



HHS Public Access

Author manuscript

Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. Author manuscript; available in PMC 2023 January 09.

Published in final edited form as:

Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2022 June ; 2022: 2161–2170. doi:10.1109/cvpr52688.2022.00221.

Self-Supervised Keypoint Discovery in Behavioral Videos

Jennifer J. Sun^{1,*}, Serim Ryou^{1,*†}, Roni H. Goldshmid¹, Brandon Weissbourd¹, John O. Dabiri¹, David J. Anderson¹, Ann Kennedy², Yisong Yue^{1,3}, Pietro Perona¹

¹Caltech

²Northwestern University

³Argo AI

Abstract

We propose a method for learning the posture and structure of agents from unlabelled behavioral videos. Starting from the observation that behaving agents are generally the main sources of movement in behavioral videos, our method, Behavioral Keypoint Discovery (B-KinD), uses an encoder-decoder architecture with a geometric bottleneck to reconstruct the spatiotemporal difference between video frames. By focusing only on regions of movement, our approach works directly on input videos without requiring manual annotations. Experiments on a variety of agent types (mouse, fly, human, jellyfish, and trees) demonstrate the generality of our approach and reveal that our discovered keypoints represent semantically meaningful body parts, which achieve state-of-the-art performance on keypoint regression among self-supervised methods. Additionally, B-KinD achieve comparable performance to supervised keypoints on downstream tasks, such as behavior classification, suggesting that our method can dramatically reduce model training costs vis-a-vis supervised methods.

1. Introduction

Automatic recognition of object structure, for example in the form of keypoints and skeletons, enables models to capture the essence of the geometry and movements of objects. Such structural representations are more invariant to background, lighting, and other nuisance variables and are much lower-dimensional than raw pixel values, making them good intermediates for downstream tasks, such as behavior classification [4, 11, 15, 39, 43], video alignment [26, 44], and physics-based modeling [7, 12].

However, obtaining annotations to train supervised pose detectors can be expensive, especially for applications in behavior analysis. For example, in behavioral neuroscience [34], datasets are typically small and lab-specific, and the training of a custom supervised keypoint detector presents a significant bottleneck in terms of cost and effort. Additionally, once trained, supervised detectors often do not generalize well to new agents with different

Correspondence to jjsun@caltech.edu.

*Equal contribution.

†Current affiliation: Samsung Advanced Institute of Technology

structures without new supervision. The goal of our work is to enable keypoint discovery on new videos without manual supervision, in order to facilitate behavior analysis on novel settings and different agents.

Recent unsupervised/self-supervised methods have made great progress in keypoint discovery [20, 21, 51] (see also Section 2), but these methods are generally not designed for behavioral videos. In particular, existing methods do not address the case of multiple and/or non-centered agents, and often require inputs as cropped bounding boxes around the object of interest, which would require an additional detector module to run on real-world videos. Furthermore, these methods do not exploit relevant structural properties in behavioral videos (e.g., the camera and the background are typically stationary, as observed in many real-world behavioral datasets [5, 15, 22, 29, 34, 39]).

To address these challenges, the key to our approach is to discover keypoints based on reconstructing the *spatiotemporal difference* between video frames. Inspired by previous works based on image reconstruction [20, 37], we use an encoder-decoder setup to encode input frames into a geometric bottleneck, and train the model for reconstruction. We then use spatiotemporal difference as a novel reconstruction target for keypoint discovery, instead of single image reconstruction. Our method enables the model to focus on discovering keypoints on the behaving agents, which are generally the only source of motion in behavioral videos.

Our self-supervised approach, **Behavioral Keypoint Discovery (B-KinD)**, works without manual supervision across diverse organisms (Figure 1). Results show that our discovered keypoints achieve state-of-the-art performance on downstream tasks among other self-supervised keypoint discovery methods. We demonstrate the performance of our keypoints on behavior classification [42], keypoint regression [20], and physics-based modeling [7]. Thus, our method has the potential for transformative impact in behavior analysis: first, one may discover keypoints from behavioral videos for new settings and organisms; second, unlike methods that predict behavior directly from video, our low-dimensional keypoints are semantically meaningful so that users can directly compute behavioral features; finally, our method can be applied to videos without the need for manual annotations.

To summarize, our main contributions are:

1. **Self-supervised method for discovering keypoints** from real-world behavioral videos, based on spatiotemporal difference reconstruction.
2. **Experiments across a range of organisms** (mice, flies, human, jellyfish, and tree) demonstrating the generality of the method and showing that the discovered keypoints are semantically meaningful.
3. **Quantitative benchmarking** on downstream behavior analysis tasks showing performance that is comparable to supervised keypoints.

2. Related work

Analyzing Behavioral Videos.

Video data collected for behavioral experiments often consists of moving agents recorded from stationary cameras [1, 4, 5, 11, 15, 22, 33, 39]. These behavioral videos contain different model organisms studied by researchers, such as fruit flies [4, 11, 15, 24] and mice [5, 17, 22, 39]. From these recorded video data, there has been an increasing effort to automatically estimate poses of agents and classify behavior [13, 14, 17, 24, 30, 39].

Pose estimation models that were developed for behavioral videos [16, 30, 35, 39] require human annotations of anatomically defined keypoints, which are expensive and time-consuming to obtain. In addition to the cost, not all data can be crowd-sourced due to the sensitive nature of some experiments. Furthermore, organisms that are translucent (jellyfish) or with complex shapes (tree) can be difficult for non-expert humans to annotate. Our goal is to enable keypoint discovery on videos for behavior analysis, without the need for manual annotations.

After pose estimation, behavior analysis models generally compute trajectory features and train behavior classifiers in a fully supervised fashion [5, 15, 17, 39, 43]. Some works have also explored using unsupervised methods to discover new motifs and behaviors [3, 18, 28, 50]. Here, we apply our discovered keypoints to supervised behavior classification and compare against baseline models using supervised keypoints for this task.

Keypoint Estimation.

Keypoint estimation models aim to localize a predefined set of keypoints from visual data, and many works in this area focus on human pose. With the success of fully convolutional neural networks [40], recent methods [8, 32, 45, 49] employ encoder-decoder networks by predicting high-resolution outputs encoded with 2D Gaussian heatmaps representing each part. To improve model performance, [32, 45, 49] propose an iterative refinement approach, [8, 36] design efficient learning signals, and [9, 47] exploit multi-resolution information. Beyond human pose, there are also works that focus on animal pose estimation, notably [16, 30, 35]. Similar to these works, we also use 2D Gaussian heatmaps to represent parts as keypoints, but instead of using human-defined keypoints, we aim to discover keypoints from video data without manual supervision.

Unsupervised Part Discovery.

Though keypoints provide a useful tool for behavior analysis, collecting annotations is time-consuming and labor-intensive especially for new domains that have not been previously studied. Unsupervised keypoint discovery [20, 21, 51] has been proposed to reduce keypoint annotation effort and there have been many promising results on centered and/or aligned objects, such as facial images and humans with an upright pose. These methods train and evaluate on images where the object of interest is centered in an input bounding box. Most of the approaches [20, 27, 51] use an autoencoder-based architecture to disentangle the appearance and geometry representation for the image reconstruction task. Our setup is

similar in that we also use an encoder-decoder architecture, but crucially, we reconstruct spatiotemporal difference between video frames, instead of the full image as in previous works. We found that this enables our discovered keypoints to track semantically-consistent parts without manual supervision, requiring neither keypoints nor bounding boxes.

There are also works for parts discovery that employ other types of supervision [21, 37, 38]. For example, [37] proposed a weakly-supervised approach using class label to discriminate parts to handle viewpoint changes, [21] incorporated pose prior obtained from unpaired data from different datasets in the same domain, and [38] proposed a template-based geometry bottleneck based on a pre-defined 2D Gaussian-shaped template. Different from these approaches, our method does not require any supervision beyond the behavioral videos. We chose to focus on this setting since other supervisory sources are not readily available for emerging domains (ex: jellyfish, trees).

In previous works, keypoint discovery has been applied to downstream tasks, such as image and video generation [21, 31], keypoint regression to human-annotated poses [20, 51], and video-level action recognition [25, 31]. While we also apply keypoint discovery to downstream tasks, we note that our work differs in approach (we discover keypoints directly on behavioral videos using spatiotemporal difference reconstruction), focus (behavioral videos of diverse organisms), and application (real-world behavior analysis tasks [7, 42]).

3. Method

The goal of B-KinD (Figure 2) is to discover semantically meaningful keypoints in behavioral videos of diverse organisms without manual supervision. We use an encoder-decoder setup similar to previous methods [20, 37], but instead of image reconstruction, here we study a novel reconstruction target based on spatiotemporal difference. In behavioral videos, the camera is generally fixed with respect to the world, such that the background is largely stationary and the agents (*e.g.* mice moving in an enclosure) are the only source of motion. Thus spatiotemporal differences provide a strong cue to infer location and movements of agents.

3.1. Self-supervised keypoint discovery

Given a behavioral video, our work aims to reconstruct regions of motion between a reference frame I_t (the video frame at time t) and a future frame I_{t+T} (the video frame T timesteps later, for some set value of T). We accomplish this by extracting appearance features from frame I_t and keypoint locations (“geometry features”) from both frames I_t and I_{t+T} (Figure 2). In contrast, previous works [20, 21, 27, 37, 38] use appearance features from I_t and geometry features from I_{t+T} to reconstruct the full image I_{t+T} (instead of difference between I_t and I_{t+T}).

We use an encoder-decoder architecture, with shared appearance encoder Φ , geometry decoder Ψ , and reconstruction decoder ψ . During training, the pair of frames I_t and I_{t+T} are fed to the appearance encoder Φ to generate appearance features, and those features are then fed into the geometry decoder Φ to generate geometry features. In our approach, the

reference frame I_t is used to generate both appearance and geometry representations, and the future frame I_{t+T} is only used to generate a geometry representation. The appearance feature h_a^t for frame I_t are defined simply as the output of $\Phi: h_a^t = \Phi(I_t)$.

The pose decoder Φ outputs K raw heatmaps $\mathbf{X}_i \in \mathbb{R}^2$, then applies a spatial softmax operation on each heatmap channel. Given the extracted $p_i = (u_i, v_i)$ locations for $i \in \{1, \dots, K\}$ keypoints from the spatial softmax, we define the geometry features h_g^t to be a concatenation of 2D Gaussians centered at (u_i, v_i) with variance σ .

Finally, the concatenation of the appearance feature h_a^t and the geometry features h_g^t and h_g^{t+T} is fed to the decoder φ to reconstruct the learning objective \hat{S} discussed in the next section: $\hat{S} = \psi(h_a^t, h_g^t, h_g^{t+T})$.

3.2. Learning formulation

3.2.1 Spatiotemporal difference—Our method works with different types of spatiotemporal differences as reconstruction targets. For example:

Structural Similarity Index Measure (SSIM) [48].: This is a method for measuring the perceived quality of the two images based on luminance, contrast, and structure features. To compute our reconstruction target based on SSIM, we apply the SSIM measure locally on corresponding patches between I_t and I_{t+T} to build a similarity map between frames. Then we compute dissimilarity by taking the negation of the similarity map.

Frame differences.: When the video background is static with little noise, simple frame differences, such as absolute difference ($S_d = |I_{t+T} - I_t|$) or raw difference ($S_d = I_{t+T} - I_t$), can also be directly applied as a reconstruction target.

3.2.2 Reconstruction loss—We apply perceptual loss [23] for reconstructing the spatiotemporal difference S . Perceptual loss compares the L2 distance between the features computed from VGG network ϕ [41]. The reconstruction \hat{S} and the target S are fed to VGG network, and mean squared error is applied to the features from the intermediate convolutional blocks:

$$\mathcal{L}_{recon} = \left\| \phi(S(I_t, I_{t+T})) - \phi(\hat{S}(I_t, I_{t+T})) \right\|_2. \quad (1)$$

3.2.3 Rotation equivariance loss—In cases where agents can move in many directions (*e.g.* mice filmed from above can translate and rotate freely), we would like our keypoints to remain semantically consistent. We enforce rotation-equivariance in the discovered keypoints by rotating the image with different angles and imposing that the predicted keypoints should move correspondingly. We apply the rotation equivariance loss (similar to the deformation equivariance in [46]) on the generated heatmap.

Given reference image I and the corresponding geometry bottleneck h_g , we rotate the geometry bottleneck to generate pseudo labels $h_g^{R^\circ}$ for rotated input images I^{R° with degree $R = \{90^\circ, 180^\circ, 270^\circ\}$. We apply mean squared error between the predicted geometry bottlenecks \hat{h}_g from the rotated images and the generated pseudo labels h_g :

$$\mathcal{L}_r = \left\| h_g^{R^\circ} - \hat{h}_g(I^{R^\circ}) \right\|_2. \quad (2)$$

3.2.4 Separation loss—Empirical results show that rotation equivariance encourages the discovered keypoints to converge at the center of the image. We apply separation loss to encourage the keypoints to encode unique coordinates, and prevent the discovered keypoints from being centered at the image coordinates [51]. The separation loss is defined as follows:

$$\mathcal{L}_s = \sum_{i \neq j} \exp\left(\frac{-(p_i - p_j)^2}{2\sigma_s^2}\right). \quad (3)$$

3.2.5 Final objective—Our final loss function is composed of three parts: reconstruction loss \mathcal{L}_{recon} , rotation equivariance loss \mathcal{L}_r , and separation loss \mathcal{L}_s :

$$\mathcal{L} = \mathcal{L}_{recon} + \mathbb{1}_{epoch > n}(w_r \mathcal{L}_r + w_s \mathcal{L}_s). \quad (4)$$

We adopt curriculum learning [2] and apply \mathcal{L}_r and \mathcal{L}_s once the keypoints are consistently discovered from the semantic parts of the target instance.

3.3. Feature extraction for behavior analysis

Following standard approaches [5, 17, 39], we use the discovered keypoints from B-KinD as input to a behavior quantification module: either supervised behavior classifiers or a physics-based model. Note that this is a separate process from keypoint discovery; we feed discovered geometry information into a downstream model.

In addition to discovered keypoints, we extracted additional features from the raw heatmap (Figure 3) to be used as input to our downstream modules. For instance, we found that the confidence and the shape information from the of the network prediction of keypoint location was informative. When a target part is well localized, our keypoint discovery network produces a heatmap with a single high peak with low variance; conversely, when a target part is occluded, the raw heatmap contains a blurred shape with lower peak value. This “confidence” score (heatmap peak value) is also a good indicator for whether keypoints are discovered on the background (blurred over the background with low confidence) or tracking anatomical body parts (peaked with high confidence), visualized in Supplementary materials. The shape of a computed heatmap can also reflect shape information of the target (*e.g.* stretching).

Given a raw heatmap \mathbf{X}_k for part k , the confidence score is obtained by choosing the maximum value from the heatmap, and the shape information is obtained by computing

the covariance matrix from the heatmap. Figure 3 visualizes the features we extract from the raw heatmaps. Using the normalized heatmap as the probability distribution, additional geometric features are computed:

$$\begin{aligned}\sigma_x^2(\mathbf{X}_k) &= \sum_{ij} (x_i - u_k)^2 \mathbf{X}_k(i, j), \\ \sigma_y^2(\mathbf{X}_k) &= \sum_{ij} (y_j - v_k)^2 \mathbf{X}_k(i, j) \\ \sigma_{xy}^2(\mathbf{X}_k) &= \sum_{ij} (x_i - u_k)(y_j - v_k) \mathbf{X}_k(i, j).\end{aligned}\tag{5}$$

4. Experiments

We demonstrate that B-KinD is able to discover consistent keypoints in real-world behavioral videos across a range of organisms (Section 4.1.1). We evaluate our keypoints on downstream tasks for behavior classification (Section 4.2) and pose regression (Section 4.3), then illustrate additional applications of our keypoints (Section 4.4).

4.1. Experimental setting

4.1.1 Datasets

CalMS21.: CalMS21 [42] is a large-scale dataset for behavior analysis consisting of videos and trajectory data from a pair of interacting mice. Every frame is annotated by an expert for three behaviors: sniff, attack, mount. There are 507k frames in the train split, and 262k frames in the test split (video frame: 1024×570 , mouse: approx 150×50). We use only the train split on videos without miniscope cable to train B-KinD. Following [42], the downstream behavior classifier is trained on the entire training split, and performance is evaluated on the test split.

MARS-Pose.: This dataset consists of a set of videos with similar recording conditions to the CalMS21 dataset. We use a subset of the MARS pose dataset [39] with keypoints from manual annotations to evaluate the ability of our model to predict human-annotated keypoints, with {10, 50, 100, 500} images for train and 1.5k images for test.

Fly vs. Fly.: These videos consists of interactions between a pair of flies, annotated per frame by domain experts. We use the Aggression videos from the Fly vs. Fly dataset [15], with the train and test split having 1229k and 322k frames respectively (video frame: 144×144 , fly: approx 30×10). Similar to [43], we evaluate on behaviors of interest with more than 1000 frames in the training set (lunge, wing threat, tussle).

Human 3.6M.: Human 3.6M [19] is a large-scale motion capture dataset, which consists of 3.6 million human poses and images from 4 viewpoints. To quantitatively measure the pose regression performance against baselines, we use the Simplified Human 3.6M dataset, which consists of 800k training and 90k testing images with 6 activities in which the human body is mostly upright. We follow the same evaluation protocol from [51] to use subjects 1,

5, 6, 7, and 8 for training and 9 and 11 for testing. We note that each subject has different appearance and clothing.

Jellyfish.: The jellyfish data is an in-house video dataset containing 30k frames of recorded swimming jellyfish (video frame: 928×1158 , jellyfish: approx 50 pix in diameter). We use this dataset to qualitatively test the performance of B-KinD on a new organism, and apply our keypoints to detect the pulsing motion of the jellyfish.

Vegetation.: This is an in-house dataset acquired over several weeks using a drone to record the motion of swaying trees. The dataset consists of videos of an oak tree and corresponding wind speeds recorded using an anemometer, with a total of 2.41M video frames (video frame: 512×512 , oak tree: varies, approx $\frac{1}{4}$ of the frame). We evaluate this dataset using a physics-based model [7] that relates the visually observed oscillations to the average wind speeds.

4.1.2 Training and evaluation procedure—We train B-KinD using the full objective in Section 3.2.5. During training, we rescale images to 256×256 and use T of around 0.2 seconds, except Human3.6M, where we use 128×128 . Unless otherwise specified, all experiments are ran with all keypoints discovered from B-KinD with SSIM reconstruction and with 10 keypoints for mouse, fly, and jellyfish, 16 keypoints for Human3.6M, 15 keypoints for Vegetation. We train on the train split of each dataset as specified, except for jellyfish and vegetation, where we use the entire dataset. Additional details are in the Supplementary materials.

After training the keypoint discovery model, we extract the keypoints and use it for different evaluations based on the labels available in the dataset: behavior classification (CalMS21, Fly), keypoint regression (MARS-Pose, Human), and physics-based modeling (Vegetation).

For keypoint regression, similar to previous works [20, 21], we compare our regression with a fully supervised 1-stack hourglass network [32]. We evaluate keypoint regression on Simplified Human 3.6M by using a linear regressor without a bias term, following the same evaluation setup from previous works [27, 51]. On MARS-Pose, we train our model in a semi-supervised fashion with 10, 50, 100, 500 supervised keypoints to test data efficiency. For behavior classification, we evaluate on CalMS21 and Fly, using available frame-level behavior annotations. To train behavior classifiers, we use the specified train split of each dataset. For CalMS21 and Fly, we train the 1D Convolutional Network benchmark model provided by [42] using B-KinD keypoints. We evaluate using mean average precision (MAP) weighted equally over all behaviors of interest.

4.2. Behavior classification results

CalMS21 Behavior Classification.—We evaluate the effectiveness of B-KinD for behavior classification (Table 1). Compared to supervised keypoints trained for this task, our keypoints (without manual supervision) is comparable when using both pose and confidence

as input. Compared to other self-supervised methods, even those that use bounding boxes, our discovered keypoints on the full image generally achieve better performance.

Keypoints discovered with image reconstruction, similar to baselines [20,37] cannot track the agents well without using bounding box information (Figure 4) and does not perform well for behavior classification (Table 1). When we provide bounding box information to the model based on image reconstruction, the performance is significantly improved, but this model does not perform as well as B-KinD keypoints from spatiotemporal difference reconstruction.

For the per-class performance (see the Supplementary materials), the biggest gap exists between B-KinD and MARS on the “attack” behavior. This is likely because during attack, the mice are moving quickly, and there exists a lot of motion blur and occlusion which is difficult to track without supervision. However, once we extract more information from the heatmap, through computing keypoint confidence, our keypoints perform comparably to MARS.

Fly Behavior Classification.—The FlyTracker [15] uses hand-crafted features computed from the image, such as contrast, as well as features from tracked fly body parts, such as wing angle or distance between flies. Using discovered keypoints, we compute comparable features without assuming keypoint identity, by computing speed and acceleration of every keypoint, distance between every pair, and angle between every triplet. For all self-supervised methods, we use keypoints, confidence, and covariance for behavior classification. Results demonstrate that while there is a small gap in performance to the supervised estimator, our discovered keypoints perform much better than image reconstruction, and is comparable to models that require bounding box inputs (Table 2).

4.3. Pose regression results

MARS Pose Regression.—We evaluate the pose estimation performance of our method in the setting where some human annotated keypoints exist (Figure 5). For this experiment, we train B-KinD in a semi-supervised fashion, where the loss is a sum of both our keypoint discovery objective (Section 3.2.5) as well as standard keypoint estimation objectives based on MSE [39]. For both black and white mouse, when using our keypoint discovery objective in a semi-supervised way during training, we are able to track keypoints more accurately compared to the supervised method [39] alone. We note that the performance of both methods converge at around 500 annotated examples.

Simplified Human 3.6M Pose Regression.—To compare with existing keypoint discovery methods, we evaluate our discovered keypoints on Simplified Human3.6M (a standard benchmarking dataset) by regressing to annotated keypoints (Table 3). Though our method is directly applicable to full images, we train the discovery model using cropped bounding box for a fair comparison with baselines, which all use cropped bounding boxes centered on the subject. Compared to both self-supervised + prior information and self-supervised + regression, our method shows state-of-the-art performance on the keypoint

regression task, suggesting spatiotemporal difference is an effective reconstruction target for keypoint discovery.

Learning Objective Ablation Study—We report the pose regression performance on Simplified Human3.6M (Table 4) by varying the spatiotemporal difference reconstruction target for training B-KinD. Here, image reconstruction also performs well since cropped bounding box is used as an input to the network. Overall, spatiotemporal difference reconstruction yield better performance over image reconstruction, and performance can be further improved by extracting additional confidence and covariance information from the discovered heatmaps.

4.4. Additional applications

We show qualitative performance and demonstrate additional downstream tasks using our discovered keypoints, on pulse detection for Jellyfish and on wind speed regression for Vegetation.

Qualitative Results.—Qualitative results (Figure 6) demonstrates that B-KinD is able to track some body parts consistently, such as the nose of both mice and keypoints along the spine; the body and wings of the flies; the mouth and gonads of the jellyfish; and points on the arms and legs of the human. For visualization only, we show only keypoints discovered with high confidence values (Section 3.3); for all other experiments, we use all discovered keypoints.

Pulse Detection.—Jellyfish swimming is among the most energetically efficient forms of transport, and its control and mechanics are studied in hydrodynamics research [10]. Of key interest is the relationship between body plan and swim pulse frequency across diverse jellyfish species. By computing distance between B-KinD keypoints, we are able to extract a frequency spectrogram to study jellyfish pulsing, with a visible band at the swimming frequency (Supplementary materials). This provides a way to automatically annotate swimming behavior, which could be quickly applied to video from multiple species to characterize the relationship between swimming dynamics and body plan.

Wind Speed Modeling.—Measuring local wind speed is useful for tasks such as tracking air pollution and weather forecasting [6]. Oscillations of trees encode information on wind conditions, and as such, videos of moving trees could function as wind speed sensors [6, 7]. Using the Vegetation dataset, we evaluate the ability of our keypoints to predict wind speed using a physics-based model [7]. This model defines the relationship between the mean wind speed and the structural oscillations of the tree, and requires tracking these oscillations from video, which was previously done manually. We show that B-KinD can accomplish this task automatically. Using our keypoints, we are able to regress the measured ground truth wind speed with an $R^2 = 0.79$, suggesting there is a good agreement between the proportionality assumption from [7] and the experimental results using the keypoint discovery model.

Limitations.—One issue we did not explore in detail, and which will require further work, is keypoint discovery for agents that may be partially or completely occluded at some

point during observation, including self-occlusion. Additionally, similar to other keypoint discovery models [27, 38, 51], we observe left/right swapping of some body parts, such as the legs in a walking human. One approach that might overcome these issues would be to extend our model to discover the 3D structure of the organism, for instance by using data from multiple cameras. Despite these challenges, our model performs comparably to supervised keypoints for behavior classification.

5. Discussion and conclusion

We propose B-KinD, a self-supervised method to discover meaningful keypoints from unlabelled videos for behavior analysis. We observe that in many settings, behavioral videos have fixed cameras recording agents moving against a (quasi) stationary background. Our proposed method is based on reconstructing spatiotemporal difference between video frames, which enables B-KinD to focus on keypoints on the moving agents. Our approach is general, and is applicable to behavior analysis across a range of organisms without requiring manual annotations.

Results show that our discovered keypoints are semantically meaningful, informative, and enable performance comparable to supervised keypoints on the downstream task of behavior classification. Our method will reduce the time and cost dramatically for video-based behavior analysis, thus accelerating scientific progress in fields such as ethology and neuroscience.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was generously supported by the Simons Collaboration on the Global Brain grant 543025 (to PP and DJA), NIH Award #R00MH117264 (to AK), NSF Award #1918839 (to YY), NSF Award #2019712 (to JOD and RHG), NINDS Award #K99NS119749 (to BW), NIH Award #R01MH123612 (to DJA and PP), NSERC Award #PGSD3-532647-2019 (to JJS), as well as a gift from Charles and Lily Trimble (to PP).

References

- [1]. Anderson David J and Perona Pietro. Toward a science of computational ethology. *Neuron*, 84(1):18–31, 2014. [PubMed: 25277452]
- [2]. Bengio Yoshua, Louradour Jérôme, Collobert Ronan, and Weston Jason. Curriculum learning. In *ICML*, 2009.
- [3]. Berman Gordon J, Choi Daniel M, Bialek William, and Shaevitz Joshua W. Mapping the stereotyped behaviour of freely moving fruit flies. *Journal of The Royal Society Interface*, 11(99):20140672, 2014. [PubMed: 25142523]
- [4]. Branson Kristin, Robie Alice A, Bender John, Perona Pietro, and Dickinson Michael H. High-throughput ethomics in large groups of drosophila. *Nature methods*, 6(6):451–457, 2009. [PubMed: 19412169]
- [5]. Burgos-Artizzu Xavier P, Dollár Piotr, Lin Dayu, Anderson David J, and Perona Pietro. Social behavior recognition in continuous video. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1322–1329. IEEE, 2012.

- [6]. Cardona Jennifer, Howland Michael, and Dabiri John. Seeing the wind: Visual wind speed prediction with a coupled convolutional and recurrent neural network. In Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, and Garnett R, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [7]. Cardona Jennifer L and Dabiri John O. Wind speed inference from environmental flow-structure interactions, part 2: leveraging unsteady kinematics. *arXiv preprint arXiv:2107.09784*, 2021.
- [8]. Chen Yilun, Wang Zhicheng, Peng Yuxiang, Zhang Zhiqiang, Yu Gang, and Sun Jian. Cascaded pyramid network for multi-person pose estimation. *CoRR*, abs/1711.07319, 2017.
- [9]. Cheng Bowen, Xiao Bin, Wang Jingdong, Shi Honghui, Huang Thomas S., and Zhang Lei. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [10]. Costello John H, Colin Sean P, Dabiri John O, Gemmell Brad J, Lucas Kelsey N, and Sutherland Kelly R. The hydrodynamics of jellyfish swimming. *Annual Review of Marine Science*, 13:375–396, 2021.
- [11]. Dankert Heiko, Wang Liming, Hoopfer Eric D, Anderson David J, and Perona Pietro. Automated monitoring and analysis of social behavior in drosophila. *Nature methods*, 6(4):297–303, 2009. [PubMed: 19270697]
- [12]. de Silva Brian M, Higdon David M, Brunton Steven L, and Kutz J Nathan. Discovery of physics from data: universal laws and discrepancies. *Frontiers in artificial intelligence*, 3:25, 2020. [PubMed: 33733144]
- [13]. Egnor SE Roian and Branson Kristin. Computational analysis of behavior. *Annual review of neuroscience*, 39:217–236, 2016.
- [14]. Eyjolfssdottir Eyrún, Branson Kristin, Yue Yisong, and Perona Pietro. Learning recurrent representations for hierarchical behavior modeling. *ICLR*, 2017.
- [15]. Eyjolfssdottir Eyrún, Branson Steve, Burgos-Artizzu Xavier P, Hoopfer Eric D, Schor Jonathan, Anderson David J, and Perona Pietro. Detecting social actions of fruit flies. In *European Conference on Computer Vision*, pages 772–787. Springer, 2014.
- [16]. Graving Jacob M, Chae Daniel, Naik Hemal, Li Liang, Koger Benjamin, Costelloe Blair R, and Couzin Iain D. Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning. *Elife*, 8:e47994, 2019. [PubMed: 31570119]
- [17]. Hong Weizhe, Kennedy Ann, Burgos-Artizzu Xavier P, Zelikowsky Moriel, Navonne Santiago G, Perona Pietro, and Anderson David J. Automated measurement of mouse social behaviors using depth sensing, video tracking, and machine learning. *Proceedings of the National Academy of Sciences*, 112(38):E5351–E5360, 2015.
- [18]. Hsu Alexander I and Yttri Eric A. B-soid, an open-source unsupervised algorithm for identification and fast prediction of behaviors. *Nature communications*, 12(1):1–13, 2021.
- [19]. Ionescu Catalin, Papava Dragos, Olaru Vlad, and Sminchisescu Cristian. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.
- [20]. Jakab Tomas, Gupta Ankush, Bilén Hakan, and Vedaldi Andrea. Unsupervised learning of object landmarks through conditional image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- [21]. Jakab Tomas, Gupta Ankush, Bilén Hakan, and Vedaldi Andrea. Self-supervised learning of interpretable keypoints from unlabelled videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [22]. Jhuang Hueihan, Garrote Estibaliz, Yu Xinlin, Khilnani Vinita, Poggio Tomaso, Andrew D Steele, and Thomas Serre. Automated home-cage behavioural phenotyping of mice. *Nature communications*, 1(1):1–10, 2010.
- [23]. Johnson Justin, Alahi Alexandre, and Fei-Fei Li. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

- [24]. Kabra Mayank, Robie Alice A, Rivera-Alba Marta, Branson Steven, and Branson Kristin. Jaaba: interactive machine learning for automatic annotation of animal behavior. *Nature methods*, 10(1):64, 2013. [PubMed: 23202433]
- [25]. Kim Yunji, Nam Seonghyeon, Cho In, and Kim Seon Joo. Unsupervised keypoint learning for guiding class-conditional video prediction. *arXiv preprint arXiv:1910.02027*, 2019.
- [26]. Liu Jingyuan, Shi Mingyi, Chen Qifeng, Fu Hongbo, and Tai Chiew-Lan. Normalized human pose features for human action video alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11521–11531, 2021.
- [27]. Lorenz Dominik, Bereska Leonard, Milbich Timo, and Ommer Björn. Unsupervised part-based disentangling of object shape and appearance. In *CVPR*, 2019.
- [28]. Luxem Kevin, Fuhrmann Falko, Kürsch Johannes, Remy Stefan, and Bauer Pavol. Identifying behavioral structure from deep variational embeddings of animal motion. *bioRxiv*, 2020.
- [29]. Marstaller Julian, Tausch Frederic, and Stock Simon. Deepbees-building and scaling convolutional neuronal nets for fast and large-scale visual monitoring of bee hives. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.
- [30]. Mathis Alexander, Mamidanna Pranav, Cury Kevin M., Abe Taiga, Murthy Venkatesh N., Mathis Mackenzie W., and Bethge Matthias. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 2018.
- [31]. Minderer Matthias, Sun Chen, Villegas Ruben, Cole Forrester, Murphy Kevin, and Lee Honglak. Unsupervised learning of object structure and dynamics from videos. *arXiv preprint arXiv:1906.07889*, 2019.
- [32]. Newell Alejandro, Yang Kaiyu, and Deng Jia. Stacked hourglass networks for human pose estimation. In *Proc. ECCV*, 2016.
- [33]. Nilsson Simon RO, Goodwin Nastacia L, Choong Jia J, Hwang Sophia, Wright Hayden R, Norville Zane, Tong Xiaoyu, Lin Dayu, Bentzley Brandon S, Eshel Neir, et al. Simple behavioral analysis (simba): an open source toolkit for computer classification of complex social behaviors in experimental animals. *BioRxiv*, 2020.
- [34]. Pereira Talmo D, Shaevitz Joshua W, and Murthy Mala. Quantifying behavior to understand the brain. *Nature neuroscience*, 23(12):1537–1549, 2020. [PubMed: 33169033]
- [35]. Pereira Talmo D, Tabris Nathaniel, Li Junyu, Ravindranath Shruthi, Papadoyannis Eleni S, Wang Z Yan, Turner David M, McKenzie-Smith Grace, Kocher Sarah D, Falkner Annegret Lea, et al. Sleep: Multi-animal pose tracking. *BioRxiv*, 2020.
- [36]. Ryou Serim, Jeong Seong-Gyun, and Perona Pietro. Anchor loss: Modulating loss scale based on prediction difficulty. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [37]. Ryou Serim and Perona Pietro. Weakly supervised keypoint discovery. *CoRR*, abs/2109.13423, 2021.
- [38]. Schmidtke Luca, Vlontzos Athanasios, Ellershaw Simon, Lukens Anna, Arichi Tomoki, and Kainz Bernhard. Unsupervised human pose estimation through transforming shape templates. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2484–2494. Computer Vision Foundation / IEEE, 2021.
- [39]. Segalin Cristina, Williams Jalani, Karigo Tomomi, Hui May, Zelikowsky Moriel, Sun Jennifer J., Perona Pietro, Anderson David J., and Kennedy Ann. The mouse action recognition system (mars): a software pipeline for automated analysis of social behaviors in mice. *bioRxiv* 10.1101/2020.07.26.222299, 2020.
- [40]. Shelhamer Evan, Long Jonathan, and Darrell Trevor. Fully convolutional networks for semantic segmentation. *IEEE TPAMI*, 39(4):640–651, 2017.
- [41]. Simonyan Karen and Zisserman Andrew. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
- [42]. Sun Jennifer J, Karigo Tomomi, Chakraborty Dipam, Mohanty Sharada P, Anderson David J, Perona Pietro, Yue Yisong, and Kennedy Ann. The multi-agent behavior dataset: Mouse dyadic social interactions. *arXiv preprint arXiv:2104.02710*, 2021.

- [43]. Sun Jennifer J, Kennedy Ann, Zhan Eric, Anderson David J, Yue Yisong, and Perona Pietro. Task programming: Learning data efficient behavior representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2876–2885, 2021.
- [44]. Sun Jennifer J, Zhao Jiaping, Chen Liang-Chieh, Schroff Florian, Adam Hartwig, and Liu Ting. View-invariant probabilistic embedding for human pose. In European Conference on Computer Vision, pages 53–70. Springer, 2020.
- [45]. Tang Wei, Yu Pei, and Wu Ying. Deeply learned compositional models for human pose estimation. In The European Conference on Computer Vision (ECCV), September 2018.
- [46]. Thewlis James, Bilen Hakan, and Vedaldi Andrea. Unsupervised learning of object landmarks by factorized spatial embeddings. In The IEEE International Conference on Computer Vision (ICCV), Oct 2017.
- [47]. Wang Jingdong, Sun Ke, Cheng Tianheng, Jiang Borui, Deng Chaorui, Zhao Yang, Liu D, Mu Yadong, Tan Mingkui, Wang Xinggang, Liu Wenyu, and Xiao Bin. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:3349–3364, 2021. [PubMed: 32248092]
- [48]. Wang Zhou, Bovik Alan C., Sheikh Hamid R., and Simoncelli Eero P. Image quality assessment: From error visibility to structural similarity. *IEEE TRANSACTIONS ON IMAGE PROCESSING*, 13(4):600–612, 2004. [PubMed: 15376593]
- [49]. Wei Shih-En, Ramakrishna Varun, Kanade Takeo, and Sheikh Yaser. Convolutional pose machines. In Proc. IEEE CVPR, 2016.
- [50]. Wiltchko Alexander B, Johnson Matthew J, Iurilli Giuliano, Peterson Ralph E, Katon Jesse M, Pashkovski Stan L, Abaira Victoria E, Adams Ryan P, and Datta Sandeep Robert. Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015. [PubMed: 26687221]
- [51]. Zhang Yuting, Guo Yijie, Jin Yixin, Luo Yijun, He Zhiyuan, and Lee Honglak. Unsupervised discovery of object landmarks as structural representations. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2018.

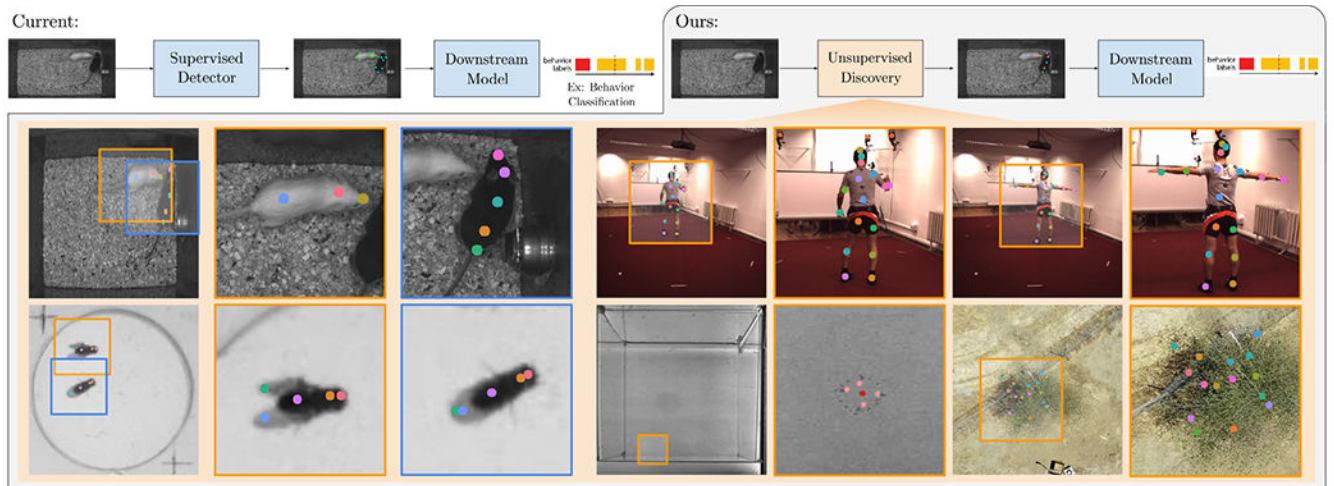


Figure 1. Self-supervised Behavioral Keypoint Discovery.

Intermediate representations in the form of keypoints are frequently used for behavior analysis. We propose a method to discover keypoints from behavioral videos without the need for manual keypoint or bounding box annotations. Our method works across a range of organisms (including mice, humans, flies, jellyfish and tree), works with multiple agents simultaneously (see flies and mice above), does not require bounding boxes (boxes visualized above purely for identifying the enlarged regions of interest) and achieves state-of-the-art performance on downstream tasks.

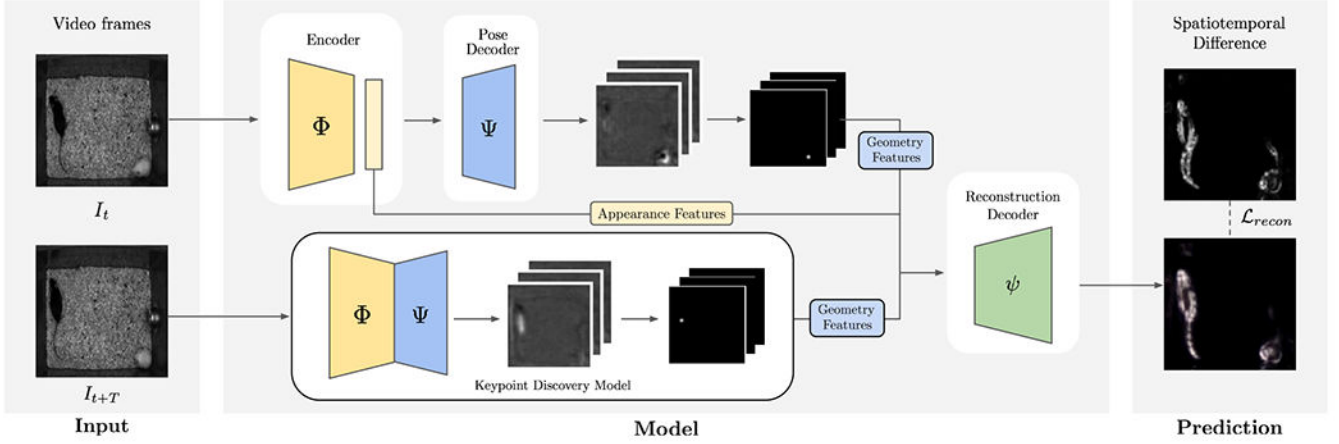


Figure 2. B-KinD, an approach for keypoint discovery from spatiotemporal difference reconstruction.

I_t and I_{t+T} are video frames at time t and $t + T$. Both frame I_t and frame I_{t+T} are fed to an appearance encoder Φ and a pose decoder Ψ . Given the appearance feature from I_t and geometry features from both I_t and I_{t+T} (Sec 3.1), our model reconstructs the spatiotemporal difference (Sec 3.2.1) computed from two frames using the reconstruction decoder ψ .

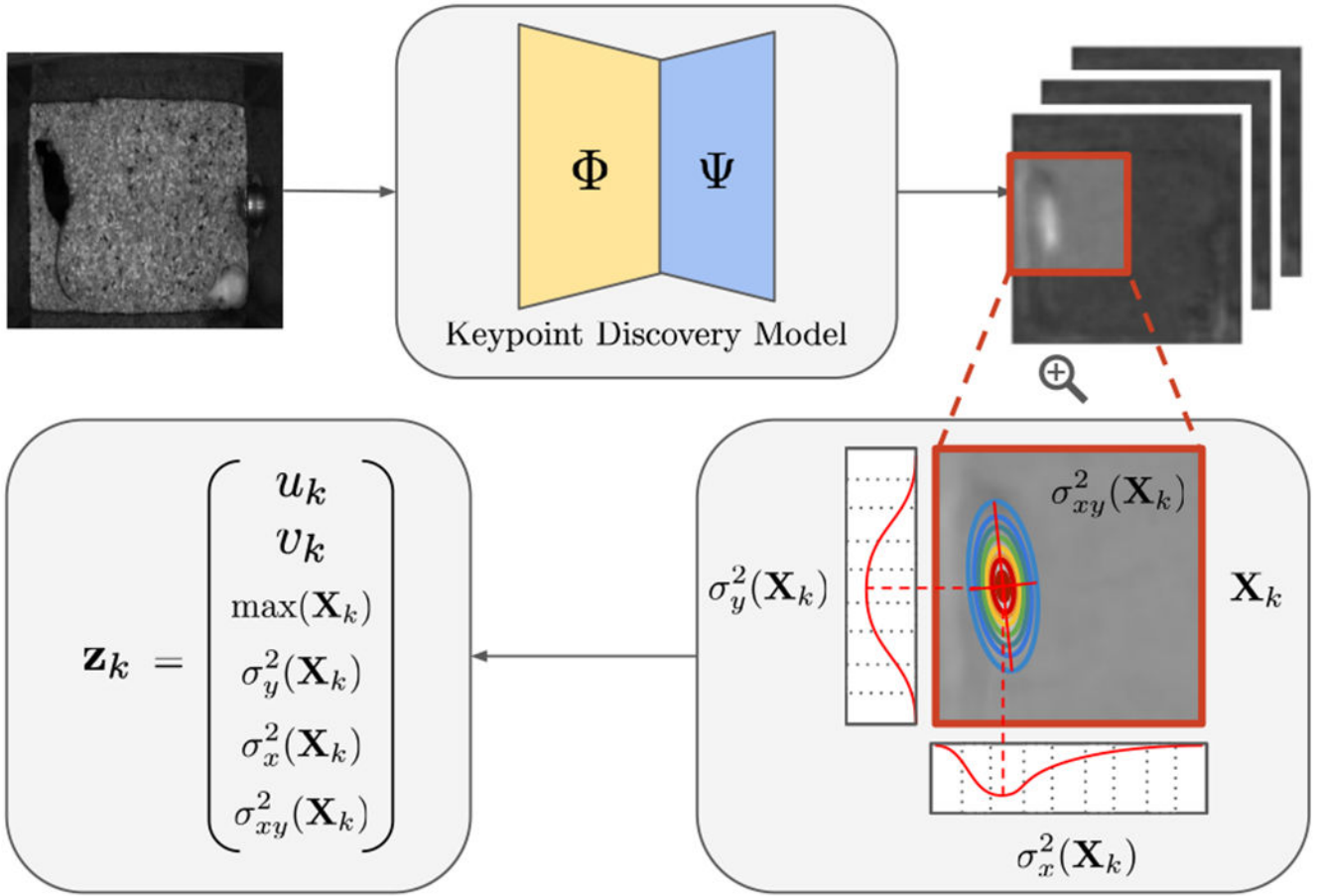


Figure 3. Behavior Classification Features.

Extracting information from the raw heatmap (Section 3.3): the confidence scores and the covariance matrices are computed from normalized heatmaps. Note that the features are computed for all x, y coordinates. We visualize the zoomed area around the target instance for illustrative purposes.

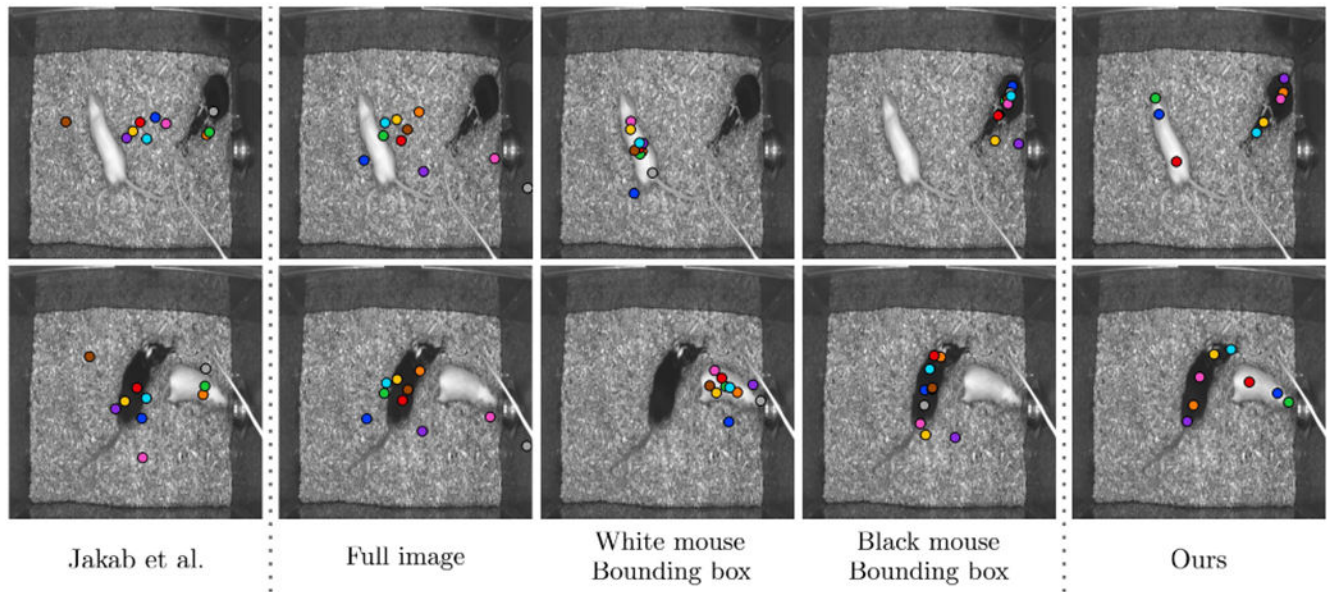


Figure 4.

Comparison with existing methods [20], full image, bounding box, and SSIM reconstruction (ours). “Jakab *et al.*” and “full image” results are based on full image reconstruction. “White mouse bounding box” and “black mouse bounding box” show the results when the cropped bounding boxes were fed to the network for image reconstruction.

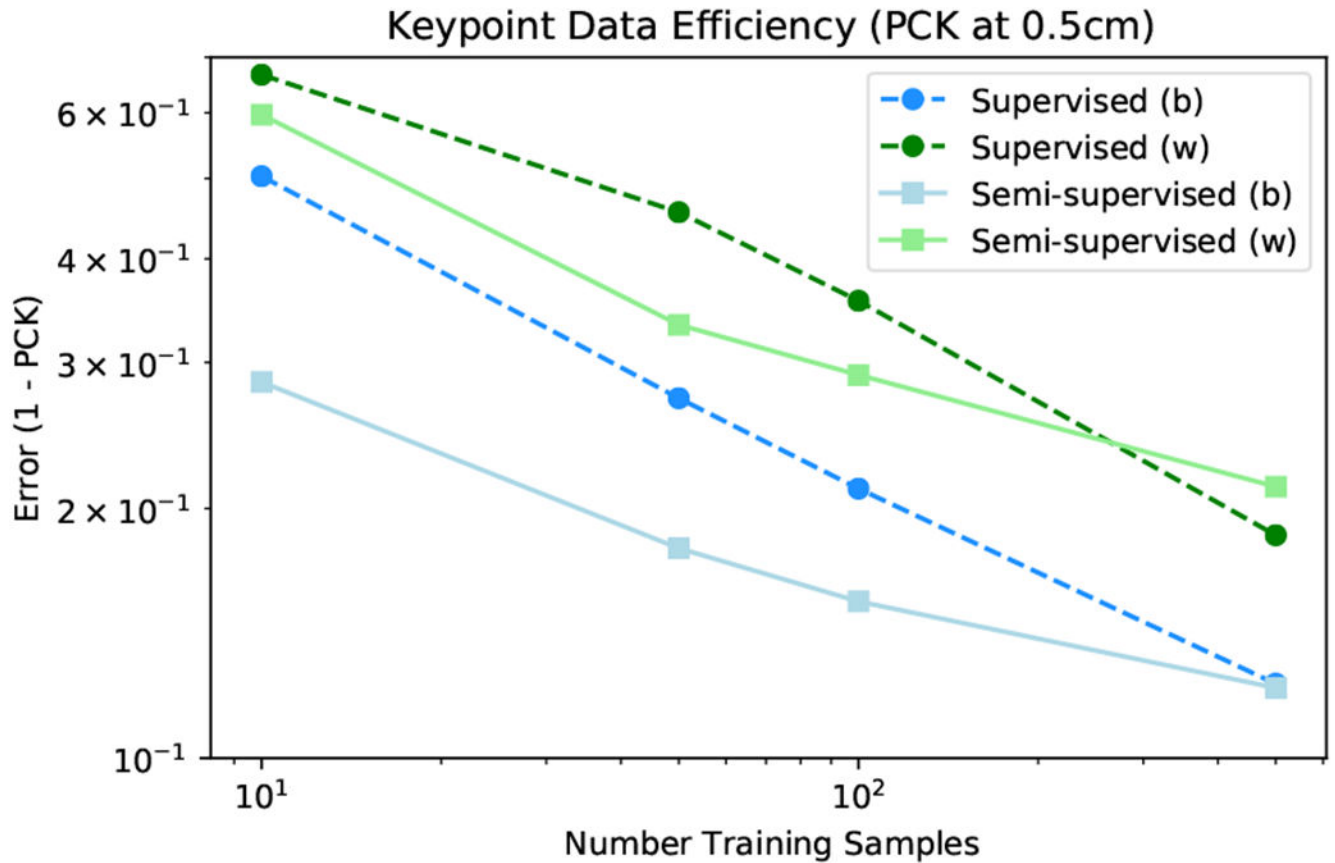


Figure 5. Keypoint data efficiency on MARS-Pose.

The supervised model is based on [39] using stacked hourglass [32], while the semi-supervised model uses both our self-supervised loss and supervision. PCK is computed at $0.5cm$ threshold, averaged across nose, ears, and tail keypoints, over 3 runs. “b” and “w” indicates the black and white mouse respectively.

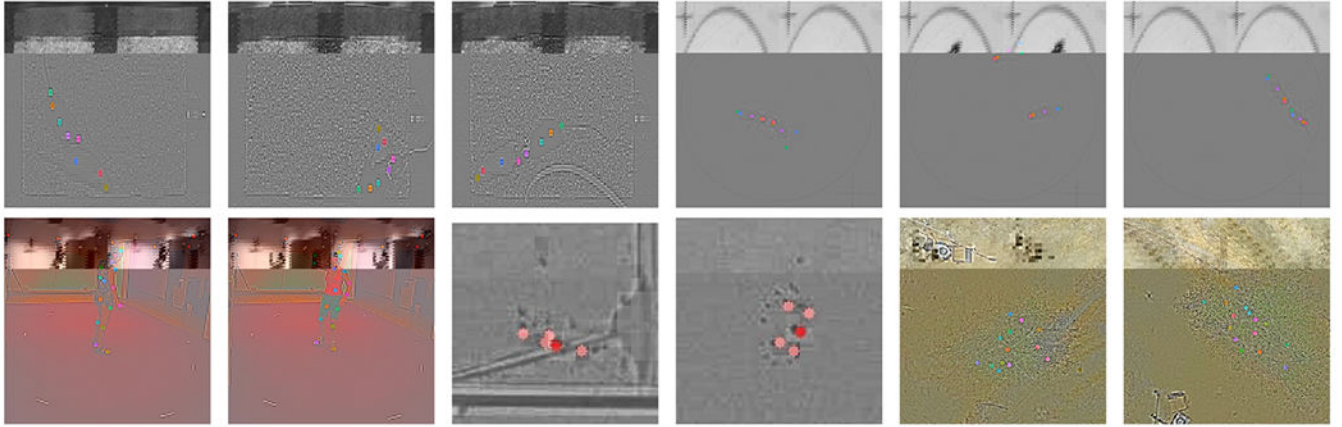


Figure 6. Qualitative Results of B-KinD.

Qualitative results for B-KinD trained on CalMS21 (mouse), Fly vs. Fly (fly), Human3.6M (human), jellyfish and Vegetation (tree). Additional visualizations are in the Supplementary materials.

Table 1.
Behavior Classification Results on CalMS21.

“Ours” represents classifiers using input keypoints from our discovered keypoints. “conf” represents using the confidence score, and “cov” represents values from the covariance matrix of the heatmap.

CalMS21	Pose	Conf	Cov	MAP
<i>Fully supervised</i>				
	✓			.856 ± .010
MARS † [39]	✓	✓		.874 ± .003
	✓	✓	✓	.880 ± .005
<i>Self-supervised</i>				
Jakab et al. [20]	✓			.186 ± .008
	✓			.182 ± .007
Image Recon.	✓	✓		.184 ± .006
	✓	✓	✓	.165 ± .012
	✓			.819 ± .008
Image Recon. bbox †	✓	✓		.812 ± .006
	✓	✓	✓	.812 ± .010
	✓			.814 ± .007
Ours	✓	✓		.857 ± .005
	✓	✓	✓	.852 ± .013

† refers to models that require bounding box inputs before keypoint estimation. Mean and std dev from 5 classifier runs are shown.

Table 2.
Behavior Classification Results on Fly.

“FlyTracker” represents classifiers using hand-crafted inputs from [15]. The self-supervised keypoints all use the same “generic features” computed on all keypoints: speed, acceleration, distance, and angle.

Fly	MAP
<i>Hand-crafted features</i>	
FlyTracker [15]	.809 ± .013
<i>Self-supervised + generic features</i>	
Image Recon.	.500 ± .024
Image Recon. bbox [†]	.750 ± .020
Ours	.727 ± .022

[†] refers to models that require bounding box inputs before keypoint estimation. Mean and std dev from 5 classifier runs are shown.

Table 3.
Comparison with state-of-the-art methods for landmark prediction on Simplified Human 3.6M.

The error is in %-MSE normalized by image size.

Simplified H36M	all	wait	pose	greet	direct	discuss	walk
<i>Fully supervised:</i>							
Newell [32]	2.16	1.88	1.92	2.15	1.62	1.88	2.21
<i>Self-supervised + unpaired labels</i>							
Jakab [21] [‡]	2.73	2.66	2.27	2.73	2.35	2.35	4.00
<i>Self-supervised + template</i>							
Schmidtke [38]	3.31	3.51	3.28	3.50	3.03	2.97	3.55
<i>Self-supervised + regression</i>							
Thewlis [46]	7.51	7.54	8.56	7.26	6.47	7.93	5.40
Zhang [51]	4.14	5.01	4.61	4.76	4.45	4.91	4.61
Lorenz [27]	2.79	–	–	–	–	–	–
Ours (best)	2.44	2.50	2.22	2.47	2.22	2.77	2.50
Ours (mean)	2.53	2.58	2.31	2.56	2.34	2.83	2.58
Ours (std)	.056	.047	.062	.048	.066	.048	.063

All methods predict 16 keypoints except for [21][‡], which uses 32 keypoints for training a prior model from the Human 3.6M dataset. B-Kind results are computed from 5 runs.

Table 4.
Learning objective ablation on Simplified Human3.6M.

%-MSE error is reported by changing the reconstruction target. Extracted features correspond to keypoint locations, confidence, and covariance. Results are from 5 B-KinD runs.

Learning Objective	%-MSE
Image Recon.	2.918 ± 0.139
Abs. Difference	2.642 ± 0.174
Difference	2.770 ± 0.158
SSIM	2.534 ± 0.056
<i>Self-supervised + extracted features</i>	
SSIM	2.494 ± 0.047

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript