

Structured Context Enhancement Network for Mouse Pose Estimation

Feixiang Zhou, Zheheng Jiang, Zhihua Liu, Fang Chen, Long Chen, Lei Tong, Zhile Yang, Haikuan Wang, Minrui Fei, Ling Li and Huiyu Zhou

Abstract—Automated analysis of mouse behaviours is crucial for many applications in neuroscience. However, quantifying mouse behaviours from videos or images remains a challenging problem, where pose estimation plays an important role in describing mouse behaviours. Although deep learning based methods have made promising advances in human pose estimation, they cannot be directly applied to pose estimation of mice due to different physiological natures. Particularly, since mouse body is highly deformable, it is a challenge to accurately locate different keypoints on the mouse body. In this paper, we propose a novel Hourglass network based model, namely Graphical Model based Structured Context Enhancement Network (GM-SCENet) where two effective modules, i.e., Structured Context Mixer (SCM) and Cascaded Multi-level Supervision (CMLS) are subsequently implemented. SCM can adaptively learn and enhance the proposed structured context information of each mouse part by a novel graphical model that takes into account the motion difference between body parts. Then, the CMLS module is designed to jointly train the proposed SCM and the Hourglass network by generating multi-level information, increasing the robustness of the whole network. Using the multi-level prediction information from SCM and CMLS, we develop an inference method to ensure the accuracy of the localisation results. Finally, we evaluate our proposed approach against several baselines on our Parkinson's Disease Mouse Behaviour (PDMB) and the standard DeepLabCut Mouse Pose datasets. The experimental results show that our method achieves better or competitive performance against the other state-of-the-art approaches.

Index Terms—Pose estimation, graphical model, multi-level information, Parkinson's Disease, Mouse Behaviour dataset.

I. INTRODUCTION

MOUSE models are an important tool to study neurodegenerative disorders such as Alzheimer [1] and Parkinson's disease [2] in neuroscience. Modelling mouse behaviour is key to understand brain functions and disease mechanisms [3], due to certain similarity and homology between humans and animals. Historically, studying mouse behaviour can be a time-consuming and difficult task because the collected data requires experts to physically engage. Recent advances in computer vision and machine learning have facilitated automated analysis for complex behaviours [4]–[6]. This makes it

possible to conduct a wide range of behavioural experiments without human intervention, which has yielded new insights into the pathologies and treatment of specific diseases [7]–[9].

In this paper, we are particularly interested in 2D mouse pose estimation, which is one of the fundamental problems in mouse behaviour analysis. Mouse pose estimation is defined as the problem of measuring mouse posture which denotes the geometrical configuration of body parts in images or videos [10]. This provides useful information for ethologically relevant behaviours where the keypoint localisation of individual body parts is significant [10]. Upon its success, pose analysis can be used to reveal mouse motor functions by providing configuring features [11], especially for studying social behaviour in experimental mice. Specifically, the spatio-temporal motion representations derived from the keypoints of each mouse could be encoded to interpret social behaviours. Mouse pose estimation is non-trivial at all due to small size, occlusions, invisible keypoints, and deformation of the body. Early approaches focused on labelling the locations of interest with unique and recognizable markers in order to simplify the pose analysis [12]–[14]. However, these systems are 'invasive' to the targeted subjects. To overcome this problem, several advanced markerless methods have been proposed for mouse pose estimation [15]–[17]. Overall, a common technique used in these methods is to fit specific skeleton or active models, such as bounding box [15], ellipse [16] and contour [18]. These methods have produced promising results, but the flexibility of such methods is limited as they require sophisticated skeleton models.

With the development of Deep Convolutional Neural Networks (DCNNs), significant progresses have been achieved in human and animal pose estimation [19]–[22]. Most of the existing systems (e.g., DeepLabCut [17], LEAP [23], DeepPoseKit [22]) are built on deep networks for human pose estimation. For example, DeepLabCut combines the feature detectors of DeeperCut [24] with readout layers, namely deconvolutional layers, for markerless pose estimation, and exploits transfer learning to train the deep model with fewer examples. Similarly, LEAP has been also developed based on an existing network, namely SegNet [25], which provides a graphical interface for labelling body parts. However, its preprocessing is computationally expensive, thus limiting the application of their system in other environments. DeepPoseKit uses two Hourglass modules with different scales to learn multi-scale geometry of the keypoints, based on the standard Hourglass network [26].

Although the system performance on different animal

F. Zhou, Z. Liu, F. Chen, L. Chen, L. Tong and H. Zhou are with School of Informatics, University of Leicester, United Kingdom. H. Zhou is the corresponding author. E-mail: hz143@leicester.ac.uk.

Z. Jiang is with School of Computing and Communications, Lancaster University, United Kingdom.

Z. Yang is with Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China.

H. Wang and M. Fei are with School of Mechanical Engineering and Automation, Shanghai University, Shanghai, China.

L. Li is with School of Computing, University of Kent, United Kingdom.

datasets reported in these works is encouraging, the set-up of their environments is simple with clean background. These established systems mainly deployed the standard human pose estimation networks for animal applications. A common pipeline of the existing deep models is to generate heatmaps (also known as score maps) of all the generated keypoints simultaneously at the final stage of the networks [17], [22], [23], [26]–[28]. However, these systems have not fully considered the differences between different body parts, which is critical to spatio-temporal analysis. This gap may be due to the fact that spatial correlations between mouse body parts are much weaker than those of human body parts due to highly deformable body structures.

We here propose a Graphical Model based Structured Context Enhancement Network (GM-SCENet) for mouse pose estimation, as shown in Fig. 1. We use the standard Hourglass network [26] as the basic structure due to its strong multi-scale feature representation capability (Fig. 1(a)). In our approach, the proposed Structured Context Mixer (SCM), shown in Fig. 1(b), first captures global spatial configurations of the full mouse body to describe the global contextual information by concatenating multi-level feature maps from the Hourglass network. In combination with the global contextual information, our proposed SCM explores the neighbouring areas at four different directions of each keypoint collected from mouse body parts. Afterwards, we model the relationship between the keypoint-specific and the global contextual information by a novel graphical model. We define these two types of contextual information as structured contextual information hereafter.

In addition, we introduce a novel Cascaded Multi-level Supervision module (CMLS) [26], [27] to jointly train the proposed SCM and the Hourglass network, as shown in Fig. 1(c). This allows our network to generate multi-level information for addressing the problems of scale variations and small subjects by strengthening contextual learning. In fact, the prediction results of the current scale/level, as prior knowledge, are used for semantic description with the supervision of high-scale feature maps. In the meantime, the extracted features and the predicted heatmaps within the CMLS module can be further refined using a cascaded structure. Finally, to integrate multi-level localisation results, we also design an inference algorithm called Multi-level Keypoint Aggregation (MLKA) to produce the final prediction results, where the locations of the selected multi-level keypoints from all the supervision layers are aggregated.

Fig. 1 shows our proposed framework for mouse pose estimation. The main contributions of our work can be summarised as follows:

- We propose a novel Graphical Model based Structured Context Enhancement Network (GM-SCENet) which consists of Structured Context Mixer (SCM), a Cascaded Multi-level Supervision module (CMLS) and the backbone Hourglass network. The proposed GM-SCENet has the ability of modelling the differences and spatial correlations between mouse body parts simultaneously.
- The proposed SCM can adaptively learn and enhance the structured contextual information of each keypoint. This is achieved by exploring the global and keypoint-

specific contextual information to describe the differences between body parts, and designing a novel graphical model to explicitly model the relationship between them.

- We also develop a novel CMLS module to jointly train SCM and the Hourglass network, which adopts multi-scale features for supervised learning and increases the robustness of the whole network. In addition, we design a Multi-Level Keypoint Aggregation (MLKA) algorithm to integrate multi-level localisation results.
- We introduce a new challenging mouse dataset for pose estimation and behaviour recognition, namely Parkinson's Disease Mouse Behaviour (PDMB). Several common behaviours and the locations of four representative body parts are included in this dataset. To the best of our knowledge, this is the first large dataset for mouse pose estimation, collected from standard CCTV cameras.

II. RELATED WORK

A. Human Pose Estimation

Many traditional methods for single-person pose estimation adopt graph structures, e.g., pictorial structures [29] or graphical models [30] to explicitly model the spatial relationships between body parts. However, the prediction results of keypoints rely heavily on hand-crafted features, which are susceptible to difficult problems such as occlusion. Recently, significant improvements have been achieved by designing DCNNs for learning high-level representations [28], [31]–[37]. DeepPose [38] is the first attempt to use deep models for pose estimation, which directly regress the keypoints coordinates using an end-to-end network. However, generating predictions in this way misses the spatial relationship in the images. Recent methods mainly adopt fully convolutional neural networks (FCNNs) to regress 2D Gaussian heatmaps [26], [27], followed by further inference using Gaussian. For instance, Newell et al. [26] first proposed the Hourglass network to gradually capture representations across scales by repeating bottom-up and top-down processes. They also adopted intermediate supervision to train the entire network, and learnt the spatial relationship with a larger receptive field. In [28], Sun et al. designed the High-Resolution Net (HRNet) that pursues high-resolution representations, fusing multi-scale feature maps of parallel multi-resolution subnetworks. In spite of their promising performance, it remains an open problem to locate keypoints accurately due to abnormal postures, small objects and occluded body parts.

B. Mouse Pose Estimation

In previous works, fitting active shape models was a popular way to approach the understanding of mouse postures. For example, De Chaumont et al. [16] employed a set of geometrical primitives for mouse modelling. Mouse posture is represented by three parts such as head, trunk and neck with corresponding spatial distances between them. Similarly, in [18], 12 B-spline deformable 2D templates were adopted to represent a mouse contour which is mathematically described with 8 elliptical parameters. However, these methods require sophisticated skeleton models, thus limiting the ability to fit to

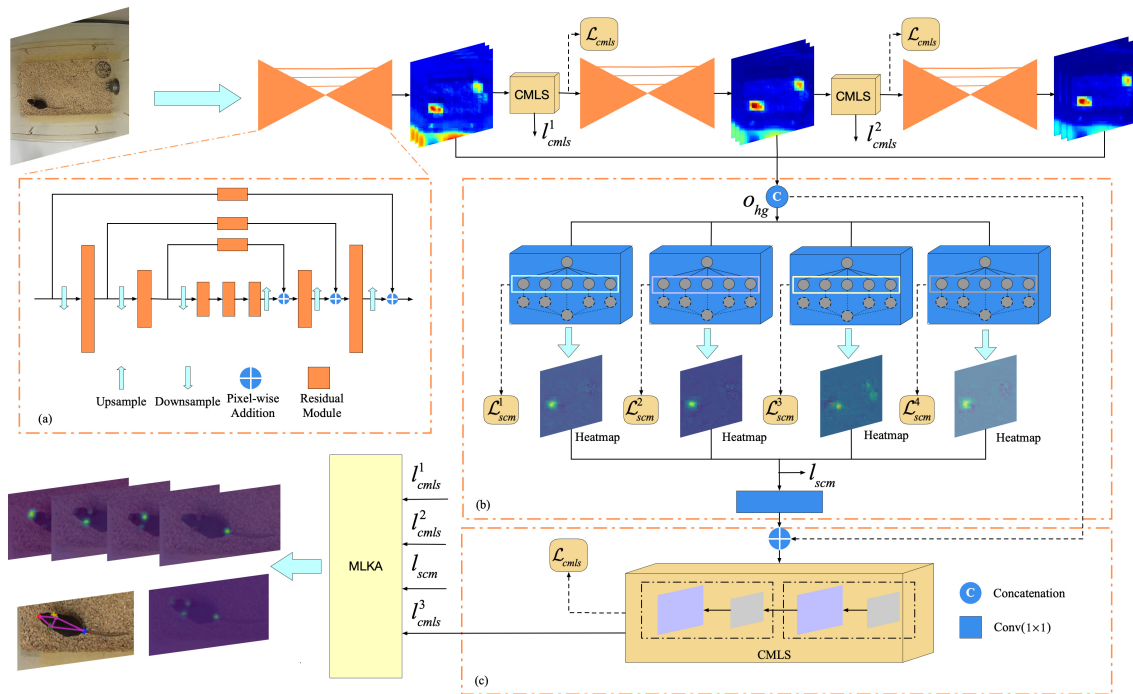


Fig. 1. Overview of the proposed framework for mouse pose estimation. (a) The structure of the Hourglass network [26]. We stack 3 Hourglass modules to capture multi-level pose features. (b) The structure of the proposed Structured Context Mixer (SCM). It first generates global contextual information by fusing multi-level features from (a). To describe the differences between different body parts, we then design 4 independent subnetworks to predict the heatmap of the corresponding keypoint, where each subnetwork explores global and keypoint-specific contextual information to represent structured contextual information of each keypoint (see Section III-A1). The blue blocks refer to the 4 subnetworks, in each of which we propose a novel graphical model to adaptively fuse the global and keypoint-specific contextual information (see Section III-A2). (c) The structure of Cascaded Multi-level Supervision (CMLS). We use this module as an intermediate supervision agent with the multi-scale outputs for joint training of SCM, the CMLS module and the Hourglass network (see Section III-B). We attach this module to the end of each Hourglass module and the proposed SCM. Moreover, during the inference phase, the prediction results (i.e. locations of keypoints) l_{cmls}^1, l_{cmls}^2 and l_{cmls}^3 generated in the CMLS modules and l_{scm} generated in SCM are used as input of our proposed Multi-level Keypoint Aggregation (MLKA) inference method to produce the final localisation results (see Section III-C). All these components are trained in an end-to-end fashion. \mathcal{L}_{cmls} in (c) represents the loss of all the keypoints. $\mathcal{L}_{scm}^1, \mathcal{L}_{scm}^2, \mathcal{L}_{scm}^3$ and \mathcal{L}_{scm}^4 in (b) represent the losses of keypoints 1-4, respectively.

different scenes. Recently, most deep learning based methods for 2D mouse or other animal pose estimation are based on human pose estimation models such as DeeperCut [24], SegNet [25], and Hourglass network [26]. In [22], Graving et al. presented a multi-scale deep learning model called Stacked DenseNet to generate confidence maps from keypoints, based on the standard Hourglass architecture. This network is able to capture multi-scale information with intermediate supervision capabilities. Nevertheless, the ability of these approaches to deal with the problems such as occlusion and small target size is still weak.

III. PROPOSED METHODS

We formulate the task of mouse pose estimation in images as the problem of learning a non-linear mapping $\mathcal{F}: \mathcal{I} \rightarrow \mathcal{G}$, where \mathcal{I} and \mathcal{G} refer to image and keypoint spaces. More formally, let $T = \{(I_i, \tilde{G}_i)\}_{i=1}^M$ be the training set of M mouse images, where $I_i \in \mathcal{I}$ denotes the input image and $\tilde{G}_i \in \mathcal{G}$ represents the corresponding ground-truth labels. \tilde{G}_i can be represented as $\tilde{G}_i = \{G_i^1, G_i^2, \dots, G_i^K\} \in \mathbb{R}^{K \times 2}$, where K is the number of the keypoints defined in the image space \mathcal{I} .

To learn \mathcal{F} , we propose a framework composed of three main parts: Structured Context Mixer, Cascaded Multi-Level Supervision module and the backbone network. The proposed

SCM aims to model the dependency between keypoint-specific and global contextual information, which focuses on both spatial correlations and differences between keypoints. In this paper, we denote $O = \{O_c\}, c \in C = \{c_g, c_t, c_l, c_b, c_r\}$ as the structured contextual feature maps of each keypoint, where c_g represents the global contextual information, c_t, c_l, c_b and c_r are keypoint-specific contextual information. Here, O_c can be represented as a set of feature vectors $O_c = \{o_c^i\}_{i=1}^N$, $o_c^i \in \mathbb{R}^L$, where N is the number of the image pixels in a contextual feature map and L is the number of the channels of the contextual feature maps. Unlike previous systems that model contextual information using direct concatenation [39], channel attention [40] and spatial attention [34], we here propose to adaptively fuse the structured contextual information by learning a set of latent feature maps $H_c = \{h_c^i\}_{i=1}^N, c \in C$ with a novel graphical model. In particular, we implement two interactive attention gates $A_c = \{a_c^i\}_{i=1}^N, a_c^i \in [0, 1]$ and $G_c = \{g_c^i\}_{i=1}^N, g_c^i \in [0, 1]$ to control the flow of the information between keypoint-specific and global contextual information. With the two-gate mechanism, our network can adaptively learn and enhance the relatively important keypoint-specific contextual features by preventing the potential loss of low- and high-level information. In SCM, the proposed graphical model jointly infers the hidden contextual features

$H = \{H_c\}, c \in C$ and the attention gates $A = \{A_c\}, G = \{G_c\}, c \in C_k = \{c_t, c_l, c_b, c_r\}$.

A. Structured Context Mixer

1) *Structured Context Information Representations*: Extracting keypoints of mouse body parts is non-trivial at all due to occlusion, small size and highly deformable body. In order to accurately locate different keypoints of the targets, most deep models take multi-scale or high-resolution information into account [22], [26], [28], whilst utilising contextual information. In general, contextual information is referred to as regions surrounding the targets, and it has been proved effective in pose estimation [26], [34] and object detection [41]. However, these models miss the differences between different keypoints to some extent, which may significantly contribute towards mouse pose estimation due to spatial correlations and restrictions. Unlike previous methods, we explore the structured contextual information of body parts to describe the differences between them, including global and keypoint-specific contextual information.

Before generating structured contextual information of each keypoint, we aggregate the low- and high-level feature maps from multiple stacks of the Hourglass network for representing the global contextual information O_{hg} of the whole subject, as shown in Fig. 1. In fact, the aggregated features also encode the local contextual information derived from the low-level layers. O_{hg} is defined as:

$$O_{hg} = \text{Concat}(o_{s1}, o_{s2}, \dots, o_{sn}) \quad (1)$$

where *Concat* represents channel-wise concatenation, o_{s1} , o_{s2} , and o_{sn} are the outputs of the first, second and n -th stack of the Hourglass network, respectively.

Afterwards, we extract keypoint-specific contextual information, where multiple discriminative regions are selected to represent the contextual information of each keypoint. This is beneficial to the accurate localisation of the keypoints. Inspired by the corner pooling scheme used in [42], we explore four contextual features to describe the keypoint-specific contextual information by the Directionmax operation, as shown in Fig. 2, using prior knowledge.

In [42], to locate the top-left corner of a bounding box, we look horizontally towards the right for the topmost boundary of the target and vertically towards the bottom for the leftmost boundary. Similarly, we look horizontally towards the left for the bottommost boundary of the target and vertically towards the top for the rightmost boundary to pursue the bottom-right corner. Different from their work, we focus on locating different mouse keypoints in our paper. The problem of locating the top-left and bottom-right corners can be considered as that of locating a point when these two corners spatially overlap. Therefore, following the top-left and bottom-right pooling used in [42], we adopt top-left pooling (Topmax and Leftmax) and bottom-right pooling (Bottommax and Rightmax) simultaneously to generate four regions to indicate keypoint-specific contextual information. For a keypoint, Topmax means that we look vertically towards the bottom to explicitly investigate the area below the keypoint, while Leftmax suggests that we look

horizontally towards the right to examine the area to the right of the keypoint. Similarly, Bottommax and Rightmax take us to look vertically towards the top and horizontally towards the left, respectively. With $H \times W$ feature maps, the computation of Topmax shown in Fig. 2 uses the following equations:

$$c_t^{(i,j)} = \begin{cases} \text{Max} \left(f_{c_t^{(i,j)}}, c_t^{(i+1,j)} \right) & \text{if } i < H \\ f_{c_t^{(H,j)}} & \text{otherwise} \end{cases} \quad (2)$$

where f_{c_t} is the feature maps that are the inputs to the Topmax operations. c_t is the feature maps of the output, i.e., keypoint-specific contextual information. $f_{c_t^{(i,j)}}$ is the vectors at (i, j) .

2) *Structured Contextual Information Modelling with Graphical Model*: Although the global contextual information defined in Eq. (1) can be used to directly infer the locations of keypoints, it pays less attention onto the local characteristics of each keypoint, i.e., keypoint-specific contextual information for describing the differences between body parts. Thus, in this subsection, we aim at modelling the dependency between global and keypoint-specific contextual information. Fig. 3 shows the structure of a graphical model based Structured Context Mixer. In Fig. 3(a), O_{hg} is the fusion of the output from multiple Hourglass modules, which preserve the multi-level feature representations. Here, we use five 3×3 Conv-BN-Relu layers to further refine the generated feature maps. One of them is considered as global contextual feature, whilst the other four maps are fed into Topmax, Leftmax, Bottommax and Rightmax shown in Fig. 2 respectively to generate keypoint-specific contextual features. Fig. 3(b) shows the proposed graphical model which seamlessly fuses two types of contextual features for describing each keypoint. In particular, we design two attention gates (A and G), which interact with each other to control the flow of the information between the coherent feature maps, i.e., global and keypoint-specific contextual feature maps.

Given the collected contextual features O , we aim to infer the hidden structured contextual feature representation H and the attention variables A and G . We formalise the problem within a conditional random field framework and define a graphical model by constructing the following conditional distribution:

$$P(H, A, G | I, \Theta) = \frac{1}{Z(I, \Theta)} \tilde{P}(H, A, G | I, \Theta) \quad (3)$$

where I represents the input image, $Z(I, \Theta) = Z$ is the partition function for normalisation, Θ is the set of parameters and $\tilde{P} = \exp\{-E(H, A, G, I, \Theta)\}$. The energy function E is defined as:

$$E(H, A, G, I, \Theta) = \Phi(H, O) + \Psi(H, A) + \Gamma(H, G) \quad (4)$$

The first term shown in Eq. (4) denotes the unary potentials related to the variables o_c^i and the corresponding latent variables h_c^i . It can be further defined as:

$$\Phi(H, O) = \sum_{c \in C} \sum_{i=1}^N \phi(h_c^i, o_c^i) \quad (5)$$

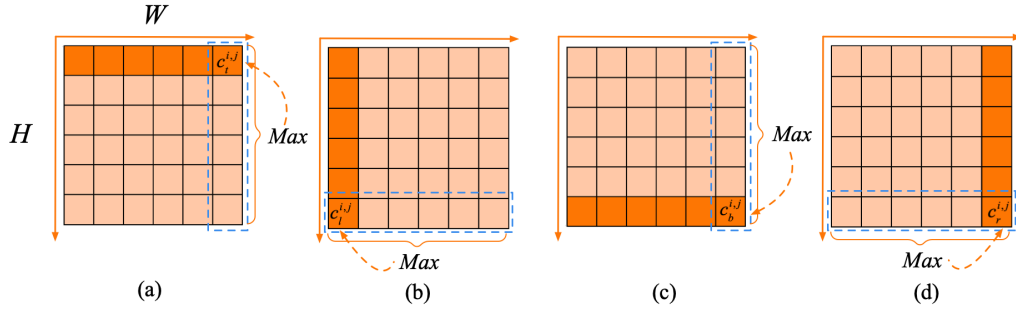


Fig. 2. Directionmax operation. (a) Topmax. (b) Leftmax. (c) Bottommax. (d) Rightmax. For Topmax and Leftmax, we scan from bottom to top for the vertical max-pooling and from right to left for the horizontal max-pooling respectively.

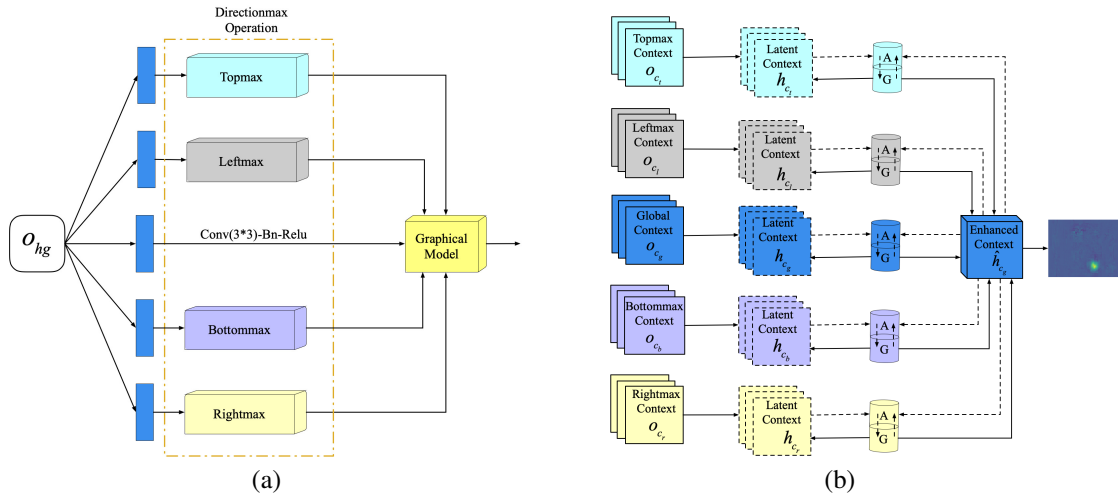


Fig. 3. Structured Context Mixer (SCM). (a) The whole structure of SCM. (b) The detailed architecture of our proposed graphical model. The arrows indicate the dependency between the variables used in our message passing algorithm, as shown in Algorithm 1. The solid arrows denote the updates from the keypoint-specific and global contextual feature maps, while the dashed ones represent the updates from the two attention gates.

According to the previous work [43], we wish to represent the latent features. Thus, we adopt a multidimensional Gaussian distribution to represent $\phi(h_c^i, o_c^i)$ as follows:

$$\phi(h_c^i, o_c^i) = -\frac{1}{2}(h_c^i - o_c^i)^T \Sigma^{-1}(h_c^i - o_c^i) \quad (6)$$

where $\phi(h_c^i, o_c^i)$ is the Manhattan Distance between h_c^i and o_c^i . Σ is covariance matrix. Here, we assume Σ is Identity Matrix I . The second and third terms shown in Eq. (4) are two branches to model the relationship between the latent keypoint-specific and the latent global contextual feature maps. For each branch, we design a gate to regulate the flow of the information between the two types of the contextual features. Inspired by [34], [43], we define two bilinear potentials, i.e. ψ_h and ζ_h , to represent the dependency between the latent keypoint-specific and the latent global contextual feature maps. $\Psi(H, A)$ and $\Gamma(H, G)$ can be defined as:

$$\Psi(H, A) = \sum_{c \neq c_g} \sum_{i,j=1}^N \psi(a_c^i, h_c^i, h_{c_g}^j) = \sum_{c \neq c_g} \sum_{i,j=1}^N a_c^i \psi_h(h_c^i, h_{c_g}^j) \quad (7)$$

$$\Gamma(H, G) = \sum_{c \neq c_g} \sum_{i,j=1}^N \zeta(g_c^i, h_c^i, h_{c_g}^j) = \sum_{c \neq c_g} \sum_{i,j=1}^N g_c^i \zeta_h(h_c^i, h_{c_g}^j) \quad (8)$$

where $\psi_h(h_c^i, h_{c_g}^j) = h_c^i \lambda_{i,j}^c h_{c_g}^j$, $\zeta_h(h_c^i, h_{c_g}^j) = h_c^i \varpi_{i,j}^c h_{c_g}^j$, $\lambda_{i,j}^c, \varpi_{i,j}^c \in \mathbb{R}^{L_c \times L_{c_g}}$, and L_c and L_{c_g} denote the number of the channels of the keypoint-specific and global contextual feature maps, respectively.

Using two branches with different gates, shown in Eqs. (7) and (8), we also consider the relationship between the two gates. Different from previous works that model the spatial relationship between pixels within a feature map [34], [44], we highlight the spatial dependency between feature maps. In our algorithm, the two gates are spatially interactive, guiding our proposed GM-SCENet to preserve useful keypoint-specific contextual information. We define:

$$\Xi(A, G) = \sum_{c \neq c_g} \sum_{i,j=1}^N \vartheta(a_c^i, g_c^j) = \sum_{c \neq c_g} \sum_{i,j=1}^N a_c^i \eta_{i,j}^c g_c^j \quad (9)$$

where $\eta_{i,j}^c \in \mathbb{R}^{L_c^a \times L_{c_g}^g}$, $L_c^a, L_{c_g}^g$ are the numbers of the channels of gate A and G , respectively.

To deduce $P(H, A, G | I, \Theta)$ shown in Eq. (3), we adopt classical mean-field approximation [45] which is effective and efficient to handle high-dimensional data. $P(H, A, G | I, \Theta)$ can be approximated by a new distribution $Q(H, A, G | I, \Theta)$, which is represented as a product of independent distributions

as follows:

$$Q(H, A, G | I, \Theta) = \prod_{i=1}^N q_i(h_c^i) q_i(h_{c_g}^i) q_i(a_c^i) q_i(g_c^i) \quad (10)$$

where $q_i(h_c^i)$, $q_i(h_{c_g}^i)$, $q_i(a_c^i)$ and $q_i(g_c^i)$ are independent marginal distributions. Afterwards, we minimise the differences between the two distributions using Kullback–Leibler (KL) divergence, which is formulated as follows:

$$\begin{aligned} D_{KL}[Q(H, A, G | I, \Theta) || P(H, A, G | I, \Theta)] \\ = \sum_x Q(x) \log Q(x) - \sum_x Q(x) \log P(x) \\ = \sum_x Q(x) \log Q(x) - \sum_x Q(x) \log \tilde{P} + \sum_x Q(x) \log Z \end{aligned} \quad (11)$$

Our target is to minimise the right hand side of the above equation. We convert Eq. (11) to the following form:

$$\begin{aligned} L(h_c^i) = \sum_{h_c^i} q_i(h_c^i) \log(q_i(h_c^i)) - \\ \sum_{h_c^i} q_i(h_c^i) \mathbb{E}_{\bar{q}_i(h_c^i)} \{\log \tilde{P}\} + C \end{aligned} \quad (12)$$

$$\begin{aligned} \bar{q}_i(h_c^i) = q_i(h_{c_g}^i) q_i(a_c^i) q_i(g_c^i) \cdot \\ \prod_{j \neq i} q_j(h_{c_g}^j) q_j(h_{c_g}^j) q_j(a_c^j) q_j(g_c^j) \end{aligned} \quad (13)$$

where $\mathbb{E}_{\bar{q}_i(h_c^i)}$ represents the expectation of distribution $\bar{q}_i(h_c^i)$, and C is the constant. Then, we minimise $L(h_c^i)$, which can be regarded as a constrained optimisation problem. Given the condition $\int q_i(h_c^i) dh_c^i = 1$, we rewrite Eq. (12) as:

$$\begin{aligned} \arg \min_{q_i} (L(h_c^i)) = \arg \min_{q_i} \left(\int q_i(h_c^i) \log q_i(h_c^i) dh_c^i - \right. \\ \left. \int q_i(h_c^i) \mathbb{E}_{\bar{q}_i(h_c^i)} \{\log \tilde{P}\} dh_c^i + \lambda_i \left(\int q_i(h_c^i) dh_c^i - 1 \right) \right) \end{aligned} \quad (14)$$

To seek the minimum, we take the derivative of $L(h_c^i)$ with respect to $q_i(h_c^i)$ and set the derivative to zero, as shown in Eq. (15). To achieve the minimisation, we can derive the mean-field update shown in Eq. (16) by combining Eqs. (3)-(10).

$$\frac{\partial L(h_c^i)}{\partial q_i(h_c^i)} = \log q_i(h_c^i) - \mathbb{E}_{\bar{q}_i(h_c^i)} \{\log \tilde{P}\} + \lambda_i = 0 \quad (15)$$

$$\begin{aligned} q_i(h_c^i) \propto \exp \left(\phi(h_c^i, o_c^i) + \right. \\ \left. \mathbb{E}_{q_i(a_c^i)} \{a_c^i\} \sum_{j=1}^N \mathbb{E}_{q_i(h_{c_g}^j)} \{\psi_h(h_c^i, h_{c_g}^j)\} + \right. \\ \left. \mathbb{E}_{q_i(g_c^i)} \{g_c^i\} \sum_{j=1}^N \mathbb{E}_{q_i(h_{c_g}^j)} \{\zeta_h(h_c^i, h_{c_g}^j)\} \right) \end{aligned} \quad (16)$$

Eqs. (12)-(16) show the process of mean-field approximation where latent variables are updated, while we fix the remaining latent variables. Following this way, $q_i(h_{c_g}^i)$, $q_i(a_c^i)$, $q_i(g_c^i)$ can be derived as:

$$\begin{aligned} q_i(h_{c_g}^i) \propto \exp \left(\phi(h_{c_g}^i, o_{c_g}^i) + \right. \\ \left. \sum_{c \neq c_g} \sum_{j=1}^N \mathbb{E}_{q_i(a_c^j)} \{a_c^j\} \mathbb{E}_{q_i(h_c^j)} \{\psi_h(h_{c_g}^i, h_c^j)\} + \right. \\ \left. \sum_{c \neq c_g} \sum_{j=1}^N \mathbb{E}_{q_i(g_c^j)} \{g_c^j\} \mathbb{E}_{q_i(h_c^j)} \{\zeta_h(h_{c_g}^i, h_c^j)\} \right) \end{aligned} \quad (17)$$

$$\begin{aligned} q_i(a_c^i) \propto \exp \left(a_c^i \mathbb{E}_{q_i(h_c^i)} \left\{ \sum_{j=1}^N \mathbb{E}_{q_i(h_{c_g}^j)} \{\psi_h(h_c^i, h_{c_g}^j)\} \right\} + \right. \\ \left. \sum_{c \neq c_g} \sum_{j=1}^N \mathbb{E}_{q_i(g_c^j)} \{\vartheta(a_c^i, g_c^j)\} \right) \end{aligned} \quad (18)$$

$$\begin{aligned} q_i(g_c^i) \propto \exp \left(g_c^i \mathbb{E}_{q_i(h_c^i)} \left\{ \sum_{j=1}^N \mathbb{E}_{q_i(h_{c_g}^j)} \{\zeta_h(h_c^i, h_{c_g}^j)\} \right\} + \right. \\ \left. \sum_{c \neq c_g} \sum_{j=1}^N \mathbb{E}_{q_i(a_c^j)} \{\vartheta(g_c^i, a_c^j)\} \right) \end{aligned} \quad (19)$$

Eqs. (16) and (17) show the posterior distributions of h_c^i and $h_{c_g}^i$ in Gaussian distributions. We denote the expectation of h_c^i , $h_{c_g}^i$, a_c^i and g_c^i by $\hat{h}_c^i = \mathbb{E}_{q_i(h_c^i)} \{h_c^i\}$, $\hat{h}_{c_g}^i = \mathbb{E}_{q_i(h_{c_g}^i)} \{h_{c_g}^i\}$, $\hat{a}_c^i = \mathbb{E}_{q_i(a_c^i)} \{a_c^i\}$ and $\hat{g}_c^i = \mathbb{E}_{q_i(g_c^i)} \{g_c^i\}$ respectively. By combining Eqs. (6), (7) and (8), we derive the following mean-field updates for the latent structured contextual representations:

$$\hat{h}_c^i = o_c^i + \hat{a}_c^i \sum_{j=1}^N \lambda_{i,j}^c \hat{h}_{c_g}^j + \hat{g}_c^i \sum_{j=1}^N \varpi_{i,j}^c \hat{h}_{c_g}^j \quad (20)$$

$$\hat{h}_{c_g}^i = o_{c_g}^i + \sum_{c \neq c_g} \sum_{j=1}^N \hat{a}_c^j \lambda_{i,j}^c \hat{h}_c^j + \sum_{c \neq c_g} \sum_{j=1}^N \hat{g}_c^j \varpi_{i,j}^c \hat{h}_c^j \quad (21)$$

Algorithm 1 Algorithm for mean-field updates in our proposed Structured Context Mixer with CNN.

Input: Global contextual feature maps \hat{h}_{c_g} initialised with corresponding feature observation o_{c_g} , keypoint-specific contextual feature maps \hat{h}_c initialized with o_c , and the number of iteration N .

Output: Enhanced structured contextual feature maps, i.e., global contextual feature maps generated in the last iteration \hat{h}_{c_g} shown in Fig. 3(b) and Eq. (21).

- 1: **for** $i \leftarrow 1$ to N **do**
- 2: $\tilde{g}_c \leftarrow \hat{h}_c \odot (\varpi_c \otimes \hat{h}_{c_g})$, $\tilde{a}_c \leftarrow \hat{h}_c \odot (\lambda_c \otimes \hat{h}_{c_g})$;
- 3: $\tilde{g}_c \leftarrow \eta_c \otimes \tilde{a}_c$, $\tilde{a}_c \leftarrow \eta_c \otimes \tilde{g}_c$;
- 4: $\hat{g}_c \leftarrow \sigma(-(\tilde{g}_c \oplus \tilde{g}_c))$;
- 5: $\hat{a}_c \leftarrow \sigma(-(\tilde{a}_c \oplus \tilde{a}_c))$;
- 6: $\tilde{h}_{c_g} \leftarrow \sum_{c \neq c_g} \hat{g}_c \odot (\varpi_c \otimes \hat{h}_c)$;
- 7: $\hat{h}_{c_g} \leftarrow \sum_{c \neq c_g} \hat{a}_c \odot (\lambda_c \otimes \hat{h}_c)$;
- 8: $\hat{h}_{c_g} \leftarrow o_{c_g} \oplus \tilde{h}_{c_g} \oplus \hat{h}_{c_g}$;
- 9: $\tilde{h}_c \leftarrow \hat{g}_c \odot (\varpi_c \otimes \hat{h}_{c_g})$;
- 10: $\hat{h}_c \leftarrow \hat{a}_c \odot (\lambda_c \otimes \hat{h}_{c_g})$;
- 11: $\hat{h}_c \leftarrow o_c \oplus \tilde{h}_c \oplus \hat{h}_c$;
- 12: **end for**
- 13: **return** Enhanced structured contextual feature maps \hat{h}_{c_g} .

In Section III-A2, we define a_c^i and g_c^i as binary variables. Therefore, the expectations of their distributions are the same

as $q(a_c^i = 1)$ and $q(g_c^i = 1)$ respectively. The expectations \hat{a}_c^i and \hat{g}_c^i can be derived using the distributions shown in Eqs. (18) and (19) and the potential functions defined in Eqs. (7), (8) and (9):

$$\hat{a}_c^i = \sigma\left(-\sum_{j=1}^N \hat{h}_c^i \lambda_{i,j}^c \hat{h}_{c_g}^j - \sum_{c \neq c_g} \sum_{j=1}^N \eta_{i,j}^c \hat{g}_c^j\right) \quad (22)$$

$$\hat{g}_c^i = \sigma\left(-\sum_{j=1}^N \hat{h}_c^i \varpi_{i,j}^c \hat{h}_{c_g}^j - \sum_{c \neq c_g} \sum_{j=1}^N \eta_{i,j}^c \hat{a}_c^j\right) \quad (23)$$

where $\sigma(\cdot)$ is the sigmoid function. Using Eq. (23), the updates for gate g_c^i depend on the values of the hidden features h_c^i and the other gate a_c^i . In particular, the spatially interactive gates allow us to adjust the distribution of the hidden features, i.e., keypoint-specific and the global contextual features shown in Eqs. (20) and (21), respectively.

3) *End-to-End Optimisation*: Following [43], [46], we convert the mean-field inference equations shown in Eqs. (20)-(23) to convolutional operations by deep neural networks for jointly training our Structured Context Mixer and the network. Our goal is to achieve end-to-end optimisation of the proposed GM-SCENet.

The mean-field updates of the two gates shown in Eqs. (22) and (23) can be implemented with deep neural networks in several steps as follows: (a) Message passing between the global contextual feature maps \hat{h}_{c_g} and the keypoint-specific contextual feature maps \hat{h}_c can be performed by $\tilde{g}_c \leftarrow \hat{h}_c \odot (\varpi_c \otimes \hat{h}_{c_g})$ and $\tilde{a}_c \leftarrow \hat{h}_c \odot (\lambda_c \otimes \hat{h}_{c_g})$, where ϖ_c and λ_c represents the convolutional kernels and the symbols \odot and \otimes denote the element-wise product and convolutional operation, respectively; (b) Message passing between the two gates is performed by $\hat{g}_c \leftarrow \eta_c \otimes \tilde{a}_c$ and $\hat{a}_c \leftarrow \eta_c \otimes \tilde{g}_c$, where η_c is a convolution kernel; (c) The normalisation of the two gates is performed by $\hat{g}_c \leftarrow \sigma(-(\tilde{g}_c \oplus \hat{g}_c))$ and $\hat{a}_c \leftarrow \sigma(-(\tilde{a}_c \oplus \hat{a}_c))$, where \oplus represents the element-wise addition operation.

Afterwards, we conduct mean-field updating of the global contextual feature maps h_{c_g} shown in Eq. (21) after having updated the two gates. This process goes through: (a) Message passing from the keypoint-specific to the global contextual feature maps under the control of gate g_c is achieved by $\hat{h}_{c_g} \leftarrow \sum_{c \neq c_g} \hat{g}_c \odot (\varpi_c \otimes \hat{h}_c)$; (b) Message passing from the keypoint-specific to the global contextual feature maps under the control of gate a_c is performed by $\hat{h}_{c_g} \leftarrow \sum_{c \neq c_g} \hat{a}_c \odot (\lambda_c \otimes \hat{h}_c)$. (c) The final update involves unary term, i.e., the observed global contextual feature maps o_{c_g} can be performed by $\hat{h}_{c_g} \leftarrow o_{c_g} \oplus \tilde{h}_{c_g} \oplus \hat{h}_{c_g}$.

Similarly, following Eq. (20), the updating of the latent feature maps corresponding to the keypoint-specific contextual information is straightforward. The mean-field updates in our proposed SCM are summarised in Algorithm 1.

B. Cascaded Multi-Level Supervision

Since mice are small and highly deformable, single-scale supervision cannot work stably in different set-ups [26], [17].

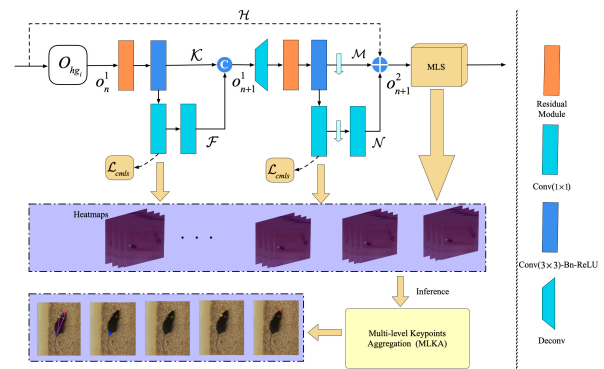


Fig. 4. Architecture of the Cascaded Multi-level Supervision module (CMLS).

In this section, we design a novel Cascaded Multi-level Supervision module (CMLS) consisting of two identical Multi-level Supervision blocks (MLS) as intermediate supervision to jointly train the proposed SCM and the Hourglass network, where multi-level features, i.e., multi-stage features of multiple scales, are used for supervised learning, as shown in Figs. 1 and 4. HRNet [28] also includes large and small scale representations, but this model performs multi-scale fusion among sub-networks without considering the multi-scale features for supervised learning. On the contrary, our MLS block provides a coarse-to-fine supervision to strengthen multi-scale feature learning. Additionally, different from [35], our CMLS module is used for module-wise supervision rather than layer-wise supervision. In other words, we adopt this module to connect different Hourglass modules and further refine the prediction results from SCM, as shown in Fig. 1. Particularly, for each MLS block, we combine the supervision information with historical information by the *Concat* operation in the first stage, before adopting a *Deconv* layer (Deconvolution) to extract high-resolution information in the second stage. We use the *Concat* here to preserve feature maps of higher dimensions. This allows the *Deconv* operation to capture more high-resolution information, which can provide more detailed features, thus benefiting the large-scale supervision in the second stage. Similar to [26], we also use identity mapping to add the input information to the output of the second stage. Furthermore, two MLS blocks are concatenated to build CMLS module where we take the output of the first MLS block o_{n+1}^2 as the input of the other MLS block. In this way, small-scale supervision information is used as prior knowledge for large-scale supervision, and then the large-scale supervision information is also considered as prior knowledge of the following MLS module. In addition, the generated multi-level information can drive our SCM to strengthen the contextual feature learning. Therefore, such coarse-to-fine process can help our GM-SCENet to generate multi-level feature maps for precisely localising the keypoints of the mouse.

The mathematical formulation of the first and second stages in our proposed CMLS are as follows:

$$o_{n+1}^1 = \text{Concat}(\mathcal{K}(o_n^1, W_n^K), \mathcal{F}(o_n^1, W_n^F)) \quad (24)$$

$$o_{n+1}^2 = \mathcal{H}(o_n^2) + \mathcal{M}(o_n^2, W_n^M) + \mathcal{N}(o_n^2, W_n^N) \quad (25)$$

where $\text{concat}(\cdot)$ denotes the channel-wise concatenation, o_n^1 and o_{n+1}^1 are the input and output of the first stage of the n -th MLS module. o_n^2 and o_{n+1}^2 are the input and output of the second stage of the n -th MLS module and $o_{n+1}^1 = o_n^2$. $\mathcal{H}(o_n^2) = o_n^2$ refers to the identity mapping. $\mathcal{K}(o_n^1, W_n^K)$ is a stack of the Residual Module and $\text{Conv}(3 \times 3)$ -Bn-ReLU. $\mathcal{F}(o_n^1, W_n^F)$ represents two 1×1 convolutions. $\mathcal{M}(o_n^2, W_n^M)$ is a stack of deconvolution and residual blocks followed by $\text{Conv}(3 \times 3)$ -Bn-ReLU and downsampling operation. $\mathcal{N}(o_n^2, W_n^N)$ is a combination of 1×1 convolution, downsampling operation and 1×1 convolution.

We use the Mean-Squared Error (MSE) based loss function [22], [34] for the training of the entire network. To represent the ground-truth keypoint, we generate a heatmap h_k for each single keypoint k , $k \in \{1, 2, \dots, K\}$ by a 2D Gaussian distribution centered at the mouse part position $l_k = (x_k, y_k)$. We apply a MSE loss function to the proposed SCM and the CMLS module (i.e., \mathcal{L}_{scm} and \mathcal{L}_{cmls} in Fig. 1) as follows:

$$\mathcal{L} = \sum_{k=1}^K \sum_{p \in \mathbb{P}} \|\hat{h}_k(p) - h_k(p)\|_2^2 \quad (26)$$

where $p \in \mathbb{P} = \{(x, y) \mid x = 1, \dots, W, y = 1, \dots, H\}$ denotes the p -th location, \hat{h}_k denotes the predicted heatmap of keypoint k . In our system, we add all the supervision losses from SCM and the CMLS module. During the inference phase, we obtain the location of the keypoint from the predicted heatmap \hat{h}_k by choosing the position with the maximum score, i.e., $\hat{l}_k = \arg \max \hat{h}_k$. Then, combining all the predicted heatmaps from the inner layers of the GM-SCENet and using Multi-Level Keypoint Aggregation introduced in Section III-C, we have the final prediction results.

C. Multi-level Keypoints Aggregation for Inference

In general, the final prediction results of deep models for pose estimation come from the last stage of the networks. Here, we argue that results generated in the inner layers may be more accurate if we apply intermediate supervision including our proposed CMLS module. Similar to the popular Non-Maximum Suppression (NMS) postprocessing step used in object detection, most top-down multi-person pose estimation also adopts this method to remove unreasonable poses [47]. Inspired by this concept, we propose an inference algorithm for single-mouse pose estimation, i.e., Multi-level Keypoint Aggregation (MLKA), to integrate prediction results generated by SCM and the CMLS module. We adopt the CMLS module as intermediate supervision for training our proposed framework, and consider all the predicted heatmaps with multiple scales as potential candidates, as shown in Fig. 4.

During the inference phase, we first obtain the positions of the keypoints from the generated heatmaps before adopting MLKA to calculate the average heatmaps of the selected multi-level keypoints, i.e., multi-scale keypoints from different CMLS modules. Specifically, we change the coordinates of different keypoints in proportion to the corresponding image sizes before choosing the keypoint candidates which are close to those from the last stage. We then use the Object Keypoint Similarity (OKS) metric to examine the similarities of the

keypoints from the same part. After having selected all the potential keypoints of different parts by the preset threshold, we aggregate all the selected keypoints of body parts by directly calculating the average of them to generate a new set of keypoints to reconstruct the skeleton. The OKS metric has been used for human pose estimation in different datasets [28]. Therefore, we use the modified OKS metric to measure the similarity between the results generated in the last stage and those predicted values in the inner layers. The modified OKS can be formed as:

$$OKS = \frac{\sum_i \exp(-d_i^2/k_i^2) \delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \quad (27)$$

where d_i^2 is the Euclidean distance between the detected i -th keypoint and the corresponding ground truth, and $\delta(\cdot)$ is the Kronecker function. $\delta(\cdot) = 1$ if the condition holds, otherwise 0. v_i is the visibility flag of the ground truth, and k_i is a per-keypoint constant that controls the falloff.

IV. EXPERIMENTAL SETUP

To evaluate the performance of our proposed methods, we implement comprehensive evaluations on two mouse pose datasets (i.e., DeepLabCut Mouse Pose¹ and our PDMB datasets) and other two animal datasets (i.e., Zebra [22] and Fly [23]) and a human pose dataset (i.e., LSP [48]). In this section, we first introduce the two mouse datasets and evaluation metrics. Then, we describe the implementation details. We also provide a supplemental material including more experimental results, and details about the used datasets (i.e., Supplementary A to H) at: <https://github.com/FeixiangZhou/GM-SCENet/blob/main/Supplementary.pdf>.

A. Datasets and Evaluation Metrics

1) *DeepLabCut Mouse Pose Dataset*: DeepLabCut Mouse Pose dataset: The dataset is recorded by two different cameras with the resolution of 640*480 and 1700*1200 pixels respectively. Most images are 640*480, while some images were cropped around mice to generate the images that are approximately 800*800. There are 1066 images from multiple experimental sessions of 7 different mice. We randomly split the dataset into a training set of 853 images and a test set of 213 images. In addition, four parts (i.e., snout, left ear, right ear and tail base) are labeled in this dataset.

2) *PDMB Dataset*: In this paper, we introduce a new dataset that is collected in collaboration with the biologists of Queen's University Belfast of United Kingdom, for a study on motion recordings of mice with Parkinson's disease (PD) [49]. The neurotoxin 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) is used as a model of PD, which has become an invaluable aid to produce experimental parkinsonism since its discovery in 1983 [50]. All the mice receive treatment of MPTP and they are housed in a controlled environment with the constant temperature of (27°C) and light condition (long fluorescent lamp of 40W), and under constant climatic

¹<https://zenodo.org/record/4008504#.X1KM7mczYW8>

conditions with free access to food and water (placed on the corner of the cage). All experimental procedures are performed in accordance with the Guidance on the Operation of the Animals (Scientific Procedures) Act, 1986 (UK) and approved by the Queen's University Belfast Animal Welfare and Ethical Review Body.

We record 4 videos for 4 mice using a Sony Action camera (HDR-AS15) (top-view) with a frame rate of 30 fps and 640*480 resolution. For each video, We divide it into 6 clips, and each clip lasts about 10 minutes. Then, we extract frames at different frequencies and build our top-view PDMB dataset with 4248, 3000, 2000 images for training, validation and testing, respectively. Particularly, our dataset is more challenging than the existing DeepLabCut Mouse Pose dataset. For example, our PDMB dataset contains a wide range of behaviours [4] (e.g., rear, groom and eat), and abnormal poses. Fig. S5 shows video frames for different behaviours. In addition, the PDMB dataset also provides an additional annotation indicates the lack of visibility. More details about the data annotation are described in Supplementary G.

3) *Evaluation Metrics*: Following [17], we use Root Mean Square Error (RMSE) as an evaluation metric. Furthermore, we introduce a new evaluation metric, i.e., Percentage of Correct Keypoints (PCK) to the PDMB and DeepLabCut Mouse Pose datasets. Different from the PCK metric used for the datasets of human pose estimation, we modify PCK and apply it to our PDMB dataset, as shown in Eq. (28). A detected keypoint is considered correct if the normalized distance between the predicted and the ground-truth keypoints falls within a certain threshold T . In our experiments, T is set to 0.2, and we also draw PCK curves with different thresholds to compare different approaches.

$$PCK@T = \frac{\sum_k \delta \left(\left\| \frac{\hat{l}_k - l_k}{\vec{c}} \right\|_2 < T \right)}{\sum_k 1} \quad (28)$$

where T represents the threshold evaluating the normalized distance between \hat{l}_k and l_k , \vec{c} represents a constant vector for normalisation.

B. Implementation Details

All the experiments are performed on a server with an Intel Xeon CPU @ 2.40GHz and two 16GB Nvidia Tesla P100 GPUs. The parameters are optimised by the Adam algorithm. For the DeepLabCut Mouse Pose dataset, we use the initial learning rate of 1e-4 and all the training and test images are resized to the resolution of 640*480. For the PDMB dataset, the initial learning rate is set to 1e-5. We adopt the validation split of the PDMB dataset to monitor the training process. The proposed GM-SCENet is implemented with Pytorch. The source code and dataset will be available at: <https://github.com/FeixiangZhou/GM-SCENet>.

V. RESULTS AND DISCUSSION

In this section, we implement comprehensive experiments to evaluate the performance of our proposed methods. Firstly, to validate the effectiveness of the proposed SCM, CMLS module

and MLKA, we conduct ablation experiments on the proposed PDMB validation and DeepLabCut Mouse pose test datasets. We use the 3-stack Hourglass network as our baseline network. Based on this, we first investigate each proposed component, followed by the comprehensive analysis for the impact of each module. Then, we compare our network with the other state-of-the-art networks on both PDMB and DeepLabCut Mouse pose test datasets.

TABLE I
ABLATION EXPERIMENTS FOR THE STRUCTURED CONTEXT MIXER ON THE PDMB VALIDATION DATASET (RMSE(PIXELS) AND PCK@0.2).

Methods	RMSE	Snout	Left ear	Right ear	Tail base	Mean
Baseline	4.89	93.38	92.67	93.97	86.87	91.72
<i>SCM(Conv(3 × 3))</i>						
SCM(Iteration 1)	4.98	94.99	93.87	95.77	83.4	92.0
SCM(Iteration 2)	3.74	96.90	96.50	96.07	88.87	94.58
SCM(Iteration 3)	3.07	97.67	97.83	97.17	91.67	96.08
<i>SCM(Conv(1 × 1))</i>						
SCM(Iteration 1)	4.79	92.55	86.83	93.60	85.83	89.70
SCM(Iteration 2)	3.91	95.13	91.97	96.90	88.93	93.23
SCM(Iteration 3)	3.83	95.89	95.20	97.53	87.80	94.11
<i>SCM(Conv(1 × 1)+Conv(3 × 3))</i>						
SCM(Iteration 1)	4.30	94.15	92.87	97.17	87.63	92.95
SCM(Iteration 2)	3.76	96.62	91.27	97.13	90.07	93.77
SCM(Iteration 3)	3.55	97.81	91.67	97.60	89.53	94.15

A. Ablation Studies on the Structured Context Mixer

In this experiment, we investigate the influence of the proposed Structured Context Mixer with different numbers of iteration during the message passing process and different kernel sizes of λ_c , ϖ_c and η_c shown in Algorithm 1. Tables I and S1 show the performance comparisons of different networks on the PDMB validation and DeepLabCut Mouse Pose test datasets, respectively. We use 3 types of convolutional kernels where Conv(3×3) + Conv(1×1) means that the gate a_c related kernels use 3×3 convolution with stride 1 and padding 1 to keep the same spatial resolution, while the gate g_c related kernels use 1×1 convolution with stride 1 and padding 1. We do not choose larger convolutional kernels because this will affect the efficiency of the whole network. As shown in Table I, with 3×3 convolutions, we achieve the lowest RMSE of 3.07 pixels and the highest mean PCK score of 96.08%. When the iteration number is set to 2, we observe that, compared with the networks with other two types of SCM, the mean PCK score increases from 93.23% and 93.77% to 94.58% respectively by adopting SCM with 3×3 convolutions. The same trend can be observed when the iteration is set to 3. Actually, larger convolutional kernels help us to increase the receptive field of the network, which enables SCM to focus on larger context regions.

With respect to the number of iterations in the phase of inferring hidden structured context features, we train the networks with 1, 2 and 3 iterations, respectively. The results on the PDMB validation set show that the performances of the networks are continuously improved with the number of iterations increasing, which demonstrates that the proposed SCM

can adaptively learn the structured contextual information of each keypoint considering the differences between mouse body parts. However, on the DeepLabCut Mouse Pose test set, the performance becomes worse when the number of iterations is set to 3, as shown in Table S1. The main reason is that the complexity of the network will increase as the number of iterations increases, requiring more samples for training. Overall, Although the network using SCM (Conv(3×3)) with 3 iterations holds the lowest RMSE and the highest PCK score on the PDMB validation set, a higher number of iterations will consume more computational overhead. Therefore, in our implementation, we set the number of iterations to 2 on both two datasets. In particular, all the convolutions in the first and second iterations share the same weights to reduce the parameters of the network for faster implementation.

For some challenging mouse keypoints on the PDMB dataset, e.g., tail base, which is occluded frequently, we receive the PCK score of 91.67%, which is 4.8% improvement compared to the baseline model. This experimentally confirms that adding SCM to explore the keypoint-specific contextual information helps to determine the positions of the occluded parts.

B. Ablation Studies on the Cascaded Multi-level Supervision Module

TABLE II
ABLATION EXPERIMENTS FOR THE CASCADED MULTI-LEVEL SUPERVISION MODULE ON THE PDMB VALIDATION DATASET (RMSE(PIXELS) AND PCK@0.2).

Methods	RMSE	Snout	Left ear	Right ear	Tail base	Mean
Baseline	4.89	93.38	92.67	93.97	86.87	91.72
MLS(All Cas1)	4.45	93.49	96.10	98.16	86.93	93.67
CMLS(Start Cas2)	5.75	92.13	97.20	97.93	86.37	93.41
CMLS(Middle Cas2)	4.98	94.78	91.07	96.63	85.60	92.02
CMLS(Final Cas2)	5.39	97.08	94.0	97.8	86.73	93.90
CMLS(All Cas2)	13.61	95.19	93.07	96.33	78.83	90.85

In this subsection, we investigate the effect of the CMLS module on the prediction results. In our experiments, we design 5 structures using the baseline model. MLS (All Cas1) refers to the case where we use the MLS block as intermediate supervision at the end of the first and second Hourglass modules and our SCM, while CMLS (All Cas2) refers to a cascaded structure composed of 2 identical MLS blocks. For CMLS (Start Cas2), CMLS (Middle Cas2) and CMLS (Final Cas2), we design the cascaded structure at the end of the first Hourglass, the second Hourglass and SCM, respectively, while the MLS blocks are kept on the other two positions. As shown in Tables II and S2, we observe that all the networks with the CMLS module except the last one can improve the PCK score of the baseline model. However, we do not witness significant decrease on the RMSE of all the mouse parts for almost all the networks. The reason is that multi-level supervision applied to the inner layers of the network would affect the localisation results of the challenging mouse parts. As we can see from Table II, CMLS (All Cas2) has the

TABLE III
ABLATION EXPERIMENTS FOR THE MULTI-LEVEL KEYPOINTS AGGREGATION ON THE PDMB VALIDATION DATASET (RMSE(PIXELS) AND PCK@0.2).

Methods	RMSE	Snout	Left ear	Right ear	Tail base	Mean
SCM+CMLS(w/o MLKA)	2.89	97.35	98.70	98.80	90.90	96.44
SCM+CMLS(with small-scale MLKA)	2.90	97.84	99.17	98.60	91.87	96.87
SCM+CMLS(with large-scale MLKA)	2.81	97.60	98.90	98.67	91.40	96.64
SCM+CMLS(with multi-scale MLKA)	2.81	97.84	99.20	98.60	91.87	96.88

PCK score of 78.83% for tail base, which is the worst PCK score among all the networks. On the other hand, compared with the baseline model, CMLS (Start Cas2) and MLS (All Cas1) obtain 4.53% and 4.19% improvement respectively for left and right ears, demonstrating that our Cascaded Multi-level Supervision module can refine the prediction results providing large-scale (high-resolution) learning information. Furthermore, MLS (All Cas1) achieves the mean PCK score of 93.67% and the score is further improved to 93.90% by a designed cascaded structure applied to the final stage of SCM. This indicates that our CMLS module outperforms the single-scale supervision scheme.

C. Ablation Studies on Multi-level Keypoints Aggregation

The CMLS module generates multi-level features for learning during the training phase, as described in Section III-B. After having obtained all the multi-level predicted keypoints, we aggregate all the chosen keypoints via the MLKA strategy. In our experiments, we design 3 schemes to evaluate the performance of the proposed MLKA where we use the network consisting of the 3-stack Hourglass network, SCM (Conv(3×3), Iteration 2) and CMLS (Final Cas2) for comparison. IN the first scheme, we generate keypoints by aggregating all the selected small-scale keypoints and the predicted keypoints from the last stage. In the other two schemes, we aggregate all the selected large- and multi-scale keypoints respectively. The threshold of the keypoint similarity in all the experiments is set to 0.2. As shown in Table III, we observe that the inference strategy combined with small-, large- and multi-scale MLKA respectively can improve the mean PCK score. In addition, the inference strategy with multi-scale MLKA has the lowest RMSE and the best mean PCK score among these three schemes on the PDMB validation set. Besides, we have 0.24% and 0.01% improvement by replacing the small- and large-scale MLKA strategies with the multi-scale MLKA respectively, and obtain 0.44% improvement by applying multi-scale MLKA to the inference phase of the baseline network.

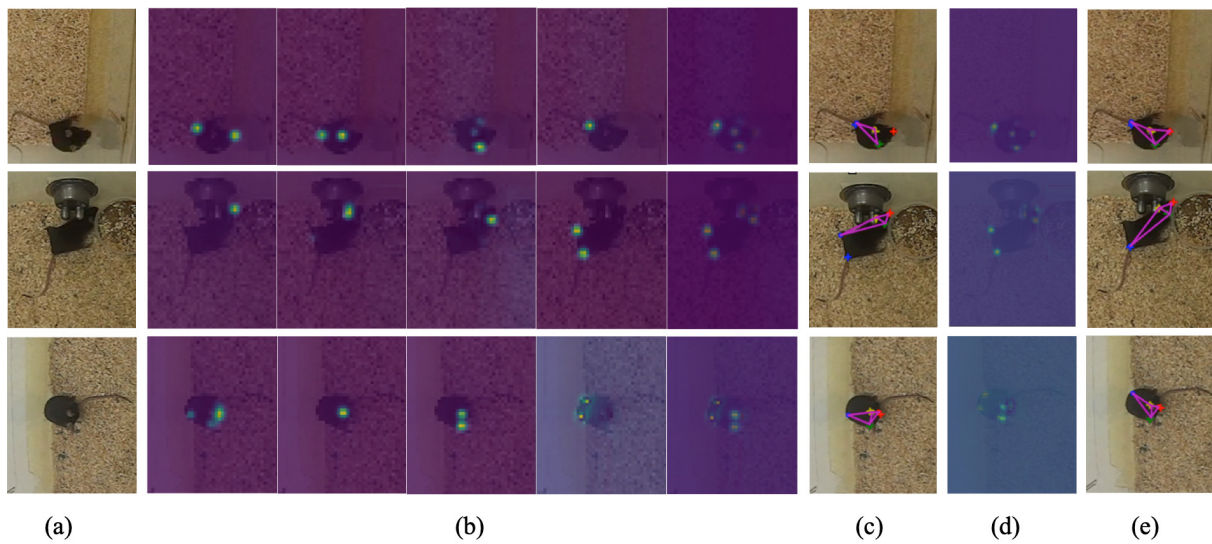


Fig. 5. Visualisation of the prediction results from the Structured Context Mixer and the Cascaded Multi-level Supervision module of our GM-SCENet. (a) The input images (we crop the original images (640*480) for better display). (b) Heatmaps predicted by the Structured Context Mixer for 4 parts (160*120). The last column in (b) are the heatmaps of all the body parts. (c) Mouse skeletons generated by the Structured Context Mixer (Dots with different colors represent the predicted results of different parts, the symbol '+' represents the ground truth, and the purple lines are the skeletons). (d) Further refined heatmaps (320*240) by the Cascaded Multi-level Supervision module. (e) Further refined skeletons by the Cascaded Multi-level Supervision module.

D. Comprehensive Analysis on Structured Context Mixer and Cascaded Multi-level Supervision Module

We also investigate the contribution of the proposed Structured Context Mixer and the Cascaded Multi-level Supervision module to the whole network on both PDMB validation and DeepLabCut Mouse Pose test sets. In addition to adding each proposed module to the baseline model separately, we further employ the two proposed modules to the baseline model simultaneously. The results are reported in Table IV. On the two datasets, our method achieves the highest mean PCK scores, 96.44% and 97.89% respectively, with the two proposed modules onboard simultaneously, which are of 4.72% and 7.15% increment compared to the baseline model. Additionally, the network combining SCM and the CMLS module achieves the lowest RMSE of 2.89 pixels on the PDMB validation set and the lowest RMSE of 3.13 pixels on the DeepLabCut Mouse Pose test set, respectively.

TABLE IV
ABLATION EXPERIMENTS FOR THE STRUCTURED CONTEXT MIXER AND THE CASCADED MULTI-LEVEL SUPERVISION MODULE ON THE PDMB VALIDATION AND DEEPLABCUT MOUSE POSE TEST DATASETS (RMSE(PIXELS) AND PCK@0.2).

Dataset	SCM	CMLS	RMSE	Snout	Left ear	Right ear	Tail base	Mean
PDMB	✓		4.89	93.38	92.67	93.97	86.87	91.72
			3.74	96.90	96.50	96.07	88.87	94.58
		✓	5.39	97.08	94.0	97.8	86.73	93.90
	✓	✓	2.89	97.35	98.70	98.80	90.90	96.44
DeepLabCut Mouse Pose	✓		3.83	89.35	90.28	90.28	93.06	90.74
			3.67	95.77	96.24	90.61	98.12	95.19
		✓	3.52	94.84	97.65	95.77	97.65	96.48
	✓	✓	3.13	96.71	98.60	97.65	98.59	97.89

We also observe that the overall performance of the network combined with only the CMLS module outperforms that of the SCM based network on the DeepLabCut Mouse Pose dataset. Different from the tail base in the PDMB dataset, the CMLS based network does not result in much decrease on the PCK score of snout which is a relatively challenging part in the DeepLabCut Mouse Pose dataset. On the contrary, the CMLS based network improves the PCK score of snout from 89.35% to 94.84%. As we discussed in Section V-B, the localisation errors of the challenging mouse parts would be amplified when only the CMLS module is used, although those of relatively easy parts can be further reduced. However, the CMLS based network can effectively deal with the problem of scaling variations by providing multi-level supervised information. Actually, the size of the mouse in the input image is different since the images in the DeepLabCut Mouse Pose dataset have different resolutions. Therefore, our network can achieve better performance by combining both SCM and the CMLS module, where SCM can adaptively learn and enhance structured contextual information by exploring the differences between different parts and the CMLS can provide multi-level supervised information to strengthen the contextual feature learning.

Finally, we explore the effect of the CMLS module applied to the last stage of SCM using qualitative evaluation on our PDMB validation set, as demonstrated in Fig. 5. We observe that the network leads to more accurate localisation results by adding the CMLS module to SCM in the challenging cases such as abnormal mouse poses, occlusion and invisibility. The network without the CMLS module produces the failing predictions as shown in Fig. 5(c). For example, in the 2nd row of Fig. 5(c), the mouse's left foot was mispredicted as the tail base due to the highly deformable body. However, adding the CMLS module to SCM results in the refined predictions as

TABLE V
COMPARISONS OF RMSE AND PCK@0.2 SCORE ON THE PDMB AND DEEPLABCUT MOUSE POSE TEST SETS.

Dataset	PDMB						DeepLabCut Mouse Pose					
Methods	RMSE	Snout	Left ear	Right ear	Tail base	Mean	RMSE	Snout	Left ear	Right ear	Tail base	Mean
Newell et al. [26]	4.29	95.69	92.75	95.05	93.38	94.22	5.80	92.96	93.43	91.55	88.73	91.67
Mathis et al. [17]	3.73	96.50	98.25	97.85	93.58	96.54	3.64	93.90	93.90	85.92	98.59	93.08
Pereira et al. [23]	4.0	94.68	94.10	96.25	93.23	94.56	5.70	88.26	94.37	94.37	96.24	93.31
Graving et al. [22]	3.23	96.96	98.89	97.30	94.64	96.95	3.48	96.24	95.31	92.02	98.12	95.42
Toshev et al. [38]	5.78	83.92	83.75	89.90	63.21	80.20	7.30	73.24	64.32	72.30	79.34	72.30
Wei et al. [27]	4.71	93.41	92.05	92.60	94.69	93.19	6.19	91.08	89.67	81.22	96.71	89.67
Chu et al. [34]	4.20	96.65	94.35	97.0	94.64	95.66	5.54	91.08	92.49	91.55	96.71	92.96
Sun et al. [28]	3.91	96.50	97.0	97.25	95.83	96.65	4.66	90.61	91.08	94.84	98.12	93.66
Artacho et al. [37]	4.37	95.38	97.25	97.10	93.13	95.72	3.87	95.31	94.84	94.84	95.77	95.19
Ours(SCM)	3.46	96.91	97.60	98.20	95.19	96.97	3.67	95.77	96.24	90.61	98.12	95.19
Ours(CMLS)	4.79	96.45	95.50	97.30	93.13	95.60	3.52	94.84	97.65	95.77	97.65	96.48
Ours(GM-SCENet)	2.84	97.87	98.65	98.80	95.89	97.80	3.13	96.71	98.60	97.65	98.59	97.89
Ours(GM-SCENet+MLKA)	2.79	97.72	98.90	99.10	96.34	98.01	3.10	97.65	98.12	98.12	98.59	98.12

shown in Fig. 5(e). It is also noticeable that the improved network produces higher responses in the ambiguous regions of the heatmaps, as depicted in Fig. 5(d), compared to the 5th column of Fig. 5(b). This difference is because the last CMLS module can preserve the spatial correlations between different parts, where the predicted heatmaps in SCM can be used as prior knowledge to combine multi-level features from the baseline model.

E. Comparison with State-of-the-art Pose Estimation Networks

In this section, we report the RMSE and PCK scores of our proposed and the other state-of-the-art methods with the threshold of 0.2 on our PDMB and DeepLabCut Mouse Pose test sets, as shown in Table V. Among them, there are three state-of-the-art approaches of animal pose estimation and six networks of human pose estimation.

On the PDMB test set, our proposed GM-SCENet achieves the state of the art RMSE of 2.84 pixels and the mean PCK score of 97.80%. Compared with the Stacked DenseNet for animal pose estimation [22], our GM-SCENet improves the mean PCK score from 96.95% to 97.80%, and the performance can be further improved to 98.10% by adding the MLKA inference algorithm. In particular, for the most challenging mouse part, i.e., tail base, our GM-SCENet achieves 1.25% improvement. The PCK scores for snout, left and right ears are also improved to 97.87%, 98.65% and 98.80%, respectively. Among these popular networks for human pose estimation, the model of [38] has the worst performance. This is because it directly predicts the numerical value of each coordinate rather than the heatmap of each part by a CNN. Also, their predictions ignore the spatial relationships between keypoints in an image. Compared with the 8-stack Hourglass network of [26], our GM-SCENet (3-stack Hourglass network based) outperforms it by a large margin, where the mean PCK score holds 3.58% improvement. The HRNet [28] achieves the best performance among the six human pose estimation methods. Nevertheless, we observe that our SCM based network can achieve competitive performance in comparison with HRNet and other

approaches, demonstrating the effectiveness of our proposed SCM for structured context enhancement. The performance can be further improved by adding the CMLS module. Figs. 6(a) and S1 show the mean PCK curves of different methods and the PCK curves for each keypoint, respectively. Our method (GM-SCENet+MLKA) achieves the best performance with different thresholds among all the methods. The final results demonstrate the superior performance of our proposed network over the other state-of-the-art methods in terms of PCK@0.2 score.

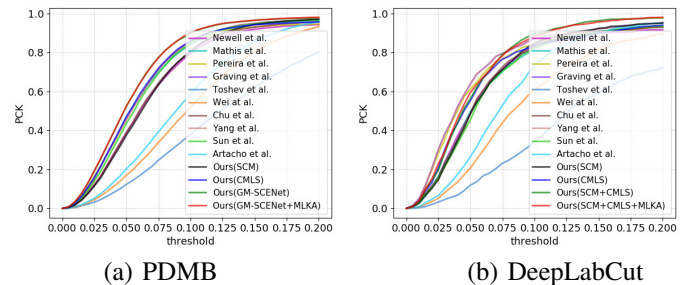


Fig. 6. Comparisons of mean PCK curves on the PDMB and DeepLabCut Mouse Pose test sets.

On the DeepLabCut Mouse Pose test set, our SCM based network achieves the mean PCK score of 95.19%, which exceeds the results of the other state-of-the-art networks except for the Stacked DenseNet. However, our CMLS based network obtains 1.06% improvement compared with the Stacked DenseNet. Furthermore, the proposed GM-SCENet improves the PCK scores with large margins of 3.29% and 2.81% on the left and right ears, compared with the closest competitors, and achieves 2.47% improvement in average, compared with the Stacked DenseNet. The use of the MLKA scheme results in further improvement on the mean PCK score and decreases on the keypoint RMSE, where our method results in an improvement of 2.93% over the best result of the methods used for human pose estimation, i.e., Unipose [37]. Figs. 6(b) and S2 show the mean PCK curves and the PCK curves of different methods for individual parts. The GM-SCENet with or without

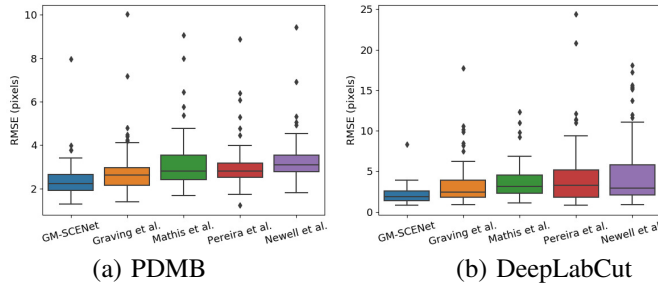


Fig. 7. Comparisons of keypoints' RMSE on the PDMB and DeepLabCut test sets. The rectangles corresponding to individual methods are used to depict the distribution of the keypoints' RMSE with the minimum and maximum values.

MLKA can achieve competitive performance, compared to the other state-of-the-art methods, on the DeepLabCut Mouse Pose test set.

It is also noteworthy that although our PDMB dataset contains more challenging mouse poses, the performance of different methods on this dataset outperforms that on the DeepLabCut Mouse Pose dataset. The reason is that our PDMB is a large database which has more images for training. For the Small DeepLabCut Mouse Pose dataset, our proposed method can achieve significant improvement on the PCK score, compared with the other state-of-the-art approaches. This shows that our method works satisfactorily on both small and relatively large datasets. Fig. S3 gives some examples of the estimated mouse poses on the PDMB and DeepLabCut test sets.

We also compare the distribution of keypoint RMSE on the PDMB test and DeepLabCut test sets (we randomly choose a group of samples rather than using the whole test set for comparison in our experiments). Here, the box plots are plotted in Fig. 7(a) and (b) to measure the effectiveness of our proposed method and other four comparison methods on the two datasets in terms of keypoints' RMSE and stability. Fig 7(a) and (b) show that our GM-SCENet sustains the lowest keypoints' RMSE with the lowest variance on the two datasets. This demonstrates that our proposed GM-SCENet can reduce the keypoints' RMSE effectively and produce more stable pose estimation results.

To further verify the generalisation capability of our proposed methods, we also conduct experiments on the Zebra, Fly and LSP datasets (Supplementary D, F). The results reported in Table S3, S4 show that our GM-SCENet can achieve competitive performance on Fly and Zebra in which the objects own a body structure similar to that of mice in comparison to the other state-of-the-art approaches. Additionally, Table S5 exhibits good generalisation of our model on human pose domain. Fig. S4 illustrates some examples of the estimated poses on the Zebra and Fly datasets.

VI. CONCLUSION

We have presented a novel end-to-end architecture called Graphical Model based Structured context Enhancement Network (GM-SCENet) for mouse pose estimation. By concen-

trating on the differences between different mouse body parts, the proposed Structured Context Mixer (SCM) based on a novel graphical model is able to adaptively learn the structured context information of individual mouse parts, including global and keypoint-specific contextual features. Meanwhile, a Cascaded Multi-level Supervision module (CMLS) has been designed to jointly train SCM and the backbone network by providing multi-level supervised information, which strengthens contextual feature learning and improves the robustness of our network. We also designed an inference strategy, i.e., Multi-level Keypoint Aggregation (MLKA), where the selected multi-level keypoints are aggregated for better prediction. Experimental results on our Parkinson's Disease Mouse Behaviour (PDMB) and the standard DeepLabCut Mouse Pose datasets demonstrate that the proposed approach outperforms several baseline methods. Additionally, our method can also be leveraged to estimate mouse pose with more keypoints and potentially extended to side-view or other datasets, especially for pose estimation of small objects with highly deformable body structures. Our future work will explore social behaviour analysis of mice with Parkinson's Disease using the proposed mouse pose estimation model.

ACKNOWLEDGMENT

The work of Huiyu Zhou is supported in part by the Royal Society-Newton Advanced Fellowship under Grant NA160342.

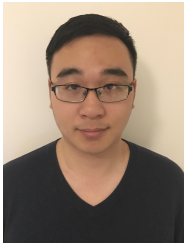
REFERENCES

- [1] L. Lewejohann, A. M. Hoppmann, P. Kegel, M. Kritzler, A. Krüger, and N. Sachser, "Behavioral phenotyping of a murine model of Alzheimer's disease in a seminaturalistic environment using RFID tracking," *Behavior Research Methods*, vol. 41, no. 3, pp. 850–856, 2009. 1
- [2] S. R. Blume, D. K. Cass, and K. Y. Tseng, "Stepping test in mice: A reliable approach in determining forelimb akinesia in MPTP-induced Parkinsonism," *Experimental Neurology*, vol. 219, no. 1, pp. 208–211, 2009. 1
- [3] A. Abbott, "Novartis reboots brain division. After years in the doldrums, research into neurological disorders is about to undergo a major change of direction," *Nature*, vol. 502, pp. 153–154, 2013. 1
- [4] Z. Jiang, D. Crookes, B. D. Green, Y. Zhao, H. Ma, L. Li, S. Zhang, D. Tao, and H. Zhou, "Context-aware mouse behavior recognition using hidden markov models," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1133–1148, 2018. 1, 9
- [5] N. G. Nguyen, D. Phan, F. R. Lumbanraja, M. R. Faisal, B. Abapihi, B. Purnama, M. K. Delimayanti, K. R. Mahmudah, M. Kubo, and K. Satou, "Applying Deep Learning Models to Mouse Behavior Recognition," *Journal of Biomedical Science and Engineering*, vol. 12, no. 02, pp. 183–196, 2019. 1
- [6] A. A. Robie, K. M. Seagraves, S. E. Egnor, and K. Branson, "Machine vision methods for analyzing social interactions," *Journal of Experimental Biology*, vol. 220, no. 1, pp. 25–34, 2017. 1
- [7] A. T. Schaefer and A. Claridge-Chang, "The surveillance state of behavioral automation," *Current Opinion in Neurobiology*, vol. 22, no. 1, pp. 170–176, 2012. 1
- [8] T. Kilpeläinen, U. H. Julku, R. Svarcbaas, and T. T. Myöhänen, "Behavioural and dopaminergic changes in double mutated human A30P*A53T alpha-synuclein transgenic mouse model of Parkinson's disease," *Scientific Reports*, vol. 9, no. 1, pp. 1–13, 2019. 1
- [9] S. R. Egnor and K. Branson, "Computational Analysis of Behavior," *Annual Review of Neuroscience*, vol. 39, no. 1, pp. 217–236, 2016. 1
- [10] M. W. Mathis and A. Mathis, "Deep learning tools for the measurement of animal behavior in neuroscience," *Current opinion in neurobiology*, vol. 60, pp. 1–11, 2020. 1

- [11] A. Arac, P. Zhao, B. H. Dobkin, S. T. Carmichael, and P. Golshani, "Deepbehavior: A deep learning toolbox for automated analysis of animal and human behavior imaging data," *Frontiers in Systems Neuroscience*, vol. 13, p. 20, 2019. 1
- [12] O. H. Maghsoudi, A. V. Tabrizi, B. Robertson, and A. Spence, "Superpixels based marker tracking vs. hue thresholding in rodent biomechanics application," *Conference Record of 51st Asilomar Conference on Signals, Systems and Computers, ACSSC 2017*, vol. 2017-Octob, pp. 209–213, 2018. 1
- [13] S. Ohayon, O. Avni, A. L. Taylor, P. Perona, and S. E. Roian Egnor, "Automated multi-day tracking of marked mice for the analysis of social behaviour," *Journal of Neuroscience Methods*, vol. 219, no. 1, pp. 10–19, 2013. 1
- [14] T. Karakostas, S. Hsiang, H. Boger, L. Middaugh, and A.-C. Granholm, "Three-dimensional rodent motion analysis and neurodegenerative disorders," *Journal of neuroscience methods*, vol. 231, pp. 31–37, 2014. 1
- [15] X. P. Burgos-artizzu, P. Doll, D. Lin, D. J. Anderson, and P. Perona, "Social behavior recognition in continuous video Supplementary Material," *2012 IEEE Conference on Computer Vision and Pattern Recognition*, vol. 5, no. 75, pp. 1322–1329, 2012. 1
- [16] F. De Chaumont, R. D. S. Coura, P. Serreau, A. Cressant, J. Chabout, S. Granon, and J. C. Olivo-Marin, "Computerized video analysis of social interactions in mice," *Nature Methods*, vol. 9, no. 4, pp. 410–417, 2012. 1, 2
- [17] 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*, vol. 21, no. 9, pp. 1281–1289, 2018. 1, 2, 7, 9, 12
- [18] K. Branson and S. Belongie, "Tracking multiple mouse contours (without too many samples)," *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, vol. 1, pp. 1039–1046, 2005. 1, 2
- [19] J. Cao, H. Tang, H.-S. Fang, X. Shen, C. Lu, and Y.-W. Tai, "Cross-domain adaptation for animal pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9498–9507. 1
- [20] X. Liu, S.-y. Yu, N. Flierman, S. Loyola, M. Kamermans, T. M. Hoogland, and C. I. D. Zeeuw, "OptiFlex: video-based animal pose estimation using deep learning enhanced by optical flow," *bioRxiv*, p. 2020.04.04.025494, 2020. 1
- [21] T. Zhang, L. Liu, K. Zhao, A. Wiliem, G. Hemson, and B. Lovell, "Omni-supervised joint detection and pose estimation for wild animals," *Pattern Recognition Letters*, vol. 132, pp. 84–90, 2020. 1
- [22] 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*, vol. 8, pp. 1–42, 2019. 1, 2, 3, 4, 8, 12
- [23] T. D. Pereira, D. E. Aldarondo, L. Willmore, M. Kislin, S. S. Wang, M. Murthy, and J. W. Shaevitz, "Fast animal pose estimation using deep neural networks," *Nature Methods*, vol. 16, no. 1, pp. 117–125, jan 2019. 1, 2, 8, 12
- [24] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepcut: A deeper, stronger, and faster multi-person pose estimation model," in *European Conference on Computer Vision*. Springer, 2016, pp. 34–50. 1, 3
- [25] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017. 1, 3
- [26] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*. Springer, 2016, pp. 483–499. 1, 2, 3, 4, 7, 12
- [27] S.-E. Wei, V. Ramakrishna, T. Kanada, and Y. Sheikh, "Convolutional Pose Machines," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2, 12
- [28] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 5686–5696, 2019. 2, 4, 7, 8, 12
- [29] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *CVPR 2011*. IEEE, 2011, pp. 1385–1392. 2
- [30] X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *Advances in neural information processing systems*, 2014, pp. 1736–1744. 2
- [31] D. Zhang, G. Guo, D. Huang, and J. Han, "Poseflow: A deep motion representation for understanding human behaviors in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6762–6770. 2
- [32] L. Dong, X. Chen, R. Wang, Q. Zhang, and E. Izquierdo, "Adore: An adaptive holons representation framework for human pose estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2803–2813, 2017. 2
- [33] S. Liu, Y. Li, and G. Hua, "Human pose estimation in video via structured space learning and halfway temporal evaluation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 7, pp. 2029–2038, 2018. 2
- [34] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1831–1840. 2, 3, 4, 5, 8, 12
- [35] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 713–728. 2, 7
- [36] J. Song, L. Wang, L. Van Gool, and O. Hilliges, "Thin-slicing network: A deep structured model for pose estimation in videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4220–4229. 2
- [37] B. Artacho and A. Savakis, "Unipose: Unified human pose estimation in single images and videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7035–7044. 2, 12
- [38] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1653–1660. 2, 12
- [39] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware cnn model," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1134–1142. 3
- [40] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141. 3
- [41] J. Nie, Y. Pang, S. Zhao, J. Han, and X. Li, "Efficient selective context network for accurate object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2020. 4
- [42] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750. 4
- [43] D. Xu, W. Ouyang, X. Alameda-Pineda, E. Ricci, X. Wang, and N. Sebe, "Learning deep structured multi-scale features using attention-gated crfs for contour prediction," in *Advances in Neural Information Processing Systems*, 2017, pp. 3961–3970. 5, 7
- [44] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3917–3925.
- [45] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in neural information processing systems*, 2011, pp. 109–117. 5
- [46] X. Chu, W. Ouyang, X. Wang *et al.*, "Crf-cnn: Modeling structured information in human pose estimation," in *Advances in Neural Information Processing Systems*, 2016, pp. 316–324. 7
- [47] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 466–481. 8
- [48] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *bmvc*, vol. 2, no. 4. Citeseer, 2010, p. 5. 8
- [49] Z. Jiang, F. Zhou, A. Zhao, X. Li, L. Li, D. Tao, X. Li, and H. Zhou, "Multi-view mouse social behaviour recognition with deep graphic model," *IEEE Transactions on Image Processing*, 2021. 8
- [50] V. Jackson-Lewis and S. Przedborski, "Protocol for the mptp mouse model of parkinson's disease," *Nature Protocols*, vol. 2, no. 1, p. 141, 2007. 8



Feixiang Zhou is currently pursuing the PhD degree with the School of Informatics, University of Leicester, U.K.



Zheheng Jiang is currently a senior research associate at the School of Computing and Communications, Lancaster University, U.K.



Zhile Yang obtained his BSc in Electrical Engineering and the MSc degree in Control Engineering both from Shanghai University (SHU) in 2010 and 2013 respectively, and he then received Ph.D. degree at the School of Electrical, Electronics and Computer Science, Queen's University Belfast (QUB), UK. He is currently an associate professor in Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. His research interests focus on artificial intelligence methods and their applications on electric vehicles, smart grid and advanced manufacturing. He is the founding chair of IEEE QUB student branch and an active member of IEEE PES, CIS and SMC societies. He is the author or co-author of more than 130 articles in peer reviewed international journals and conferences, and an active reviewer for over 50 international journals.



Zhihua Liu is currently pursuing the PhD degree with the School of Informatics, University of Leicester, U.K.



Haikuan Wang received the BSc degree from the Wuhan University in 1999, the MSc degree and the PhD degree from the Shanghai University in 2005, 2011. He is now an associate professor in the School of Mechatronic Engineering and Automation, Shanghai University. His research interests include embedded instrumentation and systems, artificial intelligence and 3D machine vision.



Fang Chen is currently pursuing the PhD degree with the Department of Mathematics, University of Leicester, U.K.



Minrui Fei received his B.S. and M.S. degrees in Industrial Automation from the Shanghai University of Technology in 1984 and 1992, respectively, and his PhD degree in Control Theory and Control Engineering from Shanghai University in 1997. Since 1998, he has been a full professor at Shanghai University. He is Chairman of Embedded Instrument and System Sub-society, and Standing Director of China Instrument & Control Society; Chairman of Life System Modeling and Simulation Sub-society, Vice-chairman of Intelligent Control and Intelligent Management Sub-society, and Director of Chinese Artificial Intelligence Association, Fellow of China Simulation Federation. His research interests are in the areas of networked control systems, machine vision, artificial intelligence, intelligent control, complex system modeling, hybrid network systems, and field control systems.



Long Chen is currently pursuing the PhD degree with the School of Informatics, University of Leicester, U.K.



Ling Li is the Director of Internationalisation at the School of Computing and also the founding coordinator of Laboratory of Brain | Cognition | Computing (BC2 Lab) of the school responsible for coordinating multidisciplinary research between Computing, Sports and local NHS hospitals. She had six-year research experience at Imperial College London with a focus to understand body sensor data (EEG, EMG, ECG, eAR-sensor, and etc.). She participated in large scale projects. She also involved in projects from government and industry (i.e. Samsung GRO award). She now serves at the editorial board of Brain Informatics and the secretary of IEEE Computing Society in UK and Ireland.



Lei Tong is currently pursuing the PhD degree with the School of Informatics, University of Leicester, U.K.



Huiyu Zhou received a Bachelor of Engineering degree in Radio Technology from Huazhong University of Science and Technology of China (1990), and a Master of Science degree in Biomedical Engineering from University of Dundee of United Kingdom (2002), respectively. He was awarded a Doctor of Philosophy degree in Computer Vision from Heriot-Watt University, Edinburgh, United Kingdom (2006). Dr. Zhou currently is a full Professor at School of Informatics, University of Leicester, United Kingdom. He has published over 350 peer-reviewed papers in the field.