1 **Selfee: Self-supervised Features Extraction of animal behaviors**

2 **Yinjun Jia^{1,2,*}, Shuaishuai Li^{3,4}, Xuan Guo^{1,2}, Junqiang Hu¹, Xiao-Hong Xu^{3,4}, Wei**
3 **Zhang^{1,2,*}**

4 ¹School of Life Sciences, IDG/McGovern Institute for Brain Research, Tsinghua
5 University, Beijing, 100084, China.

6 ²Tsinghua-Peking Center for Life Sciences, Beijing, 100084, China.

7 ³Institute of Neuroscience, State Key Laboratory of Neuroscience, Chinese Academy
8 of Sciences Center for Excellence in Brain Science and Intelligence Technology,
9 Shanghai, 200031, China.

10 ⁴Shanghai Center for Brain Science and Brain-Inspired Intelligence Technology,
11 Shanghai, 200031, China.

12 *Correspondence and requests for materials should be addressed to W.Z.
13 (wei_zhang@mail.tsinghua.edu.cn) and Y.J. (jyj20@mails.tsinghua.edu.cn)

14 **ABSTRACT**

15 Fast and accurately characterizing animal behaviors is crucial for neuroscience research.

16 Deep learning models are efficiently used in laboratories for behavior analysis.

17 However, it has not been achieved to use a fully unsupervised method to extract

18 comprehensive and discriminative features directly from raw behavior video frames for

19 annotation and analysis purposes. Here, we report a self-supervised feature extraction

20 (Selfee) convolutional neural network with multiple downstream applications to

21 process video frames of animal behavior in an end-to-end way. Visualization and

22 classification of the extracted features (Meta-representations) validate that Selfee

23 processes animal behaviors in a comparable way of human understanding. We

24 demonstrate that Meta-representations can be efficiently used to detect anomalous

25 behaviors that are indiscernible to human observation and hint in-depth analysis.

26 Furthermore, time-series analyses of Meta-representations reveal the temporal

27 dynamics of animal behaviors. In conclusion, we present a self-supervised learning

28 approach to extract comprehensive and discriminative features directly from raw video

29 recordings of animal behaviors and demonstrate its potential usage for various

30 downstream applications.

31

32 INTRODUCTION

33 Extracting representative features of animal behaviors has long been an important
 34 strategy to study the relationship between genes, neural circuits, and behaviors.
 35 Traditionally, human observations and descriptions are the primary solutions for animal
 36 behavior analysis¹. Well-trained researchers would define a set of behavior patterns and
 37 compare their intensity or proportion between experimental and control groups. With
 38 the emergence and thrive of machine learning methodology, supervised learning has
 39 been assisting human annotations and achieved impressive results²⁻⁴. Nevertheless,
 40 supervised learning is limited by prior knowledge and manually assigned labels, thus
 41 could not identify behavioral features that are not annotated.

42 Other machine learning methods were then introduced to the field which were designed
 43 to extract representative features beyond human-defined labels. These methods can be
 44 generally divided into two major categories: one estimates animal postures with a group
 45 of pre-defined key points of the body parts, and the other directly transforms raw images.
 46 The former category marks representative key points of animal bodies, including limbs,
 47 joints, trunks, and/or other body parts of interest⁵⁻⁷. Those features are usually sufficient
 48 to represent animal behaviors. However, it has been demonstrated that the key points
 49 generated by pose estimation are less efficient for direct behavior classification or two-
 50 dimensional visualization^{8,9}. Sophisticated post-processing like recurrent neural
 51 networks (RNNs)⁸, non-locomotor movement decomposition¹⁰, or feature
 52 engineerings⁹ can be applied to transform the key points into higher-level

53 discriminative features. Additionally, neglected body parts could be catastrophic. For
 54 example, the position of the proboscis of a fly is commonly neglected in behavior
 55 studies^{9,11}. Still, it is crucial for feeding¹², licking behavior during courtship¹³, and
 56 hardness detection for a substrate¹⁴. Finally, best to our knowledge, there is no
 57 demonstration of these pose-estimation methods applied to multiple animals of the
 58 same color with intensive interactions. Thus, the application of pose-estimation to
 59 mating behaviors of two black mice, a broadly adopted behavior paradigm¹⁵⁻¹⁷, could be
 60 limited because labeling body parts during mice mounting is challenging even for
 61 humans (see Discussion for more details). Therefore, using these feature extraction
 62 methods requires rigorously controlled experimental settings, additional feature
 63 engineering, and considerable prior knowledge of particular behaviors.

64 In contrast, the other category transforms pixel-level information, thus retaining more
 65 details and requiring less prior knowledge. Feature extraction of images could be
 66 achieved by wavelet transforms¹⁸ or Radon transforms¹⁹ followed by principal
 67 component analysis (PCA), and these transforms can be applied to either 2D images or
 68 depth images. However, preprocessing such as segmentation and/or registration of the
 69 images is usually required to achieve spatial invariance, a task that is particularly
 70 difficult for multi-agent videos. Additionally, these methods usually use fixed
 71 transforms and could not be adapted to different behaviors. Flourished deep learning
 72 methods, especially convolutional neural networks²⁰ (CNNs), could be adaptive to
 73 extract features from diversified datasets. Also, they have been proven more potent than

74 classic computer vision algorithms like wavelet transforms²¹ and Radon transforms²²
 75 on a famous grayscale dataset MNIST, even without supervising²³. Therefore, we
 76 attempt to adopt CNNs to achieve end-to-end feature extractions animal behaviors that
 77 are comprehensive and discriminative.

78 The cutting-edge self-supervised deep learning methods aim to extract representative
 79 features for downstream missions by comparing different augmentations of the same
 80 image and/or different images²⁴⁻²⁸. Compared with previous techniques, these methods
 81 have three major advantages. Firstly, self-supervised or unsupervised methods could
 82 completely avoid human biases. Secondly, the augmentations used to create positive
 83 samples promise invariance of the neural networks to object sizes, spatial orientations,
 84 and ambient laminations so that registration or other preprocessing is not required.
 85 Finally, the networks are optimized to export similar results for positive samples and
 86 separate negative ones, such that the extracted features are inherently discriminative.
 87 Even without negative samples, the networks can utilize differential information within
 88 batches to obtain remarkable results on downstream missions like classification or
 89 image segmentation^{27,29,30}. These advances in self-supervised learning provide a
 90 promising way to analyze animal behaviors.

91 In this work, we develop Selfee (**S**elf-supervised **F**eatures **E**xtraction) that adopts
 92 cutting-edge self-supervised learning algorithms and CNNs to analyze animal
 93 behaviors. Selfee is trained on massive unlabeled behavior video frames to avoid human
 94 bias on annotating animal behaviors, and it could capture a global character of animal

behaviors even when detailed postures are hard to see, just like human observation.

During the training process, Selfee learns to project images to a low-dimensional space without being affected by shooting conditions, image translation, and rotation, where cosine distance is proper to measure the similarities of original pictures. Selfee also provides potentials for various downstream analyses. We demonstrate that the extracted features are suitable for t-SNE visualization, k-NN-based classification, k-NN-based anomaly detection, and dynamic time warping (DTW). We also show that further integrated modeling, like the autoregressive hidden Markov model (AR-HMM), is compatible with Selfee extracted Meta-representations. After downstream analyses, Selfee provides comparable results with manual annotations on fly behavior like courtship index. We apply Selfee to fruit flies, mice, and rats, three widely used model animals, and validate our results with manual annotations. Discoveries of behavioral phenotypes in mutant flies by Selfee are proven to have biological significance. The performance of Selfee on these model species indicates its potential usage for behavioral studies of non-model animals as well as other tasks. We also provide an open-source Python package and pre-trained models of flies and mice to the community (see more in Code Availability).

RESULTS

Workflow of Selfee and its downstream analyses

Selfee is trained to generate Meta-representations at the frame level, which are then analyzed at different time scales. First, grayscale videos are decomposed into single frames, and three tandem frames are stacked into a live-frame to generate a motion-colored RGB picture (Figure 1A). These live-frames preserve not only spatial information (e.g., postures of each individual or relative distances and angles between individuals) within each channel but also temporal information across different channels. Live-frames are used to train Selfee to produce comprehensive and discriminative representations at the frame level (Figure 1B). These representations can be later used in numerous applications. For example, anomalous detection on mutant animals can discover new phenotypes compared with their genetic controls (Figure 1C). Also, the AR-HMM could be applied to model the micro-dynamics of behaviors, such as the duration of states or the probabilities of state transitions¹⁸. The AR-HMM splits videos into modules and yields behavioral state usages that visualize differences between genotypes (Figure 1D). In contrast, DTW could compare the long-term dynamics of animal behaviors and capture global differences at the video level³¹ by aligning pairs of time series and calculating their similarities (Figure 1E). These three demonstrations cover different time scales from frame to video level, and other downstream analyses could also be incorporated into the workflow of Selfee. Compared with previous machine learning frameworks for animal behavior analysis,

Selfee has three major advantages. First, Selfee and the Meta-representations could be used for various tasks. The contrastive learning process of Selfee would allow output features to be appropriately compared by cosine similarity. Therefore, distance-based applications, including classification, clustering, and anomaly detection, would be easily realized. It was also reported that with some adjustment of backbones, self-supervised learning would facilitate tasks such as pose estimation³² and object segmentation^{28,33}. Those findings indicate that Selfee could be generalized, modified, and finetuned for animal pose estimation or segmentation tasks. Second, Selfee is a fully unsupervised method developed to annotate animal behaviors. Although some other techniques also adopt semi-supervised or unsupervised learning, they usually require manually labeled pre-defined key points of the images^{8,10}; some methods also require expert-defined programs for better performance⁹. Key point selection and program incorporation require a significant amount of prior knowledge and are subject to human bias. In contrast, Selfee does not need any prior knowledge. Finally, Selfee is relatively hardware-inexpensive. Training Selfee only takes eight hours on a single RTX 3090, and the inference speed could reach 800 frames per second. Selfee could accept top-view 2D greyscale video frames as inputs so that neither depth cameras¹⁸ nor fine-calibrated multi-view camera arrays¹⁰ is required. Therefore, Selfee can be trained and used with routinely collected behavior videos on ordinary desktop workstations, warranting its accessibility by biology laboratories.

Siamese convolutional neural networks capture discriminative representations of animal posture.

Selfee contains a pair of Siamese CNNs trained to generate discriminative representations for live-frames. ResNet-50³⁴ is chosen as the backbone whose classifier layer is replaced by a three-layer multi-layer perceptron (MLP). These MLPs are called projectors which yield final representations during the inference stage. There are two branches in Selfee. The main branch is equipped with an additional predictor, while the reference branch is a copy of the main branch (the SimSiam style²⁹). Both branches contain group discriminators after projectors and perform dimension reduction on extracted features for online clustering (Figure 2B).

During the training stage, batches of live-frames are randomly transformed twice and fed into the main branch and reference branch, respectively. Augmentations applied to live-frames include crop, rotation, flip, and application of the Turbo lookup table³⁵ followed by color jitters (Figure 2A, Figure 2—figure supplement 1). The reference branch yields a representation of received frames, while the main branch predicts the outcome of the reference branch. At the same time, they both produce clustering results of the current batch. The main branch is optimized for similar predictions and clustering results as the reference branch, and the reference branch will not receive gradient information to prevent mode collapse^{27,29} (Figure 2C). In this way, Selfee is trained to be invariant to those transforms and focus on critical information to yield discriminative representations.

After the training stage, we evaluated the performance of Selfee with t-SNE visualization and k-NN classification. To investigate whether our model captures human-interpretable features, we manually labeled one clip of *Drosophila* courtship video and visualized those representations with t-SNE dimension reduction. On the t-SNE map, human-annotated courtship behaviors, including chasing, wing extension, copulation attempt, copulation, and non-interactive behaviors (“others”), separated from each other distinctively (Figure 2D).

Meta-representations can also be used for behavior classification. We manually labeled seven 10,000-frame videos (around five minutes each) as a pilot dataset. A weighed k-NN classifier was then constructed as previously reported²⁴. Seven-fold cross-validation was performed on the dataset with the k-NN classifier, which achieved a mean F₁ score of 72.4% and achieved a similar classification result as human annotations (Figure 2E, F). The classifier had the worst recall score on wing extension behaviors (67% recall), likely because of the ambiguous intermediate states between chasing and wing extension (Figure 2—figure supplement 2A). The precisions also showed that this k-NN classifier tended to have strict criteria with wing extension and copulation and relatively loose criteria with chasing and copulation attempts (Figure 2—figure supplement 2B). It was reported that independent human experts could only reach agreements on around 70% of wing extension frames³⁶, comparable to the performance of our k-NN classifier.

We then asked whether Selfee can be generalized to analyze behaviors of other species.

We finetuned fly video pre-trained Selfee with mice mating behavior data. The mating behavior of mice can be defined mainly into five categories³⁷, including social interest, mounting, intromission, ejaculation, and others (see Methods for detailed definitions). With t-SNE visualization, we found that five types of behaviors could be separated by Selfee, although mounting behaviors were rare and not concentrated (Figure 2G). We then used eight human-annotated videos to test the k-NN classification performance of Selfee-generated features. We achieved an F_1 score of 59.0% (Figure 2—figure supplement 3). Mounting, intromission, and ejaculation share similar static characteristics but are different in temporal dynamics. Therefore, we asked if more temporal information would assist the classification. Using the LightGBM classifier, we achieved a much higher classification performance by incorporating slide moving average and standard division of 81-frame time windows, the main frequencies, and their energy within 81-frame time windows. The average F_1 score of eight-fold cross-validation could reach 67.4%, and the classification results of the ensembled classifier (see Methods) were closed to human observations (Figure 2H, I). Nevertheless, it was still difficult to distinguish between mounting, intromission, and ejaculation because mounting and ejaculation are much rarer than social body contact or intromission. Selfee is more robust than the vanilla SimSiam networks when applied to the behavioral data. Behavioral data often suffer from catastrophic imbalance. For example, copulation attempts are around six-fold rarer than wing extension during fly courtship (Figure 2—figure supplement 5A). Therefore, we added group discriminators to vanilla SimSiam

networks which were reported to fight against the long-tail effect proficiently³⁸. Aside from overcoming the long-tail effect, we also found group discriminators helpful for preventing mode collapse during ablation studies (Figure 2—figure supplement 5B, C, D, and Supplementary Table 1). Additionally, the convergence can be easily reached on grayscale images of similar objects (two flies), by which CNNs may not be well trained to extract good representations. Applying the Turbo lookup table on grayscale frames brought more complexity and made color jitters more powerful on grayscale images. Selfee would capture more useful features with this Turbo augmentation (Figure 2—figure supplement 5E, F, and Figure 2—figure supplement 6).

Anomaly detection at the frame level identifies rare behaviors at the sub-second time scale.

The representations produced by Selfee could be directly used for anomaly detection without further post-processing. During the training step, Selfee learns to compare Meta-representations of frames with cosine distance which is also used for anomaly detection. When given two groups of videos, namely the query group and the reference group, the anomaly score of each live-frame in the query group is calculated by two steps (Figure 3A). First, distances between the query live-frame and all reference live-frames are measured, and the k-nearest distance is referred to as its inter-group score (IES). Without further specification, k equals 1 in all anomaly detections in this work. Some false positives occurred when only the IES was used as the anomaly score (Figure

3—figure supplement 1A). The reason could be that two flies in a chamber could be in mathematically infinite relative positions and form a vast event space. However, each group usually only contains several videos, and each video is only recorded for several minutes. For some rare postures, even though the probability of observing them is similar in both the query and reference group, they might only occur in the query group but not in the reference group. Therefore, an intra-group score (IAS) is introduced in the second step to eliminate these false-positive effects. We assume that those rare events should not be sampled frequently in the query groups either. Thus, the IAS is defined as the k-nearest distance of the query frame against all other frames within its group, except those within the time window of ± 50 frames (Figure 3—figure supplement 1B). The final anomaly score is defined as the IES minus the IAS.

To test whether our methods could detect anomalous behavior in real-world data, we performed genetic screenings within fifteen neurotransmitter-related mutant alleles or neuron-silenced lines (with UAS-Kir2.1³⁹) (Figure 3B). Their male-male interaction videos were inferred by Selfee trained on male-female courtship videos. Since we aimed to find interactions distinct from male-male courtship behaviors, a baseline of ppk23>Kir2.1 flies was established because this line exhibit strong male-male courtship behaviors⁴⁰. We compared the top-100 anomaly scores from sets of videos from experimental groups and wild-type control flies. The results revealed that one line, CCHa2-R-RB>Kir2.1, showed a significantly high anomaly score. By manually going through all anomalous live-frames, we further identified its phenotype as a brief tussle

behavior mixed with copulation attempts (Figure 3C, Video 1, 0.2x play speed). This behavior was ultra-fast and lasts for less than a quarter second (Figure 3D), making it difficult to be detected by human observers. Up to this point, we have demonstrated that the frame-level anomaly detection could capture sub-second behavior episode that human observers tend to neglect.

Selfee also revealed that *Trh* knock-out flies had an unusual close body contact during the screening. *Trh* is the crucial enzyme for serotonin biosynthesis, and its mutant flies showed a statistically significantly higher anomaly score (Figure 3B) than the wild-type control. Selfee identified 60 frames of abnormal behaviors within 42,000 input frames, occupying less than 0.15% of the total recording time. By manually going through all these frames, we concluded most of them as short-range body interactions (Figure 3E and Video 2, 0.2x play speed), and these social interactions could last for around half to one second on average (Figure 3F). Despite that serotonin signals were well-studied for controlling aggression behavior in flies⁴¹, to the best of our knowledge, the close body contact of flies and serotonergic neurons' role in this behavior has not been reported yet. A possible reason is that this behavior has no unique posture compared with other behaviors, like wing extension, and this behavior is too scarce to be noticed by human experts.

To further ask whether these close body contacts have biological significance, we performed corresponded behavior assays on mutant flies. Based on the fact that the *Trh* mutant male flies have a higher tolerance to body touch, we hypothesized that they

would have a decreased defensive behavior. As previously reported, fruit flies show robust defensive behavior to mechanical stimuli on their wings^{42,43}. Decapitated flies would kick with their hind legs when a thin probe stimulates their wings. This stimulation mimics the invasion of parasitic mites and could be used to test its defensive behavior. Our results showed that *Trh* knock-out flies had a significantly lower kicking rate than control flies (Figure 3G), indicating a reduction of self-defensive intensity. Next, we performed social behavior assay^{44,45} on the mutant flies because the close body contact can also be explained by reduced social repulsion. We measured the nearest distance, median distance, and average distance of each male flies in a forty-individual group placed in a vertical triangular chamber. By comparing median values of these distances of each replication, *Trh* knock-out flies kept significantly shorter distances from others than the control group (Figure 3H, I). The probability density function of their median distances also showed that knock-out flies had a closer social distance than control flies (Figure 3J). Therefore, we concluded that *Trh* knock-out flies had reduced social repulsion. Taken together, Selfee is capable of discovering novel features of animal behaviors with biological relevance when a proper baseline is defined.

Modeling motion structure of *Drosophila* courtship behaviors.

Animal behaviors have long-term structures beyond single-frame postures. The duration and proportions of each bout and transition probabilities of different behaviors

have been proven to have biological significance^{18,46}. To better understand those long-term characteristics, we introduce AR-HMM and DTW analyses to model the temporal structure of *Drosophila* courtship behavior. AR-HMM is a powerful method to analyze stereotyped behavioral data^{18,47,48}. It discovers modules of behaviors and describes the modules with auto-regressive matrixes. The transition probability of each state is defined by the transition matrix of the HMM (Figure 4A). In this way, AR-HMM could capture local structures of animal behaviors as well as syntaxes.

We asked if we could detect the dynamic changes of courtship behaviors of male flies by disturbing their chemosensation. *Ir76b* is an extensively studied (co)receptor that is known to mediate female pheromones detection⁴⁹⁻⁵². We used an AR-HMM model with ten modules (No.1 to 10) to analyze the courtship of *Ir76b* mutant flies and their control group and focused on state usages. PCA of state usages revealed an apparent difference between mutant flies and control flies (Figure 4B). Module No.6 showed a statistically significant difference among ten discovered modules (Figure 4C). By manually going through all the frames of module No.6, we found that it mainly contained non-interactive behaviors with minor contaminations of courtship behaviors (Video 3, 1x play speed). To validate this result, we compared it with human annotations. Although this module did not cover all non-interactive behaviors that human experts would label, they showed a similar trend between the experimental and control group (Figure 4D). We also performed AR-HMM analysis with a much larger module number. The PCA result was also distinct, and the previous module No.6 was split into five smaller

modules (No.2, 15, 24, 32, 34) containing non-interactive behaviors (Figure 4—figure supplement 1, Video 4-8, 1x play speed). This tuning indicated that AR-HMM analysis is robust regardless of the number of modules, same as a previous report¹⁸. Our results indicated that *Ir76b* mutation might affect male flies' detection of female pheromones and consequentially the temporal structure of their courtship behaviors. These findings prove that Selfee with AR-HMM could discover the differences in proportions of behaviors, similar to what was achieved with classic manual analysis such as the courtship index.

The AR-HMM modeling does not necessarily capture the difference of long-term dynamics intuitively, such as the latency of certain behaviors. To solve this problem, we introduce DTW analysis. DTW is a well-known algorithm to align time series, which returns the best-matched path and the matching similarity (Figure 4E). The alignment can be simplified as follows. When given the same start state and end state, it optimally maps all indices from the query series to the reference series monotonically. Pairs of mapped indices form a path to visualize the dynamic difference. The points upper than the diagonal line indicate that the current time point in the query group is matched to a future time point in the reference group so that the query group has faster dynamics and vice versa. In our experiments, cosine similarities of Selfee extracted representations are used to calculate warping paths.

Previously, DTW was widely applied to numerical measures of animal behaviors, including trajectory⁵³, audios⁵⁴, and acceleration⁵⁵. For the first time, we applied DTW

to image data, with the aid of Selfee, to study the prolonged dynamic of animal behaviors. We analyzed whether the vision is essential for a male fly's copulation completion. Visual cues are essential for male flies to locate female flies during courtship⁵⁶, and mutant flies of *NorpA*, which have defective visual transduction⁵⁷, have a prolonged courtship latency in our experiments (Figure 4F), similar to previously findings⁵⁸. When wild-type flies were used as the reference for the DTW, the group of *NorpA* mutant flies yielded a curve lower than the diagonal line, indicating a delay of their courtship behaviors (Figure 4G). In this way, our experiments confirm that Selfee and DTW could capture differences in long-term dynamics such as behavior latency. In conclusion, DTW and AR-HMM could capture temporal differences between control and experimental groups beyond single-frame postures, making Selfee a competent unsupervised method for traditional analyses like courtship index or copulation latency.

DISCUSSION

Here we use cutting-edge self-supervised learning methods and convolutional neural networks to extract Meta-representations from animal behavior videos. Siamese CNNs have proven their capability to learn comprehensive representations²⁹. The cosine similarity, part of its loss function used for training, is rational and well-suited to measure similarities between the raw images. Besides, convolutional neural networks are trained end-to-end so that preprocessing steps like segmentation or key points extraction is unnecessary. By incorporating Selfee with different post-processing

methods, we can identify phenotypes of animal behaviors at different time scales. In the current work, we demonstrate that the extracted representations could be used not only for straightforward distance-based analyses such as t-SNE visualization or k-NN anomaly detection but also for sophisticated post-processing methods like AR-HMM. These validations confirm that the extracted Meta-representations are meaningful and valuable.

By applying our method to mice mating behavior, we show that our Selfee outperformed some of the widely used pose-estimation methods in multi-animal behavior analysis. The famous DeepLabCut and similar methods could identify human-defined key points on animals. However, when animals of the same color are recorded at a compromised resolution and their body contacts are intensive, the current version of DeepLabCut could hardly extract useful features (Figure 1—figure supplement 1, Video 9). The reason is that it is extremely difficult to unambiguously label body parts like nose, ears and hips when two mice are close enough, a task challenging even for human experts. By contrast, Selfee could readily identify the frame as “intromission” (Figure 1—figure supplement 1) as human experts would do. These results show that our methods could capture global characteristics of behaviors like human experts, making it well-suited for processing multi-animal behavior videos, compared with pose-estimation methods.

We also demonstrate that the cutting-edge self-supervised learning model is accessible to biology labs. Our model can be trained on only one RTX 3090 GPU with a batch size

of 256 within only 8 hours with the help of the newly proposed CLD loss function³⁸ and other improvements (see Methods for further details). Furthermore, when the model pre-trained with mice videos was applied to rat behaviors, we were able to achieve a zero-shot classification of five major types of social behaviors (Figure 2—figure supplement 4). Although the F₁ score was only 49.6%, it still captured the major differences between similar behaviors, such as allogrooming and social nose contact. Thus, we have demonstrated that self-supervised learning could be easily achieved with limited computation resources and a much shorter time and could be transferred to datasets that share similar visual characteristics.

Despite those advantages, there are some limitations of Selfee. First, because each live-frame only contains three raw frames, our model could not capture much information on the animal motion. It becomes more evident when Selfee is applied to highly dynamic behaviors such as mice mating behaviors. This can be overcome by increasing the computation because commonly used 3D convolution⁵⁹ or spatial-temporal attention⁶⁰ is good at dynamic information extraction but requires much more computational resources. Second, as previously reported, CNNs are highly vulnerable to image texture⁶¹. We observed that certain types of beddings of the behavior chamber could profoundly affect the performance of our neural networks (Figure 1—figure supplement 2), so in some cases, background removal is necessary (see Methods for further details). Lastly, Selfee could only use discriminative features within each batch, without any negative samples provided, so minor irrelevant differences could be

amplified and cause inconsistent results (named mode-split). This mode split may increase variations of downstream analyses.

We can envision at least two possible future directions for Selfee. One is to optimize the backbone of neural networks to extract better features. Advanced self-supervised learning methods like DINO³³ (with visual transformers, ViTs) could separate objects from the background and extract more explainable representations. Besides, by using ViTs, the neural network could be more robust against distractive textures⁶². At the same time, more temporal information can also be incorporated for a better understanding of motions. Combining these two, equipping ViTs with spatial-temporal attention could capture more temporal information.

Another direction will be explainable behavior forecasting for a deeper understanding of animal behaviors. For a long time, behavior forecasting has been a field with extensive investigations in which RNNs, LSTMs, or transformers are usually applied^{9,60,63}. However, most of these works use coordinates of key points as inputs. Therefore, the trained model might predominantly focus on spatial movement information and discover fewer behavioral syntaxes. By representation learning, spatial information is essentially condensed so that more syntaxes might be highlighted. Transformer models for forecasting could capture correlations between sub-series as well as long-term trends like seasonality⁶⁴. These deep learning methods would provide behavioral neuroscientists powerful tools to identify behavior motifs and syntaxes that organize stereotyped motifs beyond the Markov property.

428 **ACKNOWLEDGEMENTS**

429 We thank members of the Zhang lab for discussions. This work was supported by grants
 430 31871059 and 32022029 from the National Natural Science Foundation of China, grant
 431 Z181100001518001 from the Beijing Municipal Science & Technology Commission,
 432 and a ‘Brain+X’ seed grant from the IDG/McGovern Institute for Brain Research at
 433 Tsinghua to W.Z.. W.Z. is supported by Chinese Institute for Brain Research, Beijing.
 434 W.Z. is an awardee of the Young Thousand Talent Program of China.

435 **AUTHOR CONTRIBUTIONS**

436 Y.J. and J.H. coded the Selfee neural network and other accessory parts. Y.J., X.G. and
 437 S.L. performed animal experiments and analyzed data. W.Z. and X.X. supervised the
 438 project. Y.J. and W.Z. wrote the manuscript. All authors discussed and commented on
 439 the manuscript.

440 **DECLARATION OF INTERESTS**

441 The authors declare no competing interests.

DATA AND CODE AVAILABILITY STATEMENT

Major data used in this study were uploaded to Dryad. Data could be accessed via:

[https://datadryad.org/stash/share/BnIoOnaweOn2fc-sllSO0FhJJqduXQYaNu-](https://datadryad.org/stash/share/BnIoOnaweOn2fc-sllSO0FhJJqduXQYaNu-KuPgZ394)

[KuPgZ394](https://datadryad.org/stash/share/BnIoOnaweOn2fc-sllSO0FhJJqduXQYaNu-KuPgZ394) or its DOI 10.5061/dryad.br15dvb8. We also shared our pretrained weights

with Google Drive:

<https://drive.google.com/file/d/1A3U5guNEKA3Bi9H3QnfstZDEZ6aesqcR/view?usp=sharing>.

With the uploaded dataset and pretrained weights, our experiments could be replicated. However, due to its huge size and the limited internet service resources, we are currently not able to share our full training dataset. The full dataset is as large as 400GB, which is hard to upload to a public server and will be difficult for others users to download.

For training dataset, it would be available from the corresponding author upon reasonable request (wei_zhang@mail.tsinghua.edu.cn), and then we can discuss how to transfer such a big dataset. No project proposal is needed as long as the dataset is not used for any commercial purpose.

Our Python scripts could be accessed on GitHub: <https://github.com/EBGU/Selfee>

Other software used in our project include ImageJ(<https://imagej.net/software/fiji/>) and GraphPad Prism(<https://www.graphpad.com/>).

All data used to plot graphs and charts in the manuscript can be fully accessed on Dryad (DOI 10.5061/dryad.br15dvb8).

462 REFERENCES

- 463 1 Hall, J. The mating of a fly. *Science* **264**, 1702-1714,
464 doi:10.1126/science.8209251 (1994).
- 465 2 Kabra, M., Robie, A. A., Rivera-Alba, M., Branson, S. & Branson, K. JAABA:
466 interactive machine learning for automatic annotation of animal behavior.
467 *Nature Methods* **10**, 64-67, doi:10.1038/nmeth.2281 (2013).
- 468 3 Jiang, Z., Chazot, P. L., Celebi, M. E., Crookes, D. & Jiang, R. Social Behavioral
469 Phenotyping of Drosophila With a 2D–3D Hybrid CNN Framework. *IEEE*
470 *Access* **7**, 67972-67982, doi:10.1109/ACCESS.2019.2917000 (2019).
- 471 4 Segalin, C. *et al.* The Mouse Action Recognition System (MARS): a software
472 pipeline for automated analysis of social behaviors in mice. *bioRxiv*,
473 2020.2007.2026.222299, doi:10.1101/2020.07.26.222299 (2020).
- 474 5 Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body
475 parts with deep learning. *Nature Neuroscience* **21**, 1281-1289,
476 doi:10.1038/s41593-018-0209-y (2018).
- 477 6 Günel, S. *et al.* DeepFly3D, a deep learning-based approach for 3D limb and
478 appendage tracking in tethered, adult Drosophila. *eLife* **8**, e48571,
479 doi:10.7554/eLife.48571 (2019).
- 480 7 Graving, J. M. *et al.* DeepPoseKit, a software toolkit for fast and robust animal
481 pose estimation using deep learning. *eLife* **8**, e47994, doi:10.7554/eLife.47994
482 (2019).
- 483 8 Luxem, K., Fuhrmann, F., Kürsch, J., Remy, S. & Bauer, P. Identifying
484 Behavioral Structure from Deep Variational Embeddings of Animal Motion.
485 *bioRxiv*, 2020.2005.2014.095430, doi:10.1101/2020.05.14.095430 (2020).
- 486 9 Sun, J. J. *et al.* in *2021 IEEE/CVF Conference on Computer Vision and Pattern*
487 *Recognition (CVPR)*. 2876-2885 (2021).
- 488 10 Huang, K. *et al.* A hierarchical 3D-motion learning framework for animal
489 spontaneous behavior mapping. *Nature Communications* **12**, 2784,
490 doi:10.1038/s41467-021-22970-y (2021).
- 491 11 Calhoun, A. J., Pillow, J. W. & Murthy, M. Unsupervised identification of the
492 internal states that shape natural behavior. *Nature Neuroscience* **22**, 2040-2049,
493 doi:10.1038/s41593-019-0533-x (2019).
- 494 12 Zhou, Y., Cao, L.-H., Sui, X.-W., Guo, X.-Q. & Luo, D.-G. Mechanosensory
495 circuits coordinate two opposing motor actions in *Drosophila* feeding. *Science*
496 *Advances* **5**, eaaw5141, doi:10.1126/sciadv.aaw5141 (2019).
- 497 13 Mezzera, C. *et al.* Ovipositor Extrusion Promotes the Transition from Courtship
498 to Copulation and Signals Female Acceptance in *Drosophila melanogaster*.
499 *Current Biology* **30**, 3736-3748.e3735,
500 doi:<https://doi.org/10.1016/j.cub.2020.06.071> (2020).
- 501 14 Zhang, L. *et al.* Parallel Mechanosensory Pathways Direct Oviposition
502 Decision-Making in *Drosophila*. *Current Biology* **30**, 3075-3088.e3074,

doi:<https://doi.org/10.1016/j.cub.2020.05.076> (2020).

15 Bayless, D. W. *et al.* Limbic Neurons Shape Sex Recognition and Social Behavior in Sexually Naive Males. *Cell* **176**, 1190-1205.e1120, doi:10.1016/j.cell.2018.12.041 (2019).

16 Zhang, S. X. *et al.* Hypothalamic dopamine neurons motivate mating through persistent cAMP signalling. *Nature* **597**, 245-249, doi:10.1038/s41586-021-03845-0 (2021).

17 Wei, Y.-C. *et al.* Medial preoptic area in mice is capable of mediating sexually dimorphic behaviors regardless of gender. *Nature Communications* **9**, 279, doi:10.1038/s41467-017-02648-0 (2018).

18 Wiltchko, Alexander B. *et al.* Mapping Sub-Second Structure in Mouse Behavior. *Neuron* **88**, 1121-1135, doi:<https://doi.org/10.1016/j.neuron.2015.11.031> (2015).

19 Berman, G. J., Choi, D. M., Bialek, W. & Shaevitz, J. Mapping the stereotyped behaviour of freely moving fruit flies. *Journal of the Royal Society Interface* **11** (2014).

20 Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278-2324, doi:10.1109/5.726791 (1998).

21 Romero, D., Ruedin, A. M. C. & Seijas, L. in *ICIAP*.

22 Aradhya, V. N. M., Kumar, G. H. & Noushath, S. in *2007 International Conference on Signal Processing, Communications and Networking*. 626-629.

23 Ji, X., Vedaldi, A. & Henriques, J. F. Invariant Information Clustering for Unsupervised Image Classification and Segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9864-9873 (2019).

24 Wu, Z., Xiong, Y., Yu, S. X. & Lin, D. Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination. *ArXiv abs/1805.01978* (2018).

25 Caron, M. *et al.* Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *ArXiv abs/2006.09882* (2020).

26 Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. E. A Simple Framework for Contrastive Learning of Visual Representations. *ArXiv abs/2002.05709* (2020).

27 Grill, J.-B. *et al.* Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. *ArXiv abs/2006.07733* (2020).

28 He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. B. Momentum Contrast for Unsupervised Visual Representation Learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9726-9735 (2020).

29 Chen, X. & He, K. Exploring Simple Siamese Representation Learning. *ArXiv abs/2011.10566* (2020).

30 Zbontar, J., Jing, L., Misra, I., LeCun, Y. & Deny, S. in *ICML*.

31 Myers, C., Rabiner, L. & Rosenberg, A. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **28**, 623-635,

- doi:10.1109/TASSP.1980.1163491 (1980).
- 32 Dahiya, A., Spurr, A. & Hilliges, O. in *NeurIPS 2020 Workshop on Pre-registration in Machine Learning* Vol. 148 (eds Bertinetto Luca *et al.*) 255--271 (PMLR, Proceedings of Machine Learning Research, 2021).
- 33 Caron, M. *et al.* Emerging Properties in Self-Supervised Vision Transformers. *ArXiv* **abs/2104.14294** (2021).
- 34 He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778 (2016).
- 35 Mikhailov, A. *Turbo, An Improved Rainbow Colormap for Visualization*, <<https://ai.googleblog.com/2019/08/turbo-improved-rainbow-colormap-for.html>> (2019).
- 36 Leng, X., Wohl, M., Ishii, K., Nayak, P. & Asahina, K. Quantifying influence of human choice on the automated detection of *Drosophila* behavior by a supervised machine learning algorithm. *PLoS One* **15**, e0241696, doi:10.1371/journal.pone.0241696 (2020).
- 37 McGill, T. E. Sexual Behavior in Three Inbred Strains of Mice. *Behaviour* **19**, 341-350 (1962).
- 38 Wang, X., Liu, Z. & Yu, S. X. Unsupervised Feature Learning by Cross-Level Instance-Group Discrimination. *arXiv: Computer Vision and Pattern Recognition* (2020).
- 39 Paradis, S., Sweeney, S. T. & Davis, G. W. Homeostatic Control of Presynaptic Release Is Triggered by Postsynaptic Membrane Depolarization. *Neuron* **30**, 737-749, doi:[https://doi.org/10.1016/S0896-6273\(01\)00326-9](https://doi.org/10.1016/S0896-6273(01)00326-9) (2001).
- 40 Thistle, R., Cameron, P., Ghorayshi, A., Dennison, L. & Scott, K. Contact Chemoreceptors Mediate Male-Male Repulsion and Male-Female Attraction during *Drosophila* Courtship. *Cell* **149**, 1140-1151, doi:<https://doi.org/10.1016/j.cell.2012.03.045> (2012).
- 41 Alekseyenko, Olga V. *et al.* Single Serotonergic Neurons that Modulate Aggression in *Drosophila*. *Current Biology* **24**, 2700-2707, doi:<https://doi.org/10.1016/j.cub.2014.09.051> (2014).
- 42 Li, J. *et al.* A Defensive Kicking Behavior in Response to Mechanical Stimuli Mediated by *Drosophila* Wing Margin Bristles. *The Journal of Neuroscience* **36**, 11275-11282, doi:10.1523/jneurosci.1416-16.2016 (2016).
- 43 Liu, C. *et al.* A neural circuit encoding mating states tunes defensive behavior in *Drosophila*. *Nature Communications* **11**, 3962, doi:10.1038/s41467-020-17771-8 (2020).
- 44 Simon, A. F. *et al.* A simple assay to study social behavior in *Drosophila*: measurement of social space within a group1. *Genes, Brain and Behavior* **11**, 243-252, doi:<https://doi.org/10.1111/j.1601-183X.2011.00740.x> (2012).
- 45 Au - McNeil, A. R. *et al.* Conditions Affecting Social Space in *Drosophila melanogaster*. *JoVE*, e53242, doi:doi:10.3791/53242 (2015).

- 587 46 Mueller, J. M., Ravbar, P., Simpson, J. H. & Carlson, J. M. *Drosophila*
588 *melanogaster* grooming possesses syntax with distinct rules at different
589 temporal scales. *PLOS Computational Biology* **15**, e1007105,
590 doi:10.1371/journal.pcbi.1007105 (2019).
- 591 47 Wiltchko, A. B. *et al.* Revealing the structure of pharmacobehavioral space
592 through motion sequencing. *Nature Neuroscience* **23**, 1433-1443,
593 doi:10.1038/s41593-020-00706-3 (2020).
- 594 48 Rudolph, S. *et al.* Cerebellum-Specific Deletion of the GABAA Receptor δ
595 Subunit Leads to Sex-Specific Disruption of Behavior. *Cell Reports* **33**, 108338,
596 doi:<https://doi.org/10.1016/j.celrep.2020.108338> (2020).
- 597 49 Luo, R. *et al.* Molecular basis and homeostatic regulation of Zinc taste.
598 *Protein&Cell*, doi:10.1007/s13238-021-00845-8 (2021).
- 599 50 Sánchez-Alcañiz, J. A. *et al.* An expression atlas of variant ionotropic glutamate
600 receptors identifies a molecular basis of carbonation sensing. *Nature*
601 *Communications* **9**, 4252, doi:10.1038/s41467-018-06453-1 (2018).
- 602 51 Ganguly, A. *et al.* A Molecular and Cellular Context-Dependent Role for Ir76b
603 in Detection of Amino Acid Taste. *Cell Reports* **18**, 737-750,
604 doi:<https://doi.org/10.1016/j.celrep.2016.12.071> (2017).
- 605 52 Zhang, Y. V., Ni, J. & Montell, C. The Molecular Basis for Attractive Salt-Taste
606 Coding in *Drosophila*. *Science* **340**, 1334-1338,
607 doi:10.1126/science.1234133 (2013).
- 608 53 Cleasby, I. R. *et al.* Using time-series similarity measures to compare animal
609 movement trajectories in ecology. *Behavioral Ecology and Sociobiology* **73**,
610 151, doi:10.1007/s00265-019-2761-1 (2019).
- 611 54 Kohlsdorf, D., Herzing, D. & Starner, T. Methods for discovering models of
612 behavior: A case study with wild Atlantic spotted dolphins. *Animal Behavior*
613 *and Cognition* **3**, 265-287 (2016).
- 614 55 Aurasopon, A. Dynamic Time Warping for classifying cattle behaviors and
615 reducing acceleration data size. *Agricultural Engineering International: The*
616 *CIGR Journal* **18**, 293-300 (2016).
- 617 56 Ribeiro, I. M. A. *et al.* Visual Projection Neurons Mediating Directed Courtship
618 in *Drosophila*. *Cell* **174**, 607-621.e618, doi:10.1016/j.cell.2018.06.020 (2018).
- 619 57 Bloomquist, B. T. *et al.* Isolation of a putative phospholipase c gene of
620 *drosophila*, *norpA*, and its role in phototransduction. *Cell* **54**, 723-733,
621 doi:[https://doi.org/10.1016/S0092-8674\(88\)80017-5](https://doi.org/10.1016/S0092-8674(88)80017-5) (1988).
- 622 58 Markow, T. A. & Manning, M. Mating success of photoreceptor mutants of
623 *Drosophila melanogaster*. *Behavioral and Neural Biology* **29**, 276-280,
624 doi:[https://doi.org/10.1016/S0163-1047\(80\)90612-3](https://doi.org/10.1016/S0163-1047(80)90612-3) (1980).
- 625 59 Ji, S., Xu, W., Yang, M. & Yu, K. 3D Convolutional Neural Networks for Human
626 Action Recognition. *IEEE Transactions on Pattern Analysis and Machine*
627 *Intelligence* **35**, 221-231, doi:10.1109/TPAMI.2012.59 (2013).
- 628 60 Aksan, E., Cao, P., Kaufmann, M. & Hilliges, O. Attention, please: A Spatio-

629 temporal Transformer for 3D Human Motion Prediction. *ArXiv* **abs/2004.08692**
630 (2020).
631 61 Geirhos, R. *et al.* ImageNet-trained CNNs are biased towards texture;
632 increasing shape bias improves accuracy and robustness. *ArXiv* **abs/1811.12231**
633 (2019).
634 62 Naseer, M. *et al.* Intriguing Properties of Vision Transformers. *ArXiv*
635 **abs/2105.10497** (2021).
636 63 Fragkiadaki, K., Levine, S. & Malik, J. Recurrent Network Models for
637 Kinematic Tracking. *ArXiv* **abs/1508.00271** (2015).
638 64 Wu, H., Xu, J., Wang, J. & Long, M. Autoformer: Decomposition Transformers
639 with Auto-Correlation for Long-Term Series Forecasting. *ArXiv*
640 **abs/2106.13008** (2021).

641

FIGURE LEGENDS

Figure 1 | The framework of Selfee and its downstream applications.

(A) One live-frame is composed of three tandem frames in R, G, and B channels, respectively.
 (B) Live-frames are used to train Selfee, which adopts a backbone of ResNet-50.
 (C, D, and E) Representations produced by Selfee could be used for anomaly detection that could identify unusual animal postures in the query video compared with the reference videos (C) AR-HMM (autoregressive hidden Markov model) that models the local temporal characteristics of behaviors and clusters frames into modules (states) and calculates stages usages of different genotypes (D) DTW (dynamic time warping) that aligns behavior videos to reveal differences of long-term dynamics (E) and other potential tasks including behavior classification, forecasting, or even image segmentation and pose estimation after properly modifying and finetuning of the neural networks.

Figure 2 | The network structure of Selfee and validation of Selfee with human annotations.

(A) The architecture of Selfee networks. Each live-frame is randomly transformed twice before being fed into Selfee.
 (B) Selfee adopts a SimSiam-style network structure with additional group discriminators. Loss 1 is canonical negative cosine loss, and loss 2 is the newly proposed CLD loss.
 (C) The reference branch of Selfee is not involved in backward propagation.
 (D) Visualization of fly courtship live-frames with t-SNE dimension reduction. Each dot was colored based on human annotations.
 (E) The confusion matrix of the k-NN classifier for fly courtship behavior, normalized by the numbers of each behavior in the ground truth. The average F_1 score of the nine-fold cross-validation was 72.4%, and mAP was 75.8%.
 (F) A visualized comparison of labels produced by the k-NN classifier and human annotations of fly courtship behaviors.
 (G) Visualization of live-frames of mice mating behaviors with t-SNE dimension reduction. Each dot is colored based on human annotations.
 (H) The confusion matrix of the LightGBM classifier for mice mating behaviors, normalized by the numbers of each behavior in the ground truth. For the LightGBM classifier, the average F_1 score of the nine-fold cross-validation was 67.4%, and mAP was 69.1%.

(I) A visualized comparison of labels produced by the LightGBM classifier and human annotations of mice mating behaviors.

Figure 3 | Anomalous posture detection using Selfee-produced features.

(A) The calculation process of anomaly scores. Each query frame is compared with every reference frame, and the nearest distance was named IES (the thickness of lines indicates distances). Each query frame is also compared with every query frame, and the nearest distance is called IAS. The final anomaly score of each frame equals IES minus IAS.

(B) Genetic screening of fifteen fly lines with mutations in neurotransmitter genes or with specific neurons silenced ($n \geq 6$ for each genotype). RA is short for CCHa2-R-RA, and RB is short for CCHa2-R-RB. $CCHa2-R-RB^{Gal4} > Kir2.1$, $q < 0.0001$; Trh^{Gal4} , $q = 0.0432$; one-way ANOVA with Benjamini and Hochberg correction.

(C) Examples of mixed tussles and copulation attempts identified in $CCHa2-R-RB^{Gal4} > Kir2.1$ flies.

(D) The temporal dynamic of anomaly scores during the mixed behavior, centralized at 1.67 s. SEM is indicated with the light color region.

(E) Examples of close body contact behaviors identified in Trh^{Gal4} flies.

(F) The cosine similarity between the center frame of the close body contact behaviors (1.67s) and their local frames. SEM is indicated with the light color region.

(G) The kicking index of Trh^{Gal4} flies ($n=30$) was significantly lower than w^{1118} flies ($n=27$), $p = 0.0034$, Mann Whitney test.

(H) Examples of social aggregation behaviors of Trh^{Gal4} flies and w^{1118} flies.

(I) Social distances of Trh^{Gal4} flies ($n=6$) and w^{1118} flies ($n=6$). Trh^{Gal4} flies had much closer social distances with each other compared with w^{1118} flies; nearest, $p = 0.0043$; median, $p = 0.002$; average, $p = 0.0087$; all Mann Whitney test.

(J) Distributions of the median social distance of Trh^{Gal4} flies and w^{1118} flies. Distributions were calculated within each replication. Average distributions are indicated with solid lines, and SEMs are indicated with light color regions.

Figure 4 | Time-series analyses using Selfee produced features.

(A) A brief illustration of the AR-HMM model. The local autoregressive property is determined by β_t , the autoregressive matrix, which is yield based on the current hidden state of the HMM. The transition between each hidden state is described by the transition matrix (p_{ij}).

(B) PCA visualization of state usages of wild-type flies ($n=7$) and $Ir76b^l$ mutant flies ($n=7$).

(C) State usages of ten modules. Module No. 6 showed significantly different usages in

wild-type and mutant flies; $q = 0.022$, Mann Whitney test with Benjamini and Hochberg correction.

(D) Module No.6 showed similar statistic results as manually labeled non-interactive behaviors. Module No.6, $p = 0.0006$; manually labeled non-interactive behaviors, $p = 0.0006$; Both were Mann Whitney test.

(E) A brief illustration of the DTW model. The transformation from a rounded rectangle to an ellipse could contain six steps (grey reference shapes). The query transformation lags at step 2 but surpasses at step 4. The dynamic is visualized on the right panel.

(F) *Norpa*³⁶ flies ($n=6$) showed a significantly longer copulation latency than wild-type flies ($n=7$), $p = 0.0495$, Mann Whitney test.

(G) *Norpa*³⁶ flies had delayed courtship dynamics than wild-type flies with DTW visualization.

Figure 1—figure supplement 1 | A comparison between Selfee and DeepLabCut on animals of the same color.

(A) Selfee is more robust when applied to intensive social behaviors of animals of the same color, like mating behaviors of two black mice. The image is from the intromission behavior which could be identified by Selfee equipped with the trained LightGBM classifier. In contrast, a trained DeepLabCut model identified it as a single mouse.

Figure 1—figure supplement 2 | Beddings and backgrounds that affect training and inference of Selfee.

(A) Textures on the damped filter paper would mislead Selfee to output features similar with copulation but not wing extension (ground truth).

(B) Background inconsistency would affect the training process when Selfee was applied to mice behavior data. Therefore, backgrounds were removed from all frames to avoid potential defects.

Figure 2—figure supplement 1 | Different augmentations used for Selfee training.

(A) Visualization of each augmentation. Detailed descriptions of each augmentation could be found in Methods and source codes.

Figure 2—figure supplement 2 | Difficulties on fly courtship behavior classification.

(A) Some wing extension frames are hard to distinguish from chasing behaviors. Blue indicators point at slightly extended wings.

(B) The confusion matrix of the k-NN classifier, normalized by the numbers of each behavior in inferred labels.

Figure 2—figure supplement 3 | Classification of mice mating behaviors with

Selfee extracted features.

(A) For the k-NN classifier, the average F₁ score of the nine-fold cross-validation was 59.0%, and mAP was 53.0%. The confusion matrix of the k-NN classifier, normalized by the numbers of each behavior in the ground truth.

(B) The confusion matrix of the k-NN classifier, normalized by the numbers of each behavior in inferred labels.

(C) The confusion matrix of the LightGBM classifier, normalized by the numbers of each behavior in inferred labels. The LightGBM classifier had a much better performance compared with the k-NN classifier.

Figure 2—figure supplement 4 | k-NN classification of rat behaviors with Selfee

trained on mice datasets.

(A) The average F₁ score of the nine-fold cross-validation was 49.6%, and mAP was 46.6%. The confusion matrix of the k-NN classifier, normalized by the numbers of each behavior in the ground truth.

(B) The confusion matrix of the k-NN classifier, normalized by the numbers of each behavior in inferred labels.

Figure 2—figure supplement 5 | Ablation test of Selfee training process on fly

datasets.

(A) The distribution of different behaviors in wild-type flies courtship videos.

(B) Visualization of the same live-frames as Figure 2D with t-SNE dimension reduction. Used representations were extracted by models trained without CLD loss. Each dot is colored based on human annotations. The legend is shared with panel A.

(C) The confusion matrix of the k-NN classifier, normalized by the numbers of each

behavior in the ground truth. Used representations were extracted by models trained without CLD loss.

(D) Collapse levels during the training process. Without CLD loss, Selfee suffered from catastrophic mode collapse. Details for collapse level calculation could be found in Methods.

(E) Visualization of the same live-frames as Figure 2D with t-SNE dimension reduction. Used representations were extracted by models trained without Turbo transformation. Each dot is colored based on human annotations. The legend is shared with panel A.

(F) The confusion matrix of the k-NN classifier, normalized by the numbers of each behavior in the ground truth. Used representations were extracted by models trained without Turbo transformation.

Figure 3—figure supplement 1 | Using intra-group score (IAS) to eliminate false-positive results in anomaly detections.

(A) Anomaly scores without IAS of wild-type male-male interactions with the same genotype as references. The blue region indicates the max anomaly score when using IAS; blue dots indicate anomaly scores without IAS that fall into the blue region; red dots indicate false-positive anomaly scores.

(B) The cosine similarity between the center frame of wild-type courtship behaviors (1.67s) and their local frames. SEM is indicated with the light color region. Seven videos containing 70,000 frames were split into non-overlapping 100-frame fragments for calculations. Beyond ± 50 frames, the cosine similarity dropped to a much lower level, not affecting anomaly detection.

Figure 4—figure supplement 1 | AR-HMM produced features with 40 modules using Selfee.

(A) PCA visualization of state usages of wild-type flies and *Ir76b¹* mutant flies.

(B) State usages of forty modules. Module No. 2, 15, 24, 32, 34 showed significantly different usages in wild-type and mutant flies; $q = 0.029, 0.029, 0.049, 0.049, 0.029$ respectively, Mann Whitney test with Benjamini and Hochberg correction.

(C) The collection of modules No. 2, 15, 24, 32, 34 showed similar statistic results as manually labeled non-interactive behaviors. AR-HMM module collection, $p = 0.0006$; manually labeled non-interactive behaviors, $p = 0.0006$; all Mann Whitney test.

815 MATERIALS AND METHODS

816 Fly stocks

817 All fly strains were maintained under a 12 h/12 h light/dark cycle at 25°C and 60%
 818 humidity (PERCIVAL incubator). The following fly lines were acquired from
 819 Bloomington *Drosophila* Stock Center: *CCHa1^{attP}* (84458), *CCHa1-R^{attP}* (84459),
 820 *CCHa2^{attP}* (84460), *CCHa2-R^{attP}* (84461), *CCHa2-R-RA^{Gal4}* (84603), *CCHa2-R-*
 821 *RB^{Gal4}* (84604), *CNMa^{attP}* (84485), *Oamb^{attP}* (84555), *Dop2R^{KO}* (84720), *DopEcR^{Gal4}*
 822 (84717), *SerT^{attP}* (84572), *Trh^{Gal4}* (86146), *TK^{attP}* (84579), *Ir76b^l* (51309), *Norpa³⁶*
 823 (9048), UAS-Kir2.1 (6596). *Tdc2^{RO54}* was a gift from Dr. Yufeng Pan at Southeast
 824 University, China. Taotie-Gal4 was a gift from Dr. Yan Zhu at Institute of Biophysics,
 825 Chinese Academy of Sciences, China.

827 Fly courtship behavior and male-male interaction

828 Virgin female flies were raised for 4~6 days in fifteen-fly groups, and naïve male flies
 829 were kept in isolated vials for 8~12 days. All behavioral experiments were done under
 830 25 °C and 45%~50% humidity. Flies were transferred into a customized chamber of
 831 3 mm height and 10 mm diameter by a homemade aspirator. Fly behaviors were
 832 recorded using a stereoscopic microscope-mounted with a CCD camera (Basler ORBIS
 833 OY-A622f-DC) at the resolution of 1000×500 (for two chambers at the same time), or
 834 640×480 (for individual chambers) and a frame rate of 30 Hz. Five types of behaviors
 835 were annotated manually, including “chasing” (a male fly follows a female fly), “wing
 836 extension” (a male fly extends unilateral wing and orientates to the female to sing
 837 courtship son, “copulation attempt” (a male fly bends its abdomen toward the genitalia
 838 of the female or the unstable state that male fly mounts on a female with its wings open),
 839 and “copulation” (male fly mounts on a female in a stable posture for several minutes).
 840

841 Fly defensive behavior assay

842 The kicking behavior was tested based on previously reported paradigms^{1,2}. Briefly,
 843 flies were raised in groups for 3~5 days. Flies were anesthetized on ice, and then male
 844 flies were decapitalized and transferred to 35 mm Petri dishes with damped filter paper
 845 on the bottom to keep the moisture. Flies were allowed to recover for around 30 minutes
 846 in the dishes. The probe for stimulation was homemade from a heat-melt yellow pipette
 847 tip, and the probe's tip was 0.3 mm. Each side of flies' wing margin was gently touched
 848 5 times, and the kicking behavior was recorded manually. The statistical analysis was

performed with the Mann Whitney test with GraphPad Prism Software.

Social behavior assay for flies

The social distance was tested based on the previously reported method³. Briefly, flies were raised in groups for 3 days. Flies were anesthetized paralyzed on ice, and male flies were picked and transferred to new vials (around 40 flies per vial). Flies were allowed to recover for one day. The vertical triangular chambers were cleaned with 75% ethanol and dried with paper towels. After assembly, flies were transferred into the chambers by a homemade aspirator. The photos were taken after 20 min, and the positions of each fly were manually marked in ImageJ. The social distances were measured with the lateral sides of the chambers (16.72 cm) as references, and the median values of the nearest, median, and average distance of each replication are calculated. The statistical analysis was performed with the Mann Whitney test in GraphPad Prism Software.

Mice mating behavior assay

Wild-type mice of C57BL/6J were purchased from Slac Laboratory Animal (Shanghai). Adult (8-24 weeks old) male mice were used for sexual behavior analysis. All animals were housed under a reversed 12 h/12 h light-dark cycle with water and food *ad libitum* in the animal facility at the Institute of Neuroscience, Shanghai, China. All experiments were approved by the Animal Care and Use Committee of the Institute of Neuroscience, Chinese Academy of Sciences, Shanghai, China (IACUC No. NA-016-2016). Male mice were singly housed for at least 3 days prior to sexual behavioral tests. All tests were initiated at least 1 hr after lights were switched off. Behavioral assays were recorded using infrared cameras at the frame rate of 30 Hz. Female mice were surgically ovariectomized and supplemented with hormones to induce receptivity. Hormones were suspended in sterile sunflower Selfee oil (Sigma-Aldrich, S5007) and injected 10 mg (in 50 mL oil) and 5 mg (in 50 mL oil) of 17 β -estradiol benzoate (Sigma-Aldrich, E8875) 48 h and 24 h preceding the test, respectively. On the day of the test, 50 mg of progesterone (Sigma-Aldrich, P0130; in 50 mL oil) was injected 4–6 h prior to the test. Male animals were adapted 10 min to behavioral testing rooms where a recording chamber equipped with video acquisition systems was located. A hormonal primed ovariectomized C57BL/6J female (OVX) was introduced to the home cage of male mice and videotaped for 30 min. Mating behavior tests were repeated three times with different OVX at least three days apart. Videos were manually scored using a custom-written MATLAB program. The following criteria were used for behavioral annotation: active nose contacts initiated by male mouse towards the female's genitals, body area, faces were defined collectively as “social interest”; male mouse climbs the back of the female and moves the pelvis were defined as “mount”; Rhythmic pelvic movements

after mount were defined as “intromission”; a body rigidity posture after final deep thrust were defined as “ejaculation”.

Data preprocessing, augmentation and sampling

Fly behavior videos were decomposed into frames by FFmpeg, and only the first 10,000 frames of each video were preserved and resized into images with a resolution of 224×224. For model training of *Drosophila* courtship behavior, each video was manually checked to ensure successful copulations within 10,000 frames.

Mice behavior videos were decomposed into frames by FFmpeg, and only frames of the first 30 min of each video were preserved. Frames were then preprocessed with OpenCV⁴ in Python. Behavior chambers in each video were manually marked, segmented, and resized into images of a resolution of 256×192. For background removal, the average frame of each video was subtracted from each frame, and noises were removed by a threshold of 25 and the median filter with a kernel size of 5. Finally, the contrast was adjusted with histogram equalization.

For data augmentations, crop, rotation, flip, Turbo, and color jitter were applied. For a given frame, it formed a live-frame with its preceding and succeeding frames. For flies’ behavior video, three frames were successive, and for mice, the preceding or succeeding frame is one frame away from the current frame due to their slower dynamics⁵. Each live-frame was randomly cropped into a smaller version containing more than 49% (70%×70%) of the original image; then the image was randomly (clockwise or anticlockwise) rotated for an angle smaller than the acute angle formed by the diagonal line and the vertical line, then the image would be vertically flipped, horizontally flipped, and/or applied the Turbo lookup table⁶ at the probability of 50%, respectively; and finally, the brightness, contrast, saturation, and hue were randomly adjusted within 10% variation. Notably, since the Turbo transformation is designed for grayscale images, for a motion-colored RGB image, each channel was transformed individually. After Turbo transformation, their corresponded channels were composited to form a new image.

For fly data sampling, all images of all videos were randomly ranked, and each batch contained 256 images from different videos. For mice data sampling, all images of each video were randomly ranked, and each batch contained 256 images from the same video. This strategy was designed to eliminate the inconsistency of recording conditions of mice that was more severe than flies.

Selfie neural network and its training

All training and inference were accomplished on a workstation with 128GB RAM, AMD Ryzen 7 5800x, and one NVIDIA GeForce RTX 3090. Selfie neural network was constructed based on publications and source codes of BYOL⁷, SimSiam⁸, and CLD⁹

with PyTorch¹⁰. In brief, the last layer of ResNet-50 was removed, and a 3-layer 2048-dimension MLP was added as the projector. Hidden layers of the projector were followed by batch normalization (BN) and ReLU activation, and the output layer only had BN. The predictor was constructed with a 2-layer bottleneck MLP with a 512-dimension hidden layer and a 2048-dimension output layer. The hidden layer but not the output layer of the predictor had BN and ReLU. As for the group discriminator for CLD loss, it had only one normalized fully connected layer that projected 2048-dimension output to 1024 dimensions, followed by a customized normalization layer that was described in the paper of CLD⁹.

The loss function of Selfee had two major parts. The first part was the negative cosine loss^{7,8}, and the second part was the CLD loss⁹. For a batch of n samples, Z , P , V represented the output of projector, predictor, and group discriminator of the main branch, respectively; Z' , P' , V' represented the output of the reference branch; and sg as the stop-gradient operator. After k-means clustering of V , the centroids of k classes were given by M , and labels of each sample were provided in the one-hot form as L . The hyperparameter θ was 0.07, and λ was 2. The loss function was given by the following equations:

$$\text{CosineDistance}(m, n) = 1 - \frac{m}{\|m\|_2} \cdot \frac{n}{\|n\|_2}$$

$$\text{Loss}_1 = \frac{1}{2n} \sum_{i=1}^n \text{CosineDistance}(sg(z_i), p'_i) + \frac{1}{2n} \sum_{i=1}^n \text{CosineDistance}(sg(z'_i), p_i)$$

$$\text{CrossEntropyLoss}(x, l) = -x \cdot l + \log(\|\exp(x)\|_1)$$

$$\text{Loss}_2 = \frac{1}{2} \text{CrossEntropyLoss}\left(\frac{v_i}{\theta} \cdot M'^T, l_i\right) + \frac{1}{2} \text{CrossEntropyLoss}\left(\frac{v'_i}{\theta} \cdot M^T, l'_i\right)$$

$$\text{Loss} = \text{Loss}_1 + \lambda \text{Loss}_2$$

For all training processes, the Selfee network was trained for 20,000 steps with the SDG optimizer with a momentum of 0.9 and a weight decay of 1e-4. The learning rate was adjusted in the one-cycle learning rate policy¹¹ with base learning rates and a pct start of 0.025. The model for *Drosophila* courtship behavior was initialized with ResNet-50 pre-trained on the ImageNet, and the base learning rate was 0.025 per batch size of 256. As for the mating behaviors of mice, the model was initialized with weights trained on the fly dataset, and the base learning rate was 0.05 per batch size of 256.

t-SNE visualization

Video frames for t-SNE visualization were all processed by Selfee. Embeddings of three tandem frames were averaged to eliminate potential noises. All embeddings were transformed using t-SNE provided in the scikit-learn¹² package in Python without further turning of parameters. Results were visualized with the Matplotlib¹³ package in Python, and their colors were assigned based on human annotations of video frames.

Classification

Two kinds of classification methods were implemented, including the k-NN classifier and the LightGBM (Light Gradient Boosting Machine) classifier. The weighed k-NN classifier was constructed based on the previous reports^{8,14}. LightGBM classifier¹⁵ was provided by its official package in Python. The F₁ score and mAP were calculated with the scikit-learn¹² package in Python.

For fly behavior classification, seven 10,000-frame videos were annotated manually. Seven-fold cross-validation was performed using embeddings generated by Selfee and the k-NN classifier. Inferred labels were forced to be continuous through time by using inferred labels of 21 neighbor frames to determine the final result. Then, a video independent of the cross-validation was annotated and inferred by a k-NN classifier using all 70,000 samples, and the last 3,000 frames were used for the raster plot.

For rat behavior classification, the RatSI dataset¹⁶ (a kind gift from Noldus Information Technology bv) contains nine manually annotated videos. We neglected three rarest annotated behaviors: moving away, nape attacking, and pinning, and we combined approaching and following into a larger category. Therefore, we used five kinds of behaviors, including allogrooming, approaching or following, social nose contact, solitary, and others. Nine-fold cross-validation was performed using embeddings generated by Selfee and the k-NN classifier. Inferred labels were forced to be continuous through time by using inferred labels of 81 neighbor frames to determine the final result.

For mice behavior classification, eight videos were annotated manually. Eight-fold cross-validation was performed using embeddings generated by Selfee and the k-NN classifier. To incorporate more temporal information, the LightGBM classifier and additional features were also used. Additional features include slide moving average and standard division of 81-frame time windows, the main frequencies, and their energy (using short-time Fourier transform in SciPy¹⁷) within 81-frame time windows. Early-stop was used to prevent over-fitting. Inferred labels were forced to be continuous through time by using inferred labels of 81 neighbor frames to determine the final result. Then, a video independent of the cross-validation was annotated and inferred by an ensemble classifier of eight previously constructed classifiers, and all frames were used for the raster plot.

Anomaly detection

For a group of query embeddings of sequential frames $q_1, q_2, q_3, \dots, q_n$, and a group of reference embeddings of sequential frames $r_1, r_2, r_3, \dots, r_m$, the anomaly score of each query frame was given by the following equation:

$$\text{AnomalyScore}(q_i) = \min_{j=1}^m (\text{CosineDistance}(q_i, r_j)) - \min_{|j-i| < 50} (\text{CosineDistance}(q_i, q_j))$$

A PyTorch implementation of cosine similarity¹⁸ was used for accelerated calculations. The anomaly score of each video was the average anomaly score of the top 100 anomalous frames. The statistical analysis of the genetic screening was performed with one-way ANOVA with Benjamini and Hochberg correction in GraphPad Prism Software.

If negative controls are provided, anomalous frames are defined as frames with higher anomaly scores than the maximum anomaly score of frames in negative control videos.

Autoregressive hidden Markov model (AR-HMM)

All AR-HMM models were built with the implementation of MoSeq⁵ (<https://github.com/mattjj/pyhsmm-autoregressive>). A principal component analysis (PCA) model that could explain 95% of variance of the control group was built and used to transform both control and experiment groups. The max module number was set as 10 for all experiments unless indicated otherwise. Each model was sampled for 1000 iterations. We kept other hyperparameters the same as the examples provided by this package. State usages of each module in control and experimental groups were analyzed by Mann Whitney test with SciPy¹⁷ followed with Benjamini and Hochberg correction. The state usages were also visualized after PCA dimensional reduction with scikit-learn¹² and Matplotlib¹³.

Dynamic time warping (DTW)

Dynamic time warping was modified from the Python implementation¹⁹ (<https://dynamictimewarping.github.io/python/>). Specifically, PyTorch implementation of cosine similarity¹⁸ was used for accelerated calculations.

Pose-estimation with DeepLabCut

We used the official implementation of DeepLabCut^{20,21}. For training, 120 frames of a mating behavior video were labeled manually, and 85% of them were used as the training set. Marked body parts included nose, ears, body center, hips, and bottom, following previous publications^{22,23}. The model (ResNet-50 as the backbone) was trained for 100,000 iterations, with a batch size of 16. We kept other hyperparameters the same as default settings.

REFERENCES

- 1 Li, J. *et al.* A Defensive Kicking Behavior in Response to Mechanical Stimuli Mediated by *Drosophila* Wing Margin Bristles. *The Journal of Neuroscience* **36**, 11275-11282, doi:10.1523/jneurosci.1416-16.2016 (2016).
- 2 Liu, C. *et al.* A neural circuit encoding mating states tunes defensive behavior in *Drosophila*. *Nature Communications* **11**, 3962, doi:10.1038/s41467-020-17771-8 (2020).
- 3 Au - McNeil, A. R. *et al.* Conditions Affecting Social Space in *Drosophila melanogaster*. *JoVE*, e53242, doi:doi:10.3791/53242 (2015).
- 4 Bradski, G. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.
- 5 Wiltshko, Alexander B. *et al.* Mapping Sub-Second Structure in Mouse Behavior. *Neuron* **88**, 1121-1135, doi:<https://doi.org/10.1016/j.neuron.2015.11.031> (2015).
- 6 Mikhailov, A. *Turbo, An Improved Rainbow Colormap for Visualization*, <<https://ai.googleblog.com/2019/08/turbo-improved-rainbow-colormap-for.html>> (2019).
- 7 Grill, J.-B. *et al.* Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. *ArXiv abs/2006.07733* (2020).
- 8 Chen, X. & He, K. Exploring Simple Siamese Representation Learning. *ArXiv abs/2011.10566* (2020).
- 9 Wang, X., Liu, Z. & Yu, S. X. Unsupervised Feature Learning by Cross-Level Instance-Group Discrimination. *arXiv: Computer Vision and Pattern Recognition* (2020).
- 10 Paszke, A. *et al.* in *NeurIPS*.
- 11 Smith, L. N. & Topin, N. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. *arXiv:1708.07120* (2017). <<https://ui.adsabs.harvard.edu/abs/2017arXiv170807120S>>.
- 12 Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825-2830 (2011).
- 13 Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science &*

1072 *Engineering* **9**, 90-95, doi:10.1109/MCSE.2007.55 (2007).

1073 14 Wu, Z., Xiong, Y., Yu, S. X. & Lin, D. Unsupervised Feature Learning via Non-
1074 Parametric Instance-level Discrimination. *ArXiv* **abs/1805.01978** (2018).

1075 15 Ke, G. *et al.* in *Proceedings of the 31st International Conference on Neural*
1076 *Information Processing Systems* 3149–3157 (Curran Associates Inc., Long
1077 Beach, California, USA, 2017).

1078 16 Lorbach, M. *et al.* Learning to recognize rat social behavior: Novel dataset and
1079 cross-dataset application. *Journal of Neuroscience Methods* **300**, 166-172,
1080 doi:<https://doi.org/10.1016/j.jneumeth.2017.05.006> (2018).

1081 17 Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in
1082 Python. *Nature Methods* **17**, 261-272, doi:10.1038/s41592-019-0686-2 (2020).

1083 18 Musgrave, K., Belongie, S. J. & Lim, S.-N. PyTorch Metric Learning. *ArXiv*
1084 **abs/2008.09164** (2020).

1085 19 Toni, G. Computing and Visualizing Dynamic Time Warping Alignments in R:
1086 The dtw Package. *Journal of Statistical Software* **31**, doi:10.18637/jss.v031.i07
1087 (2009).

1088 20 Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body
1089 parts with deep learning. *Nature Neuroscience* **21**, 1281-1289,
1090 doi:10.1038/s41593-018-0209-y (2018).

1091 21 Lauer, J. *et al.* Multi-animal pose estimation and tracking with DeepLabCut.
1092 *bioRxiv*, 2021.2004.2030.442096, doi:10.1101/2021.04.30.442096 (2021).

1093 22 Segalin, C. *et al.* The Mouse Action Recognition System (MARS): a software
1094 pipeline for automated analysis of social behaviors in mice. *bioRxiv*,
1095 2020.2007.2026.222299, doi:10.1101/2020.07.26.222299 (2020).

1096 23 Sun, J. J. *et al.* in *2021 IEEE/CVF Conference on Computer Vision and Pattern*
1097 *Recognition (CVPR)*. 2876-2885 (2021).

1098

FIGURES AND TABLES

Figure 1

Figure 2

Figure 3

1109 **Figure 4**

1110

1111

1112

1113 **Figure 1—figure supplement 1**

A

Ground Truth

Intramission

DeepLabCut

Selfee with LightGBM

Intramission

1114

1115 **Figure 1—figure supplement 2**

A

Ground Truth

Wing extension

Wing extension

k-NN classification

Wing extension

Copulation

B

Original

Background

Mice

1116

1117

Figure 2—figure supplement 1

Figure 2—figure supplement 2

Figure 2—figure supplement 3

Figure 2—figure supplement 4

Figure 2—figure supplement 5

1129 **Figure 2—figure supplement 6 | Ablation test of Selfee training process on fly datasets.**

1130

Model	Pretrained ResNet-50 with random projectors		Selfee		Selfee without CLD loss		Selfee without Turbo transformation	
Evaluation	Mean F1 Score	Mean AP	Mean F1 Score	Mean AP	Mean F1 Score	Mean AP	Mean F1 Score	Mean AP
Replication 1	0.586	0.580	0.724	0.758	0.227	0.227	0.604	0.550
Replication 2	0.597	0.570	0.676	0.683	0.163	0.200	0.574	0.551
Replication 3	0.596	0.586	0.714	0.754	0.172	0.214	0.517	0.497
Best	0.597	0.586	0.724	0.758	0.227	0.227	0.604	0.551

1131

Figure 3—figure supplement 1

Figure 4—figure supplement 1