Supplementary Material

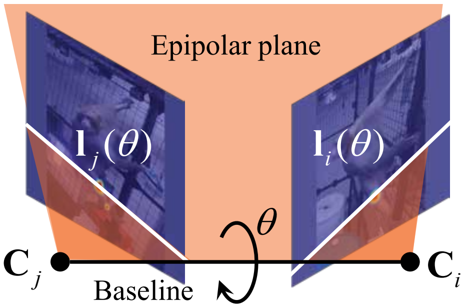

Figure 11: Two epipolar lines are induced by an epipolar plane, which can be parametrized by the rotation θ about the baseline where $\mathbf{C}_i$ and $\mathbf{C}_j$ are the camera optical centers.

A. Proof of Theorem 1

Proof. A point in an image corresponds to a 3D ray $\mathbf{L}$ emitted from the camera optical center $\mathbf{C}$ (i.e., inverse projection), and λ corresponds to the depth. $\mathbf{K}$ is the intrinsic parameter. The geometric consistency, or zero reprojection error, is equivalent to proving $\mathbf{L}_i^*, \mathbf{L}_j^* \in \mathbf{\Pi}$ where $\mathbf{\Pi}$ is an epipolar plane rotating about the camera baseline $\overline{\mathbf{C}_i\mathbf{C}_j}$ as shown in Figure 11, and $\mathbf{L}_i^*$ and $\mathbf{L}_j^*$ are the 3D rays produced by the inverse projection of correspondences $\mathbf{x}_i^* \leftrightarrow \mathbf{x}_j^*$, respectively, i.e., $\mathbf{L}_i^* = \mathbf{C}_i + \lambda \mathbf{R}_i^\mathsf{T}\mathbf{K}^{-1}\widetilde{\mathbf{x}}_i^*$. The correspondence from the keypoint distributions are:

$$\mathbf{x}_i^* = \underset{\mathbf{x}}{\operatorname{argmax}}\; P_i(\mathbf{x}) \tag{12}$$

$$\mathbf{x}_j^* = \underset{\mathbf{x}}{\operatorname{argmax}}\; P_j(\mathbf{x}), \tag{13}$$

$Q_i(\theta) = Q_{j\to i}(\theta)$ implies:

$$\begin{aligned}\theta^* &= \underset{\theta}{\operatorname{argmax}} \sup_{\mathbf{x}\in \mathbf{l}_i(\theta)} P_i(\mathbf{x}) \\ &= \underset{\theta}{\operatorname{argmax}} \sup_{\mathbf{x}\in \mathbf{l}_i(\theta)} P_{j\to i}(\mathbf{x}) \\ &= \underset{\theta}{\operatorname{argmax}} \sup_{\mathbf{x}\in \mathbf{l}_j(\theta)} P_j(\mathbf{x}).\end{aligned} \tag{14}$$

This indicates the correspondence lies in epipolar lines induced by the same θ^*, i.e,. $\mathbf{x}_i^* \in \mathbf{l}_i(\theta^*)$ and $\mathbf{x}_j^* \in \mathbf{l}_j(\theta^*)$. Since $\mathbf{l}_j(\theta^*) = \mathbf{F}\widetilde{\mathbf{x}}_i^*$, $\mathbf{l}_i(\theta^*)$ and $\mathbf{l}_j(\theta^*)$ are the corresponding epipolar lines. Therefore, they are in the same epipolar plane, and the reprojection error is zero. □

B. Cropped Image Correction and Stereo Rectification

We warp the keypoint distribution using stereo rectification. This requires a composite of transformations because the rectification is defined in the full original image. The transformation can be written as:

$${}^{\bar h}\mathbf{H}_h = \left({}^{\bar h}\mathbf{H}_{\bar c}\right)\left({}^{\bar c}\mathbf{H}_{\bar b}\right)\mathbf{H}_r \left({}^{c}\mathbf{H}_b\right)^{-1}\left({}^{h}\mathbf{H}_c\right)^{-1}. \tag{15}$$

The sequence of transformations takes a keypoint distribution of the network output P to the rectified keypoint distribution $\overline{P}$: heatmap→cropped image→original image→rectified image→rectified cropped image→rectified heatmap.

Given an image $\mathcal{I}$, we crop the image based on the bounding box as shown in Figure 12: the left-top corner is (u_x, u_y) and the height is h_b. The transformation from the image to the bounding box is:

$${}^{c}\mathbf{H}_b = \begin{bmatrix} s & 0 & w_x - su_x \\ 0 & s & w_y - su_y \\ 0 & 0 & 1 \end{bmatrix} \tag{16}$$

where $s = h_c/h_b$, and (w_x, w_y) is the offset of the cropped image. It corrects the aspect ratio factor. $h_c = 364$ is the height of the cropped image, which is the input to the network. The output resolution (heatmap) is often different from the input, $s_h = h_h/h_c \neq 1$, where h_h is the height of the heatmap. The transformation from the cropped image to the heatmap is:

$${}^{h}\mathbf{H}_c = \begin{bmatrix} s_h & 0 & 0 \\ 0 & s_h & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{17}$$

The rectified transformations $\left({}^{\bar h}\mathbf{H}_{\bar c}\right)$ and $\left({}^{\bar c}\mathbf{H}_{\bar b}\right)$ can be defined in a similar way.

The rectification homography can be computed as $\mathbf{H}_r = \mathbf{K}\mathbf{R}_n\mathbf{R}^\mathsf{T}\mathbf{K}^{-1}$ where $\mathbf{K}$ and $\mathbf{R} \in SO(3)$ are the intrinsic parameter and 3D rotation matrix and $\mathbf{R}_n$ is the rectified rotation of which x-axis is aligned with the epipole, i.e., $\mathbf{r}_x = \dfrac{\mathbf{C}_j - \mathbf{C}_i}{\|\mathbf{C}_j - \mathbf{C}_i\|}$ where $\mathbf{R}_n = \begin{bmatrix}\mathbf{r}_x^\mathsf{T} \\ \mathbf{r}_y^\mathsf{T} \\ \mathbf{r}_z^\mathsf{T}\end{bmatrix}$ and other axes can be computed by the Gram-Schmidt process.

The fundamental matrix between two rectified keypoint distributions $\overline{P}_i$ and $\overline{P}_j$ can be written as:

$$\begin{aligned}\mathbf{F} &= \mathbf{K}_j^{-\mathsf{T}} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}_\times \mathbf{K}_i^{-1} \\ &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1/f_y^j \\ 0 & 1/f_y^i & p_y^j/f_y^j - p_y^i/f_y^i \end{bmatrix}\end{aligned} \tag{18}$$

where $[\cdot]_\times$ is the skew symmetric representation of cross product, and

$$\mathbf{K}_i = \begin{bmatrix} f_x^i & 0 & p_x^i \\ 0 & f_y^i & p_y^i \\ 0 & 0 & 1 \end{bmatrix}. \tag{19}$$

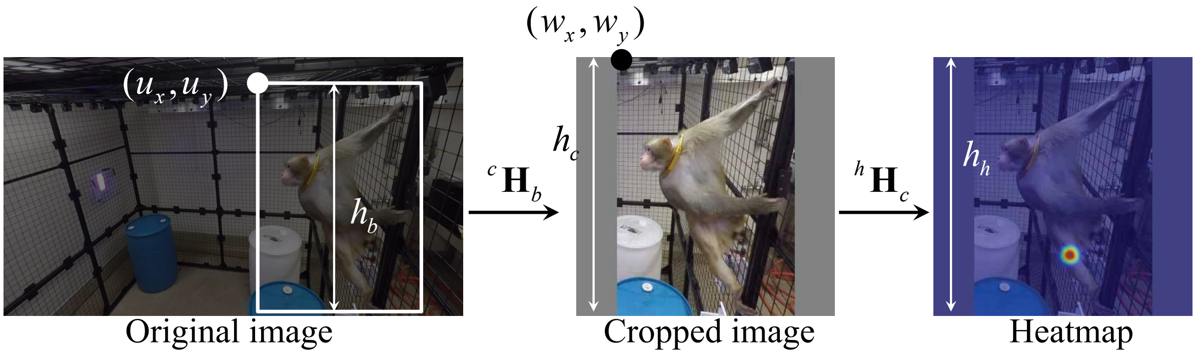

Figure 12: A cropped image is an input to the network where the output is the keypoint distribution. To rectify the keypoint distribution (heatmap), a series of image transformations need to be applied.

Subjects	P	$\lvert\mathcal{D}_L\rvert$	$\lvert\mathcal{D}_U\rvert$	$\lvert\mathcal{D}_L\rvert/\lvert\mathcal{D}_U\rvert$	C	FPS	Camera type
Monkey	13	85	63,000	0.13%	35	60	GoPro 5
Humans	14	30	20,700	0.14%	69	30	FLIR BlackFly S
Dog I	12	100	138,000	0.07%	69	30	FLIR BlackFly S
Dog II	12	75	103,500	0.07%	69	30	FLIR BlackFly S
Dog III	12	80	110,400	0.07%	69	30	FLIR BlackFly S
Dog IV	12	75	103,500	0.07%	69	30	FLIR BlackFly S

Table 3: Summary of multi-camera dataset where P is the number of keypoints, C is the number of cameras, $|\mathcal{D}_L|$ is the number of labeled data, and $|\mathcal{D}_U|$ is the number of unlabeled data.

This allows us to derive the re-scaling factor of a and b in Equation (7):

$$a = \frac{s^i f_y^i}{s^j f_y^j} \tag{20}$$

$$b = s_h s^i \left((\overline{u}_y^j - p_y^j) \frac{f_y^i}{f_y^j} + p_y^i - \overline{u}_y^i \right) \tag{21}$$

where $\overline{u}_y^i$ is the bounding box offset of the rectified coordinate.

C. Evaluation Dataset

All cameras are synchronized and calibrated using structure from motion [18]. The input of most pose detector models except for [8] is a cropped image containing a subject, which requires specifying a bounding box. We use a kernelized correlation filter [20] to reliably track a bounding box using multiview image streams given initialized 3D bounding box from the labeled data.