box corresponds to a macaque, and x^i_{conf} is the predicted confidence score.

Bounding box regression adopts an IoU-based loss that jointly considers overlap, center distance, and aspect ratio consistency:

$$\begin{aligned}\mathcal{L}_{\text{box}} = \frac{1}{N}\sum_{i=1}^{N}\big[&1-\text{IoU}+\frac{\rho^2(\mathbf{b}_i,\mathbf{b}^{\text{gt}}_i)}{c^2(\mathbf{b}_i,\mathbf{b}^{\text{gt}}_i)} \\ &+\frac{v_i^2}{(1-\text{IoU})+v_i}\big]\end{aligned} \tag{9}$$

Where IoU (Intersection over Union) [29] represents the intersection over union between the predicted box and the target box, $\mathbf{b}_i$ and $\mathbf{b}^{\text{gt}}_i$ respectively denote the i-th predicted and ground-truth bounding boxes, ρ represents the euclidean distance between box centers, and c is the diagonal length of the smallest enclosing box. The term v_i encodes the consistency of aspect ratios:

$$v_i = \frac{4}{\pi^2}\left(\arctan\left(\frac{w^{\text{gt}}_i}{h^{\text{gt}}_i}\right) - \arctan\left(\frac{w_i}{h_i}\right)\right)^2 \tag{10}$$

Where w and h denote the width and height of the corresponding bounding boxes.

Identity prediction is supervised using a multi-class cross-entropy loss:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y^j_i \log\left(\sigma(x^j_i)\right) \tag{11}$$

Where C is the number of macaque identities, y^j_i is the ground-truth label indicating whether the i-th detection belongs to the j-th individual, x^j_i is the predicted identity score, and σ denotes the sigmoid activation.

Action recognition is trained with a multi-label binary cross-entropy loss:

$$\begin{aligned}\mathcal{L}_{\text{act}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\big[&y^j_i\log\big(\sigma(x^j_i)\big) \\ &+ (1-y^j_i)\log\big(1-\sigma(x^j_i)\big)\big]\end{aligned} \tag{12}$$

Where K is the number of action categories. This formulation allows multiple action labels to be associated with a single individual when necessary.

IV. EXPERIMENTAL RESULTS

A. Datasets

We evaluate PSTRN on two complementary macaque behavior datasets that represent distinct real-world challenges. The Dual-view Macaque Behavior Dataset (DMBD) [30] captures single, cage-housed macaques under frequent structural occlusions from cage bars, while the Spatiotemporal Action Localization of Monkeys in a Group (SALMG) [31] dataset features multiple interacting rhesus monkeys within a single cage, where limited viewpoints, partial inter-individual occlusions, and frequent entry into or exit from the field of view introduce identity ambiguities. These datasets capture complementary single- and multi-macaque conditions with distinct sources of visual ambiguity, enabling a systematic evaluation of PSTRN across varying levels of scene complexity.

B. Implementation Details

Experiments were conducted on a workstation with 24 GB RAM and an NVIDIA GeForce RTX 2070 GPU, using the PyTorch deep learning framework. The datasets were split into training and testing sets following standard protocols. During training, data augmentation strategies including random horizontal flipping and Mosaic augmentation were applied to improve model robustness. The network was optimized using stochastic gradient descent with an initial learning rate of 0.01 and trained for 40 epochs.

C. Ablation Study of PSTRN

TABLE I and TABLE II summarize the ablation results on the DMBD and SALMG datasets. We evaluate a series of model variants to assess the effects of key design choices on localization and action recognition performance. Attention scales specify the number of feature scales at which cross-frame attention is applied, where 1 denotes the last scale only

TABLE I: Ablation results on DMBD

Methods	Attention Scales	Half Input	Positional Encoding	mAP_{50} (%)	Acc (%)	F1 (%)
CFA-no	✗	✗	✗	99.5	88.6	67.2
CFA-1L	1	✗	✗	99.4	95.5	77.6
CFA-3L	3	✗	✗	99.5	95.9	78.1
CFA-1L-h	1	✓	✗	99.4	96.6	82.0
PSTRN	3	✓	✗	99.5	**97.5**	**85.1**
CFA-1L-Pos	1	✗	✓	99.5	95.6	78.3
CFA-3L-Pos	3	✗	✓	99.4	95.6	77.7
CFA-1L-Pos-h	1	✓	✓	99.5	96.5	82.0
CFA-3L-Pos-h	3	✓	✓	99.5	95.0	76.2

TABLE II: Ablation results of SALMG

Methods	Attention Scales	Half Input	Positional Encoding	mAP_{50} (%)	Acc (%)	F1 (%)
CFA-no	✗	✗	✗	**93.4**	73.9	62.6
CFA-1L	1	✗	✗	91.0	85.0	72.6
CFA-3L	3	✗	✗	92.2	80.6	73.7
CFA-1L-h	1	✓	✗	89.0	81.0	73.6
PSTRN	3	✓	✗	92.6	**90.3**	**82.0**
CFA-1L-Pos	1	✗	✓	89.5	80.0	71.1
CFA-3L-Pos	3	✗	✓	89.9	82.6	74.1
CFA-1L-Pos-h	1	✓	✓	90.0	80.7	75.7
CFA-3L-Pos-h	3	✓	✓	85.6	86.3	77.6

TABLE III: Performance comparison on DMBD

Methods	mAP_{50} (%)	Acc (%)	F1 (%)
SI-2D	99.4	89.6	78.7
Faster R-CNN+C3D	97.0	92.1	73.0
Faster R-CNN+SlowFast	97.0	92.4	71.1
Faster R-CNN+ViVit	97.0	89.5	71.8
SIPEC	98.1	90.1	69.7
PSTRN	**99.5**	**97.5**	**85.1**

Baseline results are taken from [30].

TABLE IV: Performance comparison on SALMG

Methods	mAP_{50} (%)	Acc (%)	F1 (%)
Single frame	90.6	88.6	78.8
YOLOv5l+I3D	92.0	82.1	63.7
YOLOv5l+SlowFast	92.0	83.4	66.1
JDC-MF	92.0	90.2	81.4
PSTRN	**92.6**	**90.3**	**82.0**

Baseline results are taken from [31].

and 3 denotes all scales. Half input reduces the channel dimensionality of features before attention-based fusion. Positional encoding denotes the optional addition of spatial information to guide the attention. Performance is measured by mAP_{50} (mean Average Precision at 50% IoU) [32] for localization and F1 score with accuracy (ACC) for action recognition.

The results show that localization performance remains highly consistent across all variants on the DMBD dataset, reflecting the relatively simple single-macaque setting. In the more challenging SALMG dataset, mAP_{50} exhibits greater variability. The baseline CFA-no achieves the highest mAP_{50} at 93.4%, and introducing attention-based mechanisms generally leads to a slight decrease in localization accuracy, possibly due to increased model complexity. Nevertheless, PSTRN attains 92.6% mAP_{50}, remaining close to the top performance.

Despite the limited impact on localization, temporal modeling plays a critical role in action understanding. The performance on DMBD shows that CFA-1L increases the F1 score from 67.2% to 77.6% relative to CFA-no. Results on SALMG further demonstrate its effectiveness, with CFA-1L raising F1 from 62.6% to 72.6%, highlighting the importance of explicit temporal relational modeling for handling occlusions and motion variations.

Applying cross-frame attention on all three feature scales, rather than only on the final scale, yields modest gains on DMBD but more pronounced improvements on SALMG, indicating that multi-scale temporal modeling is particularly beneficial under increased visual complexity.

Introducing half-resolution input improves recognition performance on both datasets. In DMBD, CFA-1L-h increases F1 from 77.6% to 82.0%, and PSTRN further reaches the highest 85.1%. In SALMG, CFA-1L-h provides a modest gain to 73.6%, while PSTRN achieves 81.9%, indicating that applying attention on a reduced channel representation improves temporal aggregation efficiency, while the residual full-resolution features help preserve spatial detail for action recognition.

Positional encoding does not consistently improve recognition and can reduce performance, which may be because absolute positional priors are less compatible with the dynamic and non-rigid spatial configurations of macaque behavior, especially during social interactions.

The results indicate that cross-frame attention, multi-scale modeling, and half-resolution input each contribute to improving behavior recognition, and their combination in PSTRN

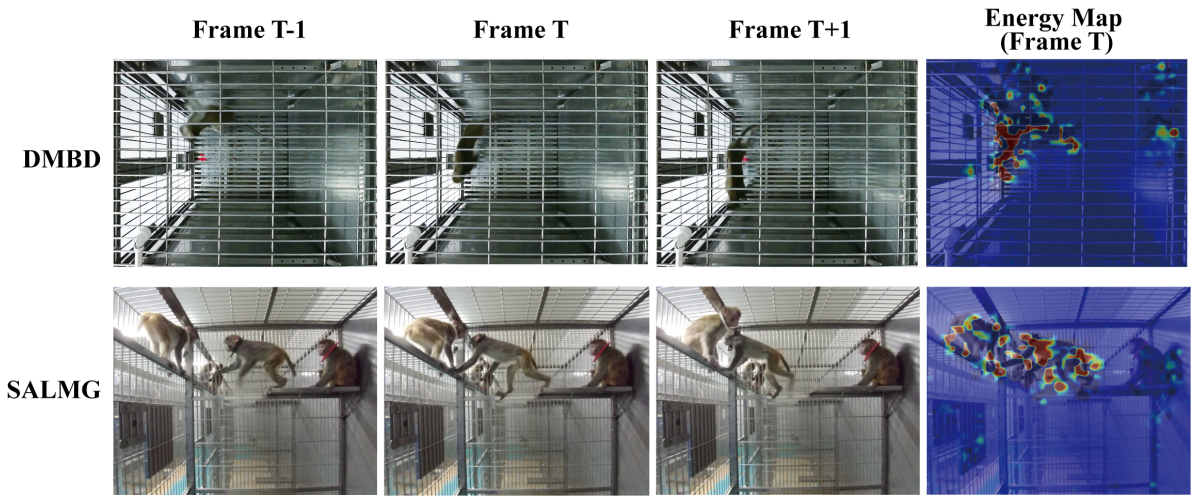

Fig. 3: Visualization of cross-frame attention energy maps.

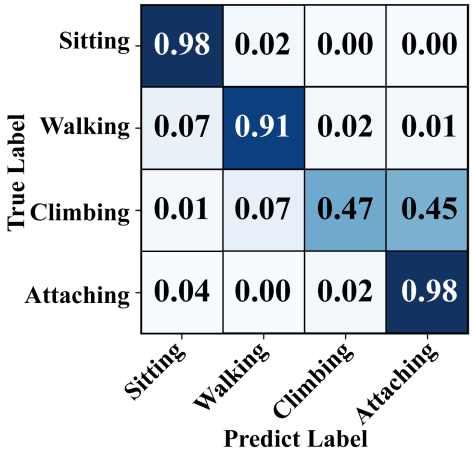

Fig. 4: Class-level Action Recognition Confusion Matrices on DMBD.

produces the best overall performance without compromising detection accuracy.

D. Comparison with existing approaches on DMBD and SALMG

TABLE III and TABLE IV compare PSTRN with representative single-frame methods, pipeline-based approaches, and multi-frame approaches on the DMBD and SALMG datasets.

On DMBD, PSTRN consistently outperforms all competing methods across localization and action recognition metrics. Compared with the single-frame method SI-2D, PSTRN increases F1 from 78.7% to 85.1% and accuracy from 89.6% to 97.5%, demonstrating that joint spatio-temporal modeling provides substantial benefits even in relatively simple single-macaque scenarios. Pipeline-based approaches that decouple detection and action recognition achieve lower recognition performance, indicating limited temporal modeling capability.

Performance differences are more pronounced on the multi-macaque SALMG dataset. Single-frame methods and pipeline-based approaches methods show reduced F1 due to partial occlusion and inter-individual interactions. It is worth noting that JDC-MF also adopts multi-frame input but relies on convolution-based temporal aggregation without cross-frame attention or 3D convolutional feature fusion. In contrast, PSTRN achieves the highest overall F1 on SALMG, improving from 81.4% for JDC-MF to 82.0%, highlighting its effectiveness in handling partially occluded, interacting individuals.

Additionally, PSTRN achieves real-time inference at 35 FPS on the NVIDIA RTX 2070 platform, demonstrating its suitability for continuous behavioral monitoring in standard laboratory environments.

E. Qualitative Analysis of Context-Aware Cross-frame Attention

To illustrate how Cross-frame Attention aggregates motion cues across time, Fig. 3 visualizes three consecutive frames together with the energy map of the middle frame. The first row presents samples from the DMBD single-macaque dataset, while the second row shows examples from the SALMG multi-macaque dataset.

Under top-down viewpoints with cage-bar occlusion, limb motion in the DMBD examples is often ambiguous or partially invisible, causing the energy map to concentrate on the torso as the most temporally stable region across frames. In contrast, the SALMG examples depict a multi-macaque scenario in which two monkeys exhibit clear motion while another remains stationary. In this case, the energy map emphasizes the heads and limbs of the moving monkeys while suppressing stationary individuals and rigid torso regions. This contrast indicates that Cross-frame Attention adapts its focus according to temporal reliability, suppressing static or ambiguous regions while emphasizing motion patterns that remain consistent across frames.

F. Class-level Action Recognition Analysis via Confusion Matrices

To further analyze the class-level recognition performance of the proposed model, we present confusion matrices on

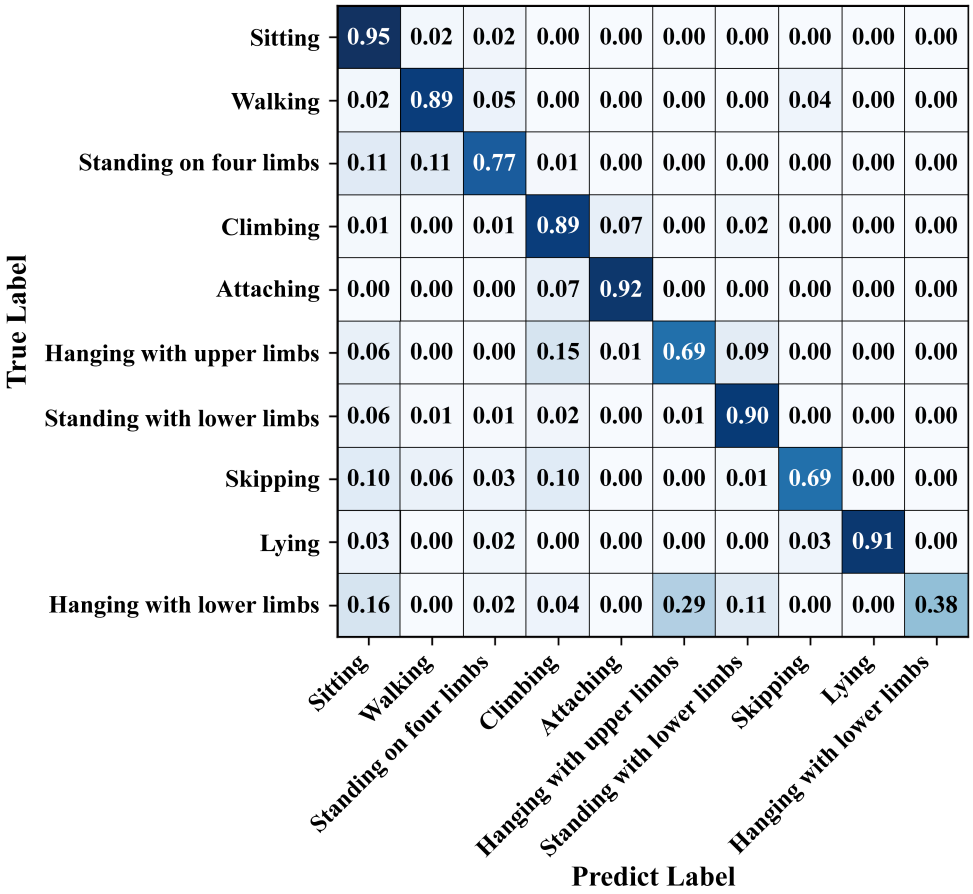

Fig. 5: Class-level Action Recognition Confusion Matrices on SALMG.

DMBD and SALMG, highlighting common misclassification patterns across action categories under different visual and contextual conditions.

Fig. 4 presents the confusion matrix on the DMBD. The results show a clear and well-structured classification pattern across the four action categories.Sitting, Walking, and Attaching are recognized with high accuracy, indicating that the model effectively captures distinct motion patterns and spatial context cues even under partial occlusion from cage structures. The primary source of confusion arises between Climbing and Attaching, which constitute a pair of fine-grained and semantically adjacent actions. While Attaching is recognized reliably, Climbing is more frequently misclassified as Attaching, which may be attributed to their subtle temporal differences and the long-tailed distribution of Climbing in the dataset [30].

Fig. 5 shows the confusion matrix on the more challenging SALMG. Actions characterized by stable body support and consistent spatial configurations across frames, such as Sitting, Lying, and Standing on four limbs, maintain high recognition accuracy. In contrast, moderate confusion occurs between Climbing and Hanging with upper limbs, as both involve similar upper-body contact with the cage and exhibit minimal visual distinction across frames. Despite the use of cross-frame attention, their shared static configuration and partial mutual occlusion lead to residual ambiguity. Nevertheless, the dominant diagonal structure and overall prediction distribution indicate that the model preserves robust action discrimination even in crowded and partially occluded multi-animal scenarios.

V. Conclusion

This paper proposes the Primate Spatio-Temporal Relational Network (PSTRN) for unified macaque behavior analysis in cage environments. PSTRN combines hierarchical spatial encoding with cross-frame relational attention to capture temporally ordered motion dynamics across multiple scales, enabling joint prediction of location, identity, and actions in a single network. Experiments on single- and multi-macaque benchmarks demonstrate that PSTRN achieves state-of-the-art action recognition F1 scores while maintaining competitive localization accuracy and real-time inference. Qualitative and class-level analyses show that the model effectively leverages temporal context to resolve ambiguities under occlusion, though distinguishing visually similar static postures remains challenging. These results highlight PSTRN as an effective and practical tool for automated, comprehensive behavioral monitoring in preclinical neuroscience and pharmacology studies.

Acknowledgment

This work was supported by the National Key Research and Development Program of China (grant number 2023YFF0720000 and 2023YFF0720002); the Strate-

gic Priority Research Program of the Chinese Academy of Sciences (grant number XDA0450000, XDA0450200 and XDA0450202); and the Youth Innovation Promotion Association CAS (grant number Y201930).

REFERENCES

[1] R. A. Gibbs, J. Rogers, M. G. Katze, R. Bumgarner, G. M. Weinstock, E. R. Mardis, K. A. Remington, R. L. Strausberg, J. C. Venter, and R. K. Wilson, "Evolutionary and biomedical insights from the rhesus macaque genome," *science*, vol. 316, no. 5822, pp. 222–234, 2007.

[2] Y. Yao, P. Bala, A. Mohan, E. Bliss-Moreau, K. Coleman, S. M. Freeman, C. J. Machado, J. Raper, J. Zimmermann, B. Y. Hayden *et al.*, "Openmonkeychallenge: dataset and benchmark challenges for pose estimation of non-human primates," *International journal of computer vision*, vol. 131, no. 1, pp. 243–258, 2023.

[3] S. Sun, J. Li, S. Wang, J. Li, J. Ren, Z. Bao, L. Sun, X. Ma, F. Zheng, and S. Ma, "CHIT1-positive microglia drive motor neuron ageing in the primate spinal cord," *Nature*, vol. 624, no. 7992, pp. 611–620, 2023.

[4] T. Kiss, W. E. Hoffmann, L. Scott, T. T. Kawabe, A. J. Milici, E. A. Nilsen, and M. Hajós, "Role of thalamic projection in NMDA receptor-induced disruption of cortical slow oscillation and short-term plasticity," *Frontiers in psychiatry*, vol. 2, p. 14, 2011.

[5] N. Vogt, "Correlating behavior and neural activity at high resolution," *Nature methods*, vol. 15, no. 7, pp. 479–479, 2018.

[6] S. Schneider, J. H. Lee, and M. W. Mathis, "Learnable latent embeddings for joint behavioural and neural analysis," *Nature*, vol. 617, no. 7960, pp. 360–368, 2023.

[7] 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*, vol. 16, no. 1, pp. 117–125, 2019.

[8] B. R. Eisenreich, B. Y. Hayden, and J. Zimmermann, "Macaques are risk-averse in a freely moving foraging task," *Scientific reports*, vol. 9, no. 1, p. 15091, 2019.

[9] M.-S. Liu, J.-Q. Gao, G.-Y. Hu, G.-F. Hao, T.-Z. Jiang, C. Zhang, and S. Yu, "MonkeyTrail: A scalable video-based method for tracking macaque movement trajectory in daily living cages," *Zoological Research*, vol. 43, no. 3, p. 343, 2022.

[10] D. J. Anderson and P. Perona, "Toward a science of computational ethology," *Neuron*, vol. 84, no. 1, pp. 18–31, 2014.

[11] P. C. Bala, B. R. Eisenreich, S. B. M. Yoo, B. Y. Hayden, H. S. Park, and J. Zimmermann, "Automated markerless pose estimation in freely moving macaques with OpenMonkeyStudio," *Nature communications*, vol. 11, no. 1, p. 4560, 2020.

[12] Y. Tang, Y. Zheng, C. Wei, K. Guo, H. Hu, and J. Liang, "Video representation learning for temporal action detection using global-local attention," *Pattern Recognition*, vol. 134, p. 109135, 2023.

[13] M. Marks, Q. Jin, O. Sturman, L. von Ziegler, S. Kollmorgen, W. von der Behrens, V. Mante, J. Bohacek, and M. F. Yanik, "Deep-learning-based identification, tracking, pose estimation and behaviour classification of interacting primates and mice in complex environments," *Nature machine intelligence*, vol. 4, no. 4, pp. 331–340, 2022.

[14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[15] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[16] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[17] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 20–36.

[18] A. Mathis, S. Schneider, J. Lauer, and M. W. Mathis, "A primer on motion capture with deep learning: Principles, pitfalls, and perspectives," *Neuron*, vol. 108, no. 1, pp. 44–65, 2020.

[19] T. D. Pereira, J. W. Shaevitz, and M. Murthy, "Quantifying behavior to understand the brain," *Nature neuroscience*, vol. 23, no. 12, pp. 1537–1549, 2020.

[20] I. Martinez-Alpiste, J.-B. de Tailly, J. M. Alcaraz-Calero, K. A. Sloman, M. E. Alexander, and Q. Wang, "Machine learning-based understanding of aquatic animal behaviour in high-turbidity waters," *Expert Systems with Applications*, vol. 255, p. 124804, 2024.

[21] C. Testard, S. Tremblay, and M. Platt, "From the field to the lab and back: Neuroethology of primate social behavior," *Current opinion in neurobiology*, vol. 68, pp. 76–83, 2021.

[22] J. Feng, H. Luo, and D. Fang, "A progressive deep learning framework for fine-grained primate behavior recognition," *Applied Animal Behaviour Science*, vol. 269, p. 106099, 2023.

[23] Z. Zhu, L. Wang, W. Tang, Z. Liu, N. Zheng, and G. Hua, "Learning disentangled classification and localization representations for temporal action localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 3644–3652.

[24] J. Redmon, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[25] H. Wang, S. Zhang, S. Zhao, J. Lu, Y. Wang, D. Li, and R. Zhao, "Fast detection of cannibalism behavior of juvenile fish based on deep learning," *Computers and Electronics in Agriculture*, vol. 198, p. 107033, 2022.

[26] Q. Li, Z. He, X. Liu, M. Chu, Y. Wang, X. Kang, and G. Liu, "Lameness detection system for dairy cows based on instance segmentation," *Expert Systems with Applications*, vol. 249, p. 123775, 2024.

[27] J. Wu, Y. Li, L. Wang, K. Wang, R. Li, and T. Zhou, "Skeleton based temporal action detection with yolo," in *Journal of Physics: Conference Series*, vol. 1237, no. 2. IOP Publishing, 2019, p. 022087.

[28] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 3. IEEE, 2006, pp. 850–855.

[29] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[30] Z. Chen, Z. Jiang, Q. Guan, and X. Ma, "End-to-end motion detection via multi-scale spatial-temporal feature fusion for dual-view 3d macaque behavior quantification," *Expert Systems with Applications*, p. 129765, 2025.

[31] K. Liang, Z. Chen, S. Yang, Y. Yang, C. Qin, and X. Ma, "The joint detection and classification model for spatiotemporal action localization of primates in a group," *Neural Computing and Applications*, vol. 35, no. 25, pp. 18 471–18 486, 2023.

[32] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–338, 2010.