Multiview Supervision By Registration

Yilun Zhang
University of Pennsylvania
zhyilun@seas.upenn.edu

Hyun Soo Park
University of Minnesota
hspark@umn.edu

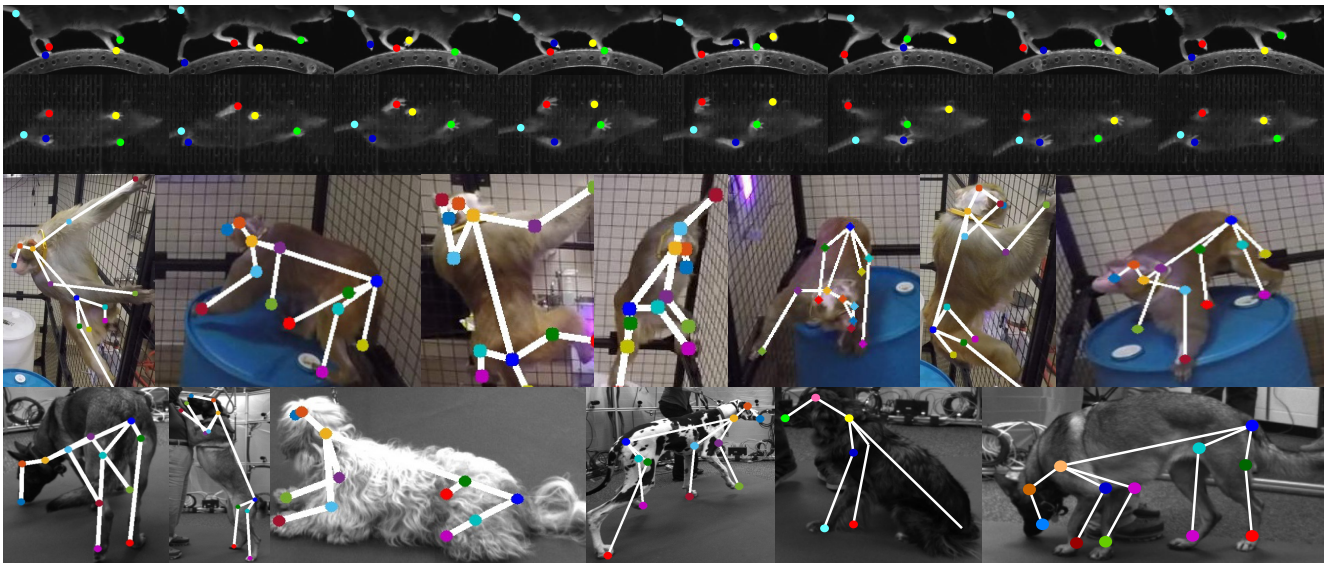


Figure 1: This paper presents a semi-supervised learning method to train a keypoint detector by leveraging multiview tracking. This keypoint detector can localize a set of joints for non-humans species such as mouse, monkey, and dogs, where attaining a large scale annotated data is extremely challenging.

Abstract

This paper presents a semi-supervised learning framework to train a keypoint detector using multiview image streams given the limited number of labeled instances (typically <4%). We leverage three self-supervisionary signals in multiview tracking to utilize the unlabeled data: (1) a keypoint in one view can be supervised by other views via epipolar geometry; (2) a keypoint detection must be consistent across time; (3) a visible keypoint in one view is likely to be visible in the adjacent view. We design a new end-to-end network that can propagate these self-supervisionary signals across the unlabeled data from the labeled data in a differentiable manner. We show that our approach outperforms existing detectors including DeepLabCut tailored to the keypoint detection of non-human species such as monkeys, dogs, and mice.

1. Introduction

Enabling computational measurements of the motor behaviors of animals gives rise to scaling up neuroscientific experiments with an unprecedented precision, leading to deeper understanding of our behaviors (humans). For instance, human surrogate models, such as monkeys and mice, have been studied to identify the neural-behavioral pathway through their *free-ranging* activities (including several social interactions), which is largely homologous to humans. While non-invasive markerless motion capture is a viable solution to measure such behaviors, it still remains blind to animal behaviors because of lack of a large-scale annotated dataset unlike human subjects (e.g., MS COCO [21] and MPII [1]).

Recently, subject-agnostic pose tracking approaches based on deep neural networks such as DeepLabCut [24] have shown remarkable generalization power, allowing a *smart* pose interpolation: a pre-trained network based on a generic large image dataset (e.g., ImageNet [29]) is refined to learn a pose variation from a few hundreds of annotated images in a video, and then, the refined network tracks the poses in the rest video by detection. It is relatively labor-effective (comparing to labeling millions of images) and resilient to a target, i.e., the keypoints on body, foot, and finger of cheetah, insects, and mouse can be reliably tracked. However, their application to the free-ranging behaviors[1] is challenging because such motion introduces a larger pose variation and self-occlusion, and therefore, considerable amount of annotations is needed. Figure 6(d-e) illustrates its performance degradation as the range of motion increases (i.e., mice $\ll$ monkeys).

[1]Their approaches are designed to track restricted motion, e.g., the animal's head be immobile and attached to a recording rig [33].

This paper presents a new semi-supervised learning approach for a pose detector that leverages the complementary relationship between multiview geometry and visual tracking given the limited labeled data. We hypothesize that the annotation efforts can be substantially reduced by utilizing three self-supervisionary signals embedded in multiview image streams[2]. (1) Multiview supervision: the pose detection from two views must satisfy the epipolar constraint, i.e., the detected keypoint in one view must lie in the corresponding epipolar line transferred from the other view given their fundamental matrix [11]. We integrate the cross-view supervision [37] by matching the keypoint distributions from two views via their common epipolar plane. This eliminates the necessity of 3D reconstruction[3]. (2) Temporal supervision: a pose changes continuously. We incorporate the dense tracking to warp the keypoint distribution between consecutive frames to supervise them to each other [8, 36]. (3) Visibility supervision: free-ranging activities inherently involve with frequent self-occlusions, producing spurious and degenerate detection. Inspired by the observation that the keypoint visibility varies smoothly across views [16], we use the spatial proximity of the cameras to supervise the visibility map in one view from the adjacent views. These three supervisionary signals are combined to form an end-to-end system that effectively uses both labeled and unlabeled data.

Our system takes as input multiview image streams with a small set of annotated frames, and outputs a pose detection network that predicts the keypoint locations on the rest unlabeled data. We propose a new formulation of multiview semi-supervised learning by matching keypoint distributions conditioned on a visibility map across frames and views. The formulation is implemented using a novel network design composed of three pathways that can minimize the distribution mismatches in the form of four losses: label loss, cross-view loss, tracking loss, and visibility loss. We demonstrate that the resulting network shows strong performance in terms of the keypoint detection accuracy in the presence of significant occlusion given a small set of labeled data ($<$4%).

Our approach inherits the flexible nature of epipolar geometry, which can be applied to various camera configurations. The distribution matching through their fundamental matrix eliminates the requirement of 3D reconstruction that involves with alternating reconstruction [4, 6, 30, 34] or data driven depth prediction [17, 32, 39]. Finally, our design is network-agnostic, i.e., any pose detection network producing a probability map representation can be used with a trivial modification such as DeepPose [31], CPM [7, 35], and Hourglass [26].

To our knowledge, this is the first paper that leverages the spatiotemporal relationship of multiview image streams to train a pose detector. The core contributions include: (1) a new differentiable formulation of multiview spatiotemporal self-supervision for the unlabeled data; (2) a visibility supervision based on camera spatial proximity to prevent from spurious propagation of the self-supervision; (3) its realization using an end-to-end network that is flexible to camera configurations; and (4) strong performance on the realworld data of non-human species on monkeys, dogs, and mice with a small set of the labeled data.

[2]Similar insight has been used to reconstruct a reliable long-term 3D trajectories with the multiview videos [10, 16, 38].

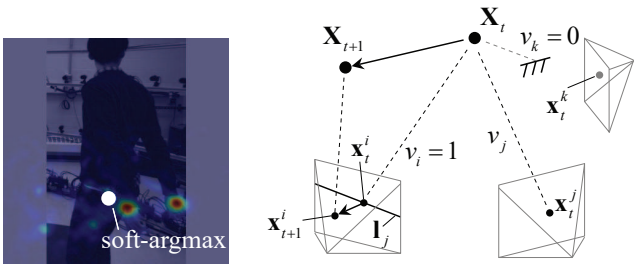
[3]This is analogous to the fundamental matrix computation without 3D estimation [11, 23].

2. Related Work

This paper studies designing a pose detector given the limited labeled data by leveraging multiview epipolar geometry and temporal consistency. These two supervisionary signals are by large studied in isolation.

Temporal Supervision The tracking results such as optical flow [3], MOSSE [5], and discriminative correlation filters [13], provides an auxiliary information that can be used to enforce the temporal consistency across a continuous sequence [8, 36]. A challenge is that it suffers from tracking drift induced by object deformation, which substantially limits its validity. Such challenge has been addressed by learning the temporal evolution of tracking patches [22, 27] using recurrent neural networks. This generates a compromised network that minimizes the inconsistency in the learned trajectories, which suppresses the low-quality detection from the tracking drift. A pitfall of this approach is the requirement of per-frame annotation to supervise the recurrent network. This requirement can be relaxed by using supervision-by-registration approach [8] that achieves higher detection rate even with the limited labeled data. However, its application towards the pose detection for non-human species is still challenging because: (1) supervision from optical flow involves with the tracking drift caused by occlusion, and therefore, long-term tracking is infeasible; (2) the soft-argmax operation for computing the track coordinate may lead to noisy supervision in the cases where the pose detection is erroneous (e.g., multiple peaks) as shown in Figure 2(a). This multi-modality of pose recognition escalates when the keypoint is invisible. This strongly influences tracking accuracy, especially for a small-sized target; (3) the argmax operation takes into account only for the peak location where the non-maximum local peaks may play a role.

Multiview Supervision Multiview images possess highly redundant yet distinctive visual information that can be used to self-supervise the unlabeled data. Bootstrapping is a common practice: to use multiview images to robustly reconstruct the geometry using the correspondences and to project to the unlabeled images to provide a pseudo-label, which has been shown highly effective [6, 30, 34]. A pitfall of this approach is that it involves an iterative process over learning and reconstruction. Another approach is to separately learn depth from a single view image in isola-



(a) Spurious soft-argmax (b) Multiview supervision

Figure 2: (a) Soft-argmax produces a biased keypoint estimate when the keypoint distribution is multimodal. (b) We use three self-supervisionary signals: cross-view supervision ($\mathbf{l}_j^\intercal \widetilde{\mathbf{x}}_t^j$), temporal supervision ($\mathbf{x}_t^i = W_{t+1\to t}(\mathbf{x}_{t+1}^i)$), and visibility supervision ($v_i \approx v_j$).

tion that can be used for self-supervision [17, 32, 39]. This relies on the depth prediction where the accuracy of the trained model is bounded by the accuracy of reconstruction/prediction. Yao et al. [37] introduces a new framework that bypasses 3D reconstruction during the training process through the epipolar constraint, i.e., the epipolar constraint is transformed to the distribution matching. The problem of this approach is that its performance is highly dependent on the pre-trained model. It has no reasoning about outliers, i.e., the recognition network converges to a trivial solution if the outliers dominate the distribution of the multiview pose detection.

Our main hypothesis is that these two supervisions are complementary. We formulate the spatiotemporal supervision that can benefit from both and address each limitation. (1) We use dense optical flow tracking to address noisy supervision, i.e., it is unlikely that the noisy prediction is temporally correlated. (2) We leverage the end-to-end epipolar distribution matching to avoid the multimodality issue that arises using the soft-argmax operation. This is differentiable, and therefore, trainable. (3) The multiview image streams can alleviate the tracking drift [16, 38], i.e., it is unlikely that the tracking drift occurs in a geometrically consistent fashion. (4) Visibility map can assist to determine the validity of the tracking without explicit outlier rejection.

3. Notation and Multiview Conditions

Consider multiview image streams, $\mathcal{I} = \{\mathbf{I}_t^i\}$ where $\mathbf{I}_t^i$ is the image of the i^{th} camera at t time instant. We denote the set of synchronized images at t time instant across all views with $\mathcal{I}_t = \{\mathbf{I}_t^1, \cdots \mathbf{I}_t^n\}$ that satisfy the epipolar constraint [23] where n is the number of cameras. $\mathcal{I}^i = \{\mathbf{I}_1^i, \cdots \mathbf{I}_T^i\}$ is the set of images from the i^{th} camera for all time instances where T is the total time instances[4]. A subset of these images are manually annotated (keypoint location) $\mathcal{I}_L$, and the rest remain unlabeled $\mathcal{I}_U$, i.e., $\mathcal{I} = \mathcal{I}_L \cup \mathcal{I}_U$.

A 3D keypoint $\mathbf{X}_t \in \mathbb{R}^3$ at t time instant travels to $\mathbf{X}_{t+1}$. The point is projected onto the i^{th} and j^{th} images ($\mathbf{I}_t^i$ and $\mathbf{I}_t^j$) to form the 2D projections $\mathbf{x}_t^i, \mathbf{x}_t^j \in \mathbb{R}^2$ as shown in Figure 2(b):

$$\widetilde{\mathbf{x}}_t^i \cong \mathbf{P}^i \widetilde{\mathbf{X}}_t, \quad \widetilde{\mathbf{x}}_t^j \cong \mathbf{P}^j \widetilde{\mathbf{X}}_t, \quad \widetilde{\mathbf{x}}_{t+1}^i \cong \mathbf{P}^i \widetilde{\mathbf{X}}_{t+1}, \tag{1}$$

where $\mathbf{P}^i \in \mathbb{R}^{3\times 4}$ is the i^{th} camera projection matrix, and $\widetilde{\mathbf{x}}$ is the homogeneous representation of $\mathbf{x}$ [11].

To be geometrically consistent across multiview image streams, the projections of the moving 3D keypoint need to satisfy the following three constraints:

Cross-view Constraint The keypoint $\mathbf{x}_t^i$ must lie in the epipolar line of the corresponding point $\mathbf{x}_t^j$ in the j^{th} view [11], i.e., $(\widetilde{\mathbf{x}}_t^j)^\intercal \mathbf{F}_{ij} \widetilde{\mathbf{x}}_t^i = \mathbf{l}_j^\intercal \widetilde{\mathbf{x}}_t^i = 0$ where $\mathbf{F}_{ij}$ is the fundamental matrix between the i^{th} and j^{th} views, and $\mathbf{l}_j \in \mathbb{P}^2$ is the epipolar line transferred from $\mathbf{x}_t^j$.

Tracking Constraint The pixel brightness on $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$ must be persistent, $\mathbf{I}_t^i(\mathbf{x}_{t+1}^i + \Delta\mathbf{x}) = \mathbf{I}_{t+1}^i(\mathbf{x}_{t+1}^i)$ where $\Delta\mathbf{x}$ is the backward optical flow at $\mathbf{x}_{t+1}^i$.

Visibility Constraint The visible keypoint in one view is likely visible in adjacent view, i.e., $v_i \approx v_j$ if $\|\mathbf{C}_i - \mathbf{C}_j\| < \epsilon$ where $v_i \in [0, 1]$ is the probability of the keypoint being visible to the i^{th} camera, and $\mathbf{C}_i$ is the optical center of the i^{th} camera. For instance, $v_i = v_j = 1$ and $v_k = 0$ in Figure 2(b).

4. Multiview Supervision by Registration

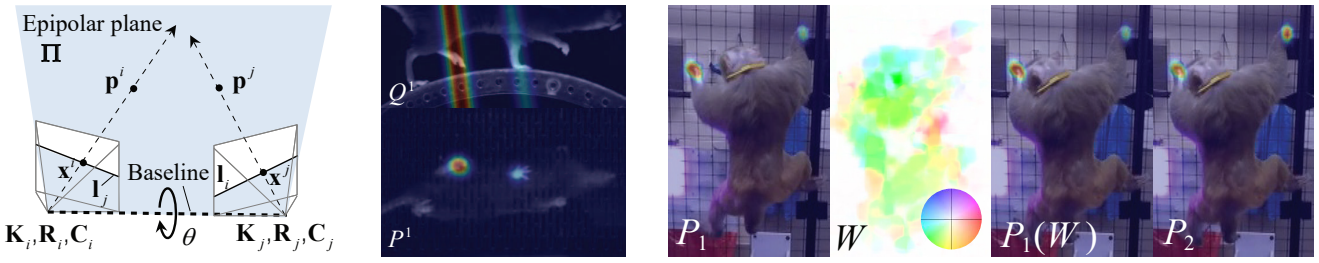
We build a keypoint detector producing the keypoint distribution $\phi(\mathbf{I}; \mathbf{w}) \in [0, 1]^{W\times H\times C}$ and its visibility map $\psi(\mathbf{I}; \mathbf{w}_v) \in [0, 1]^{W\times H\times C}$. These two distributions are combined to produce a posterior per-pixel keypoint distribution:

$$\xi(\mathbf{I}) = \phi(\mathbf{I}; \mathbf{w})\psi(\mathbf{I}; \mathbf{w}_v) \tag{2}$$

where W, H, and C are the width, height, and the number of keypoints including the background. The keypoint distribution is parametrized by the weight $\mathbf{w}$, and the visibility map is parametrized by the weight $\mathbf{w}_v$. We denote the probability evaluated at $\mathbf{x}$ as $P_t^i(\mathbf{x}) = \phi(\mathbf{I}_t^i; \mathbf{w})\big|_{\mathbf{x}}$ and $V_t^i(\mathbf{x}) = \psi(\mathbf{I}_t^i; \mathbf{w}_v)\big|_{\mathbf{x}}$. In the inference phase, the resulting detected keypoint location is the peak in the posterior distribution ξ.

We learn $\mathbf{w}$ and $\mathbf{w}_v$ from the labeled and unlabeled data where $|\mathcal{I}_L| \ll |\mathcal{I}_U|$ where a supervised learning approach alone likely to be highly biased. To utilize the unlabeled data, we leverage the three multiview constraints in Section 3. However, integrating these into an end-to-end training is challenging because of *representation mismatch*. The raster representation of the keypoint distribution $\phi(\mathbf{I}; \mathbf{w})$ differs from the vector representation of the

[4] We consider a stationary multi-camera system [16, 38] while the spatiotemporal constraint of epipolar geometry and temporal coherence still applies for a moving synchronized multi-camera system, e.g., social cameras [2].



(a) Epipolar plane (b) Cross-view supervision (c) Temporal supervision

Figure 3: (a) A keypoint distribution can be transformed to the common epipolar plane distribution, allowing cross-view supervision. (b) The keypoint distribution of the hind right foot in view 1 (P^1) is transformed to Q^i and projected onto the side view (top). (c) The keypoint distribution can be warped $P_1(W)$ using dense optical flow (W) to supervise the next frame P_2, which is multimodal distribution.

constraints (e.g., $\mathbf{l}^\mathsf{T}\widetilde{\mathbf{x}} = 0$). Conversion between these two representations requires the argmax operation:

$$\mathbf{x}^* = \underset{\mathbf{x}}{\text{argmax}} \; P_t^i(\mathbf{x}). \tag{3}$$

The argmax in Equation (3) is non-differentiable, and therefore, embedding the constraints makes the network not trainable. This precludes from an end-to-end training for multiview supervision, leading to offline alternating reconstruction [4, 6, 30, 34] or additional depth prediction [17, 32, 39] that often suffer from suboptimality [37]. Whilst the differentiable soft-argmax can alleviate this issue to some extent, it is highly sensitive to spurious and noisy keypoint detection (e.g., multimodal probability map as shown in Figure 2(a)). In subsequent sections, we address this challenge by transforming the constraints into a distribution matching with the raster representation as a whole by minimizing KL divergence [19].

4.1. Cross-view Supervision

A set of images at the same time instant, $\mathcal{I}_t$, we supervise their keypoint distributions based on the epipolar constraint. Inspired by Yao et al. [37], we reformulate the epipolar geometry in terms of distribution matching over their common epipolar planes. Consider a keypoint in the i^{th} image, $\mathbf{x}^i$, that corresponds to the keypoint in the j^{th} image $\mathbf{x}^j$. Their inverse projections (the 3D ray emitted from the camera center and passing the keypoint location $\mathbf{x}_i$) can be written as $\mathbf{p}_i(\lambda) = \lambda \mathbf{R}_i^\mathsf{T}\mathbf{K}_i^{-1}\widetilde{\mathbf{x}}_i + \mathbf{C}_i$ where $\mathbf{K}_i \in \mathbb{R}^{3\times3}$, $\mathbf{R}_i \in SO(3)$, and $\mathbf{C}_i \in \mathbb{R}^3$ are the intrinsic parameter, rotation, and optical center of the i^{th} camera, and $\lambda > 0$ is the depth of the point on the ray as shown in Figure 3(a). To satisfy the epipolar constraint, their inverse projections must lie in a common epipolar plane ($\mathbf{\Pi} \in \mathbb{P}^3$), i.e., $\mathbf{\Pi}^\mathsf{T}\widetilde{\mathbf{p}}_i = \mathbf{\Pi}^\mathsf{T}\widetilde{\mathbf{p}}_j = 0$.

Using the fact that the common epipolar plane can be parametrized by its rotation about the baseline, i.e., surface normal $\mathbf{\Pi}(\theta \in \mathbb{S})$, we transform the keypoint distribution to the epipolar plane distribution, obtained by the max-pooling over the epipolar line:

$$Q^i(\theta) = \underset{\mathbf{x}\in\mathbf{l}_j(\theta)}{\text{argmax}} \; P^i(\mathbf{x}), \tag{4}$$

where $\mathbf{l}_j(\theta)$ is the epipolar line that is the projection of the common epipolar plane, and Q^i is the epipolar plane distribution. The bottom row in Figure 3(b) illustrates the keypoint distribution of the right hind foot in view 1 (P^1). It is transformed to the epipolar plane distribution Q^1 using the max-pooling over the epipolar lines. We visualize the projection of Q^1 onto the second view (the top row), i.e., the hind foot must lie in the most probable location in the second view. Note that the multimodal keypoint distribution does not produce additional spurious supervision to the other view.

Equation (4) allows measuring geometric discrepancy of keypoint distributions across views. Therefore, the unlabeled data can be self-supervised to each other by minimizing their cross entropy with the raster representation:

$$L_{\text{C}}(\mathcal{I}_t) = \sum_{i,j\in\mathcal{C}} D_{\text{KL}}(Q_i||Q_j), \tag{5}$$

where $\mathcal{C}$ is the camera index set of $\mathcal{I}_t$.

4.2. Temporal Supervision

Given a sequence of images from the i^{th} camera, $\mathcal{I}^i$, we supervise the keypoint distribution at t^{th} time instant using that of neighboring images in time, i.e.,

$$P_{t_1}^i(\mathbf{x}) \approx P_{t_2}^i(W_{t_2\to t_1}(\mathbf{x})) \tag{6}$$

where $W_{t_2\to t_1}$ is the pre-computed dense optical flow from t_2 to t_1 frames, i.e., $P_{t_2}^i(W_{t_2\to t_1}(\mathbf{x}))$ is the warped distribution of $P_{t_2}^i$. We use a kernelized correlation filter [13] with inverse compositional mapping [3] to track all pixels offline while online optical flow computation [8, 20] can be complementary to our approach with a trivial modification.

Using Equation (6), we design a tracking loss for the temporal supervision:

$$L_{\text{T}}(\mathcal{I}^i) = \sum_{t_1,t_2\in[0,T]} D_{\text{KL}}(P_{t_1}^i||P_{t_2}^i(W_{t_2\to t_1})), \tag{7}$$

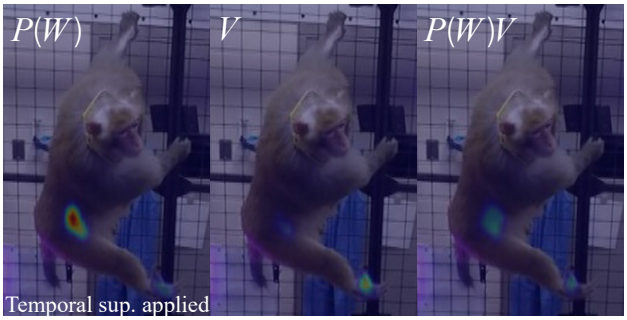


Figure 4: We integrate a visibility inference to validate the multiview supervisory signals. The left hind paw is occluded by torso, which is conditioned by the visibility map (middle), resulting in the reduction of the keypoint probability. This prevents from influencing the occluded keypoint detection across views.

where T is the number of frames.

A key innovation of Equation (7) against existing optical flow supervision [8,21,36] is that it eliminates the necessity of the argmax operation by warping the keypoint distribution as a whole. In practice, we find that having sufficient time difference between frames improves training performance and efficiency. For instance, a high framerate video of a monkey who stays still for a majority of time generates less informative temporal supervision and is prone to noise, i.e., $W_{t+1\to t} = I$ where I is the identity mapping. On the other hand, when the frame difference is too large, significant tracking drift is likely to occur. We address this by selectively applying the temporal supervision on the two frames that have the sufficient magnitude of the integral dense optical flow, i.e., $\epsilon_m < \sum_{x\in\mathcal{X}} \|W_{t_2\to t_1}(\mathbf{x})\| < \epsilon_M$ where $\mathcal{X}$ is the domain of an image, and ϵ_m and ϵ_M are lower and upper bounds of the magnitude of the integral dense optical flow. Figure 3(c) illustrates the temporal supervision using dense optical flow. The left wrist keypoint distribution P_1 is warped to form $P_1(W)$. This unimodal distribution can supervise the ambiguous prediction in P_2 with two modes.

4.3. Visibility Supervision

Free-ranging activities inherently involve with self-occlusion, e.g., a hand is occluded by the torso at a certain view. Without precise reasoning about the visibility of keypoints, the cross-view and temporal supervisions can be highly fragile because there is no mechanism to prevent from such error propagation over the unlabeled multiview images[5]. For instance, the temporal supervision via the optical flow of the occluded hand can mislead the hand location to the torso location in other visible images. To reject such error, RANSAC [9] with geometric verification (e.g., reprojection error) has been used. However, the operation is non-differentiable, and therefore, it requires alternating offline reconstruction and training [30].

Instead, we design a new module that integrates the visibility inference as a part of the training process. The key idea is that a keypoint is likely to be visible if it is visible from the adjacent cameras. This provides a spatial prior on the visibility map across views:

$$L_{\mathrm{V}}(\mathcal{I}_t) = \sum_{i,j\in\mathcal{C}} \delta_{i,j} \| \max\, V_t^i - \max\, V_t^j \|^2, \qquad (8)$$

where $\delta_{i,j}$ is Kronecker delta that is one if the distance between the optical centers of the i^{th} and j^{th} cameras is smaller than ϵ_C, i.e., $\|\mathbf{C}_i - \mathbf{C}_j\| < \epsilon_C$, and zero otherwise, and $\mathcal{C}$ is the camera index set of $\mathcal{I}_t$. Equation (8) is a necessary condition that penalizes the difference in visibility maps for adjacent cameras, i.e., it is valid when the location of the maximum visibility map coincides with the peak of the keypoint probability. In practice, the visibility is highly correlated with the keypoint distribution where L_{V} is effective. For instance, Figure 4 illustrates the visibility supervision across views. The left hind paw is occluded by torso, which is conditioned by the visibility map (middle), resulting in the reduction of the keypoint probability. This prevents from influencing the occluded keypoint detection across views.

4.4. Label Supervision

We supervise the keypoint distribution and visibility map using a set of the labeled data as follows:

$$L_{\mathrm{L}}(\mathcal{I}_L) = \sum_{\mathbf{I}\in\mathcal{I}_L} D_{\mathrm{KL}}(\overline{P}_{\mathbf{I}_t^i} || P_t^i) + D_{\mathrm{KL}}(\overline{V}_{\mathbf{I}_t^i} || V_t^i), \qquad (9)$$

where $\overline{P}_{\mathbf{I}_t^i}$ and $\overline{V}_{\mathbf{I}_t^i}$ are the ground truth keypoint distribution and its visibility of image $\mathbf{I}_t^i$. The ground truth keypoint distribution is obtained by convolving a scaled Gaussian at the ground truth keypoint location. For the visibility map, it is computed via ray-casting on a discretized 3D voxel space. See Appendix for more details of ground truth visibility map generation.

4.5. Overall Loss

The resulting keypoint detector is learned using both labeled and unlabeled data by minimizing the following overall loss:

$$L(\mathbf{w}, \mathbf{w}_v) = L_{\mathrm{L}}(\mathcal{I}) + \lambda_{\mathrm{C}} \sum_{t=1}^{T} L_{\mathrm{C}}(\mathcal{I}_t) + \lambda_{\mathrm{T}} \sum_{i\in\mathcal{C}} L_{\mathrm{T}}(\mathcal{I}^i) + \lambda_{\mathrm{V}} \sum_{t=1}^{T} L_{\mathrm{V}}(\mathcal{I}_t), \qquad (10)$$

where L_{L}, L_{C}, and L_{T}, and L_{V} are the losses for the labeled supervision, cross-view supervision, temporal supervision, and visibility supervision, respectively. λ_C, λ_T and λ_V are the weights that control their importance.

[5] A similar observation has been made for long-term trajectory reconstruction [16].

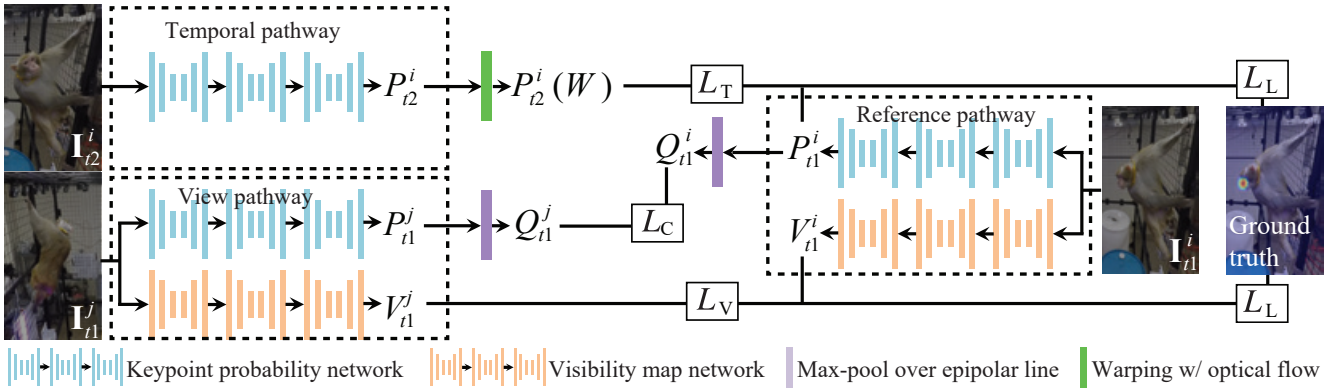


Figure 5: We design a network composed of three pathways: reference, temporal, and view pathways to utilize both labeled and unlabeled data. Each pathway is composed of two subnetworks producing keypoint distribution and visibility map. The labeled loss L_{L} is computed from the reference pathway by comparing to the ground truth annotation (keypoint and visibility) if available. The temporal and reference pathways measure the tracking loss L_{T} by warping the keypoint distribution using the dense optical flow ($P^i_{t_2}(W_{t_2 \to t_1})$), and the view and reference pathways measure the cross-view loss L_{C} by transforming the keypoint distribution to the epipolar plane distribution, i.e., $Q^i_{t_1} \leftrightarrow Q^j_{t_1}$.

5. Implementation

We design a network that is composed of three pathways: reference, view, and temporal pathways as shown in Figure 5. Each pathway takes as an input image with the size of $368 \times 368 \times 3$ and produce the keypoint probability and visibility map with the size of $46 \times 46 \times 21$. They all share the network weights $\mathbf{w}$ and $\mathbf{w}_v$. The reference and view pathways are designed to measure the cross-view loss L_{C} and visibility supervision loss L_{V} for two adjacent views by transforming to the epipolar plane distribution. The reference and temporal pathways measure the tracking loss L_{T} by warping the keypoint distribution using the dense optical flow. The label loss is measured for the reference pathway if the input image is labeled. We use the convolutional pose machine [7] as a base network to implement $\phi(\cdot)$ and $\psi(\cdot)$ while any existing pose detector can be complementary. See Appendix for network training. The code is publicly available: https://github.com/msbrpp/MSBR

Network Initialization by Bootstrapping To alleviate the noisy initialization of the detector, which occurs frequently when the unlabeled data dominate, we take a few practical steps. (1) With a subset of the labeled data in the same time instant, we triangulate the keypoint in 3D with RANSAC. This 3D keypoint is projected onto all multiview images, which can greatly augment the labeled data reliably. (2) Based on the 3D keypoints with volume estimation, we compute the visibility of labeled data using ray-casting, which provides the visibility map label for all views. (3) With the augmented labeled data with their visibility, we train the network in a fully supervised manner. This process is called bootstrapping [30], which provides a good initialization to train our triple network. (4) We retrain the pre-trained network with the unlabeled data with cross-view, tracking, and visibility losses.

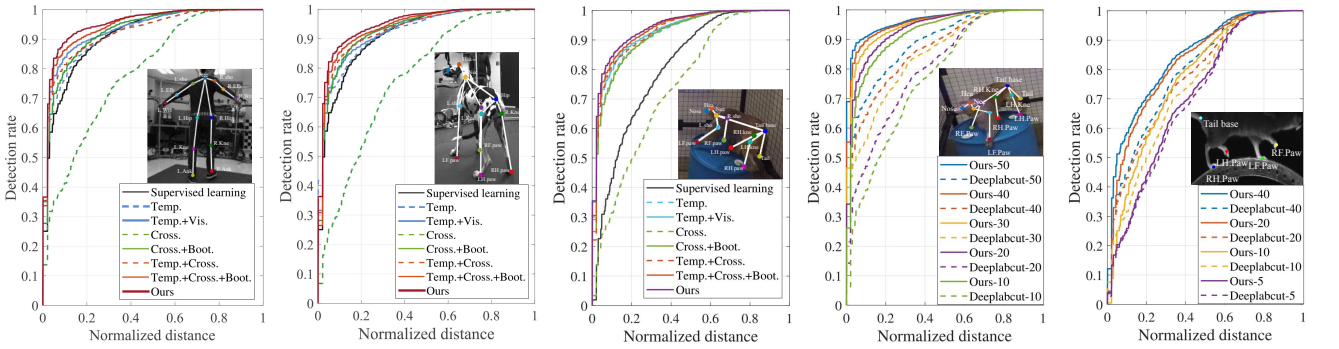
6. Experiments and Results

Datasets We evaluate our approach using realworld multiview image streams of non-human and human species without a pre-trained model captured by multi-camera systems. (1) **Monkey subject** 35 cameras running at 60 fps are installed in a large cage ($9' \times 12' \times 9'$) that allows the free-ranging behaviors of monkeys. There are diverse monkey activities include grooming, hanging, and walking. The camera produces 1280×960 images. The ground truth of keypoint and visibility is manually labeled. (2) **Dog subjects** Multi-camera system composed of 69 synchronized HD cameras (1024×1280 at 30 fps) are used to capture the behaviors of multiple breeds of dogs including Dalmatian and Golden Retrievers. The ground truth is manually labeled. (3) **Mouse subject** We use a multiview mouse locomotion dataset used to evaluate DeepLabCut [25]. A single camera with a mirror generates multiview synchronized images of a head-fixed mouse running on a treadmill. The scene is captured at 200 Hz and the keypoints are fully annotated manually[6]. (4) **Human subject I** A multiview behavioral imaging system composed of 69 synchronized HD cameras capture human activities at 30 fps with 1024×1280 resolution. We select 51 consecutive synchronized frames from 10 camera as training streams. Two end frames are used for the labeled data (20 images) and the rest images are used for the unlabeled data (490). The human pose detectors are used to triangulate the 3D pose to provide the ground truth. (5) **Human subject II** We test our approach on two publicly available datasets for human subjects: Panoptic Studio dataset [15] and Human3.6M [14]. For the Panoptic Studio dataset, we use 31 HD videos (1920×1080 at 30 Hz). The scenes includes diverse subjects with social interactions that introduce severe

[6]The data were prepared by Rick Warren in Sawtell lab [25].

	Human subject I					Dog subject					Monkey subject					
Method	Sho	Elb	Wri	Kne	**AUC**	Nec	F.Leg	Paw	H. Leg	**AUC**	Nec	F.Leg	Paw	Hip	H. Leg	**AUC**
Supervised learning	81.7	37.9	33.6	86.1	91.6	96.1	80.3	34.8	82.1	91.3	94.5	67.4	31.5	**96.9**	68.9	75.3
Temp.	86.4	44.6	32.5	93.4	91.7	94.2	83.2	31.6	83.3	92.0	94.2	82.8	37.4	90.3	83.7	87.4
Temp. + Vis.	92.7	48.4	41.1	97.8	93.3	96.9	91.5	38.1	88.9	92.5	94.9	87.4	45.8	91.6	87.9	89.2
Cross.	62.4	31.7	19.8	44.7	78.7	85.3	68.7	23.6	61.4	70.3	89.7	60.2	29.6	50.9	63.7	68.9
Cross. + Boot.	85.0	41.5	38.6	97.6	92.6	96.6	88.2	35.3	91.2	92.9	94.2	87.4	38.2	91.7	86.2	87.6
Temp. + Cross.	88.8	70.6	40.2	97.5	92.2	96.1	89.1	37.2	92.3	92.9	97.6	92.1	47.2	90.4	93.5	90.3
Temp. + Cross + Boot.	89.4	77.1	57.5	98.6	92.2	**98.9**	92.5	52.8	95.8	**93.8**	97.9	94.8	48.7	92.0	95.1	91.6
Ours	**92.9**	**77.2**	**65.4**	**98.9**	**95.1**	**98.9**	**94.2**	**53.2**	**95.8**	**94.8**	**98.7**	**95.2**	**50.1**	93.5	**95.7**	**92.2**

Table 1: We conduct an ablation study on human, dog, and monkey subjects using the PCKh measure.



(a) Human subject (b) Dog subject (c) Monkey subject (d) Monkey subject (e) Mouse subject

Figure 6: (a-c) We conduct ablation study using a PCK measure on human, dog, and monkey subjects. (d-e) We compare DeepLabCut (ResNet 50) [24] with ours on monkey and mouse subjects.

Method	Nec	F.Leg	Paw	Hip	H. Leg	**AUC**
Supervised learning	94.5	67.4	31.5	**96.9**	68.9	75.3
Simon et al. (argmax)	96.1	68.4	32.3	95.7	70.2	76.5
Dong et al. (soft-argmax)	85.7	32.9	10.6	87.2	37.8	59.6
Temporal sup. (flow warping)	94.2	82.8	37.4	90.3	83.7	87.4
Ours	**98.7**	**95.2**	**50.1**	93.5	**95.7**	**92.2**

Table 2: We compare our approach with Simon et al. [30] and Dong et al [8] on the monkey dataset.

social occlusion. The Human3.6M dataset is captured by 4 HD cameras that includes variety of single actor activities, e.g., sitting, running, and eating/drinking.

Metric We use a measure of the probability of correct keypoint (PCK) and PCKh that accounts for 50% of head length as a correct match. Area under curve (AUC) on PCK is also used to measure overall accuracy given fixed threshold.

Ablation Study We conduct ablation study to analyze the effect of each component in our network. (1) supervised learning with the labeled data; (2) semi-supervised learning with temporal supervision; (3) temporal supervision + visibility supervision; (4) cross-view supervision; (5) cross-view supervision + visibility supervision + bootstrapping; (6) cross-view supervision + temporal supervision; (7) cross-view supervision + temporal supervision + bootstrapping; (8) ours (cross-view supervision + temporal supervision + visibility supervision + bootstrapping). Except for the fully supervised learning, all network designs utilizes the unlabeled data.

Table 1 and Figure 6(a-c) summarize the result of ablation study on human, dog, and monkey subjects. Our approach achieves 95.1% on the Human dataset and 94.8% AUC on the Dog dataset, which outperforms the other 2 unsupervised baselines, temporal supervision and cross-supervision, by 3.4% and 16.4% AUC respectively on the Human dataset, and by 2.8% and 18.8% AUC on the Dog dataset. In addition, visibility probability improves temporal supervision by 1.8% AUC on the Human dataset and 2.65% AUC on the Dog dataset. Similarly, data augmentation improves cross-view supervision by 16.6% AUC on the Human dataset and 13.9% AUC on the Dog dataset.

Comparison with Soft-argmax We conduct an experiment to assess the performance of soft-argmax based approach. In Table 2, the soft-argmax approach (Dong et al.) is compared with our temporal supervision using dense flow warping on the monkey dataset. Our supervision approach significantly outperforms the soft-argmax with large margin (27.8%), which is also verified in Yao et al. [37]. The soft-argmax leads to highly biased keypoint coordinate when the prediction is spurious due to the nature of weighted average.

Comparison with Semi-supervised Learning We compare our approach with existing semi-supervised learning frameworks that use (1) temporal supervision [8] and (2) cross-view supervision [37] on two publicly available human subject datasets (Panoptic Studio and Human3.6M). No pre-trained model is used for the comparison.

Table 3 summarizes the PCKh measure of methods including fully supervised learning with the labeled data. Leveraging semi-supervised learning enhances the detection accuracy (there exists significant performance degradation of cross-view supervision due to long interval between the annotated frames). This shows that our approach leverages the unlabeled data better through the tight integration of temporal and cross-view supervisions. Also we test the generalizability of the trained pose detector by applying to

	Unlabeled data detection								Unseen data detection							
Panoptic Studio dataset	Nec	Sho	Elb	Wri	Hip	Kne	Ank	**AUC**	Nec	Sho	Elb	Wri	Hip	Kne	Ank	**AUC**
Supervised learning	93.5	78.2	36.8	28.6	98.7	83.5	92.4	88.5	94.2	75.4	32.9	23.6	97.2	78.6	89.4	85.5
Temp.	98.1	88.3	43.6	33.5	97.8	92.7	96.6	92.3	96.7	80.7	37.8	28.2	**97.8**	86.2	92.7	90.1
Yao et al. [37]	98.6	68.2	38.3	23.5	28.9	45.2	69.2	72.5	93.6	64.5	35.8	24.5	34.9	42.8	70.2	70.8
Ours	**98.8**	**93.1**	**78.5**	**66.8**	**98.5**	**98.3**	**98.9**	**95.6**	**97.2**	**88.3**	**68.3**	**52.4**	97.6	**89.3**	**94.7**	**91.4**
Human3.6M	Nec	Sho	Elb	Wri	Hip	Kne	Ank	**AUC**	Nec	Sho	Elb	Wri	Hip	Kne	Ank	**AUC**
Supervised learning	92.1	75.3	41.8	26.5	93.7	82.5	90.4	86.2	90.1	76.3	38.9	20.8	93.8	78.6	83.2	84.8
Temp.	95.4	88.6	46.5	35.2	96.5	95.6	95.2	91.6	91.7	81.4	42.3	25.6	**93.9**	83.4	87.5	86.9
Yao et al. [37]	95.8	50.8	31.5	18.5	32.6	40.8	65.3	69.9	89.6	48.3	29.7	20.5	29.8	34.9	60.7	65.2
Ours	**97.9**	**92.5**	**76.7**	**64.3**	**97.2**	**97.6**	**96.9**	**94.8**	**93.2**	**92.8**	**67.3**	**49.6**	93.7	**87.6**	**89.5**	**88.7**

Table 3: We compare our approach with existing semi-supervised learning frameworks: (1) temporal supervision and (2) cross-view supervision [37]. We evaluate on two public human datasets (Panoptic Studio and Human3.6M) using PCKh measure. We test the generalizability by applying on unseen subjects.

Monkey subject																
	DeepLabCut [24]								Ours							
# annotations	Nose	Hea	Nec	F.Leg	Paw	Hip	H. Leg	**AUC**	Nose	Hea	Nec	F.Leg	Paw	Hip	H. Leg	**AUC**
10	92.1	93.5	90.6	59.4	28.2	**97.3**	63.2	73.9	**93.2**	94.6	**91.4**	**83.2**	**43.9**	92.1	**85.5**	**89.1**
20	**95.9**	95.7	95.2	68.3	30.8	**98.3**	70.1	78.7	95.1	**99.3**	**98.7**	**95.2**	**50.1**	93.5	**95.7**	**92.2**
30	95.3	95.8	96.7	73.7	33.2	**98.5**	75.6	80.3	**95.4**	**99.1**	**98.5**	**95.9**	**54.8**	95.7	**96.0**	**93.8**
40	96.5	96.2	96.8	77.8	39.7	**97.9**	78.7	83.8	96.5	**99.5**	**99.2**	**96.3**	**55.7**	94.8	**96.3**	**95.3**
50	96.5	96.5	97.1	81.9	42.6	**98.3**	82.3	85.4	**96.6**	**99.4**	**99.0**	**96.4**	**56.3**	95.1	**96.7**	**96.2**
Mouse subject																
	DeepLabCut [24]								Ours							
# annotations	LF. paw	LH. paw	Tail	RF. paw	RH. paw	**MAE**	**RMSE**	**AUC**	LF. paw	LH. paw	Tail	RF. paw	RH. paw	**MAE**	**RMSE**	**AUC**
5	51.1	53.7	73.1	51.3	53.3	6.7	8.7	63.5	**57.6**	**58.5**	**76.6**	**57.9**	**58.1**	**6.1**	**8.4**	**65.7**
10	60.0	61.9	78.5	60.6	61.1	5.8	7.9	69.8	**68.4**	**69.5**	**82.8**	**67.8**	**69.8**	**4.9**	**7.3**	**73.6**
20	64.5	65.2	80.7	64.9	66.4	5.4	7.7	74.2	**73.9**	**75.6**	**85.4**	**74.5**	**75.0**	**4.4**	**6.5**	**79.5**
40	67.3	67.1	82.1	66.7	67.3	5.0	7.5	75.9	**78.8**	**79.0**	**87.9**	**78.4**	**79.2**	**3.9**	**5.9**	**81.4**

Table 4: We compare our approach with DeepLabCut [24] that leverages a pre-trained model as varying the number of annotations. RMSE and MAE are measured in term of confidence map size (46×46).

Methods	Nec	Sho	Elb	Wri	Hip	Kne	Ank	**AUC**
Simon et al.	92.3	82.2	43.5	35.4	91.6	85.3	89.2	90.3
Kocabas et al.	97.6	87.3	42.7	30.6	84.2	91.1	90.3	88.5
Rhodin et al.	**98.5**	85.2	56.6	42.1	**97.8**	90.4	91.7	91.9
Ours	97.9	**92.5**	**76.7**	**64.3**	97.2	**97.2**	**96.9**	**94.8**

Table 5: We compare our approach with three baselines: (1) Simon et al. [30], (2) Kocabas et al. [18], and (3) Rhodin et al. [28] on Human3.6 dataset.

the unseen subjects who are not used as unlabeled data. For Panoptic Studio, Dance 1 is used for the labeled and unlabeled and Dance 2 is used for the unseen data, and for Human3.6M, Eating and Discussion are used for the labeled and unlabeled data, and Greeting is used for the unseen data. The trend is similar to the unlabeled data, i.e., our approach shows stronger generalization power.

Comparison with DeepLabCut We compare our approach with DeepLabCut [24] that leverages a pre-trained model (ResNet 50 [12] trained on ImageNet [29]). In particular, we focus on non-human subjects (monkeys and mice) to reflect the strength of DeepLabCut. Two datasets differ in range of motion. For the mouse locomotion, the head of the mouse is stabilized where the range of motion is restricted to leg motion on the treadmill. On the other hand, the monkey activities are completely unconstrained, which produces severe self-occlusion and pose variation.

Table 4 and Figure 6(d-e) summarize the performance comparison with respect to the number of annotations. A notable difference is that the performance gap of the monkey activities is much higher than that of the mouse, e.g., for 10 annotated data, our approach outperforms 15% for the monkey and 3.5% for the mouse. This indicates that our approach is more resilient to large appearance change induced by free-ranging activities.

Qualitative Evaluation We show the qualitative result in Figure 1. See Appendix and Supplementary Video for extensive qualitative result.

7. Summary

We present a new semi-supervised learning framework to train a keypoint detector from multiview image streams. We integrate three self-supervisionary signals to effectively utilize a large amount of the unlabeled multiview data: (1) the cross-view supervision that enforces geometric consistency through the epipolar constraint across views; (2) the temporal supervision that constrains keypoint detection to be in accordance with dense optical flow; and (3) the visibility supervision that validates the detected keypoint in the presence of severe self-occlusion. We embed these supervisions into a new network design composed of three pathways in a differentiable fashion, allowing end-to-end training. We demonstrate that our approach outperforms existing semi-supervised learning approaches [8, 37] and DeepLabCut [24] that uses a pre-trained model. The resulting network precisely detects the keypoints of both non-human and human subjects with highly limited labeled data ($< 4\%$).

8. Acknowledgements

This work is supported by NSF IIS 1846031 and NSF IIS 1755895.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 1

[2] I. Arev, H. S. Park, Y. Sheikh, J. K. Hodgins, and A. Shamir. Automatic editing of footage from multiple social cameras. *SIGGRAPH*, 2014. 3

[3] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *IJCV*, 2004. 2, 4

[4] G. Bertasius, S. X. Yu, H. S. Park, and J. Shi. Exploiting visual-spatial first-person co-occurrence for action-object detection without labels. In *ICCV*, 2017. 2, 4

[5] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui. Visual object tracking using adaptive correlation filters. In *CVPR*, 2010. 2

[6] A. Byravan and D. Fox. SE3-nets: Learning rigid body motion using deep neural networks. In *ICRA*, 2016. 2, 4

[7] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *CVPR*, 2016. 2, 6

[8] X. Dong, S.-I. Yu, X. Weng, S.-E. Wei, Y. Yang, and Y. Sheikh. Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In *CVPR*, 2018. 2, 4, 5, 7, 8

[9] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *ACM Comm.*, 1981. 5

[10] Y. Furukawa and J. Ponce. Dense 3d motion capture from synchronized video streams. In *CVPR*, 2008. 2

[11] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004. 2, 3

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 8

[13] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *TPAMI*, 2015. 2, 4

[14] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 2014. 6

[15] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015. 6

[16] H. Joo, H. S. Park, and Y. Sheikh. Map visibility estimation for large-scale dynamic 3d reconstruction. In *CVPR*, 2014. 2, 3, 5

[17] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2, 3, 4

[18] M. Kocabas, S. Karagoz, and E. Akbas. Self-supervised learning of 3d human pose using multi-view geometry. *arXiv*, 2019. 8

[19] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 1951. 4

[20] C.-H. Lin and S. Lucey. Inverse compositional spatial transformer networks. In *CVPR*, 2017. 4

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollàr, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 5

[22] H. Liu, J. Lu, J. Feng, and J. Zhou. Two-stream transformer networks for video-based face alignment. *TPAMI*, 2018. 2

[23] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 1981. 2, 3

[24] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 2018. 1, 7, 8

[25] A. Mathis and R. A. Warren. On the inference speed and video-compression robustness of deeplabcut. In *bioRxiv*, 2018. 6

[26] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 2

[27] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas. A recurrent encoder-decoder network for sequential face alignment. In *ECCV*, 2016. 2

[28] H. Rhodin, M. Salzmann, and P. Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *ECCV*, 2018. 8

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 1, 8

[30] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 2, 4, 5, 6, 7, 8

[31] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014. 2

[32] S. Tulsiani, T. Zhou, A. A. Efros, and J. Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, 2017. 2, 3, 4

[33] M. Velliste, S. Perel, M. C. Spalding, A. S. Whitford, and A. B. Schwartz. Cortical control of a prosthetic arm for self-feeding. *Nature*, 2008. 1

[34] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki. Sfm-net: Learning of structure and motion from video. In *arXiv*, 2017. 2, 4

[35] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016. 2

[36] A. G. Xiaolong Wang. Unsupervised learning of visual representations using videos. In *ICCV*, 2015. 2, 5

[37] Y. Yao, Y. Jafarian, and H. S. Park. Monet: Multiview semi-supervised keypoint via epipolar divergence. In *ICCV*, 2019. 2, 3, 4, 7, 8

[38] J. S. Yoon, Z. Li, and H. S. Park. 3d semantic trajectory reconstruction from 3d pixel continuum. In *CVPR*, 2017. 2, 3

[39] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 2, 3, 4