Supplementary Materials: Epipolar Transformer for Multi-view Pose Estimation

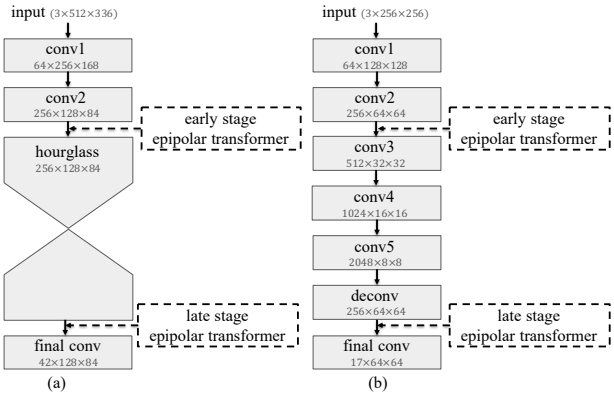


Figure 8: Early stage or late stage where we can add epipolar transformer to the backbone model. (a) Hourglass networks [23] on InterHand. (b) ResNet-50 detector [37] on Human3.6M [13].

A. Dealing with Image Transformations

As the epipolar transformer relies on camera calibration, any spatial transformation applied on the image needs to be reflected in the calibration parameters.

Data augmentation: Data augmentation like rotation, scaling and cropping can still be performed with the epipolar transformer. The projection matrix needs to be updated accordingly when the image is transformed with an affine transformation parameterized by $A \in \mathbb{R}^{2\times 2}$ and $b \in \mathbb{R}^2$:

$$M := \begin{bmatrix} A & b \\ \mathbf{0}^T & 1 \end{bmatrix} M \tag{4}$$

Different scaling and cropping parameters can be applied separately to the reference view and source view.

Scaling of projection matrices: We should pay special attention to scale the projection matrices for resizing or pooling images.

Suppose the input image is spatially down-sampled s_x and s_y times along the x-axis and y-axis, the projection matrix is updated as follows:

$$M := \begin{bmatrix} 1/s_x & 0 & (1-s_x)/2s_x \\ 0 & 1/s_y & (1-s_y)/2s_y \\ 0 & 0 & 1 \end{bmatrix} M \tag{5}$$

The coordinates are aligned with the center of pixels rather than the top-left corners, which is important for extracting features at precise locations in the epipolar transformer.

B. 2D Prediction Visualization in Video

In skeletons_1min.mp4, we visualize 2D predicted skeletons on Human3.6M [13] testing set. All methods are with ResNet-50 [12] and image size 256×256, corresponding to the following three entries in Table 6 respecitively:

1. R50 256×256 + triangulate
2. R50 256×256 + crossview[28] + triangulate
3. R50 256×256 + ours + triangulate

In twohand_30.mp4, we visualize 2D predicted hands on InterHand testing set. The video consists of the visualizations of the ground truth, the 5.46 mm baseline and 4.91 mm model with epipolar transformer in Table 3.

C. Stages to Add Epipolar Transformer

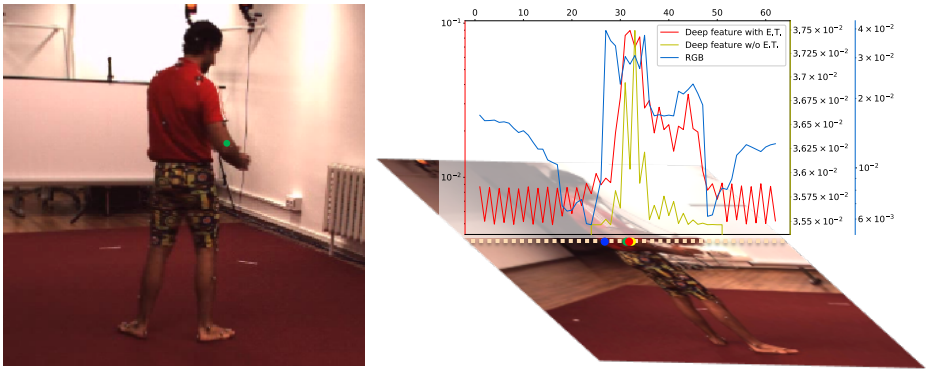
In our experiments, we compared the performance between adding epipolar transformer in the early stage and adding in the late stage (see Table 2, paragraph **Which stage to insert the epipolar transformer**). Figure 8 illustrates the precise places of inserting the epipolar transformer for the "early" and "late" settings in an one-stage Hourglass network [23] on InterHand and ResNet-50 [12] simple baseline [37] on Human3.6M [13] respectively.

D. Epipolar Transformer Visualization on Human3.6M [13]

We show the visualizations of feature matching similarity for images from Human3.6M [13] in Figure 9, Figure 10, and Figure 11. In Figure 9, we show a visualization of the matching results of an easy case in Human3.6M [13]. Note that in this case, our prediction aligns well with the ground truth. Using the features from the baseline, that is, deep features extracted without using the epipolar transformer, the matched point denoted by the yellow dot is also accurate. However, for more difficult cases in Figure 10 and Figure 11, the deep features learned with the epipolar transformer can perform more accurate matching.

E. Epipolar Transformer Visualization on InterHand

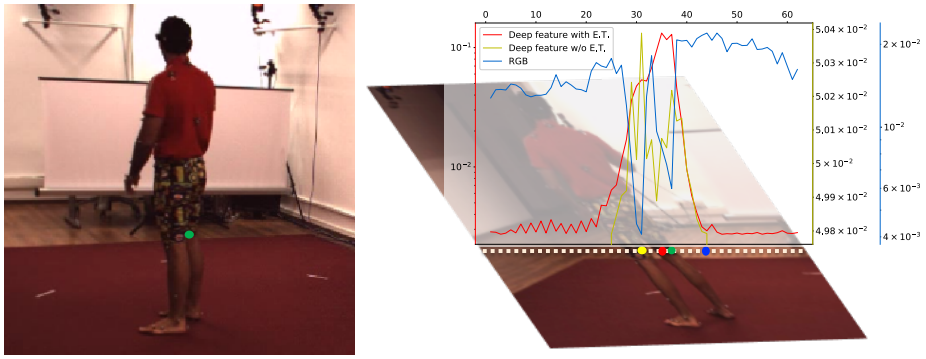
We show visualization of feature matching similarity for images from InterHand. As shown in Figure 12, the color features have multiple peaks in similarity due to the different fingers having similar colors. On the other hand, for most cases deep features trained through the epipolar transformer are able to discriminate the correct finger from all the other similar looking fingers.



Reference view Source view

Right elbow selected, denoted in green.

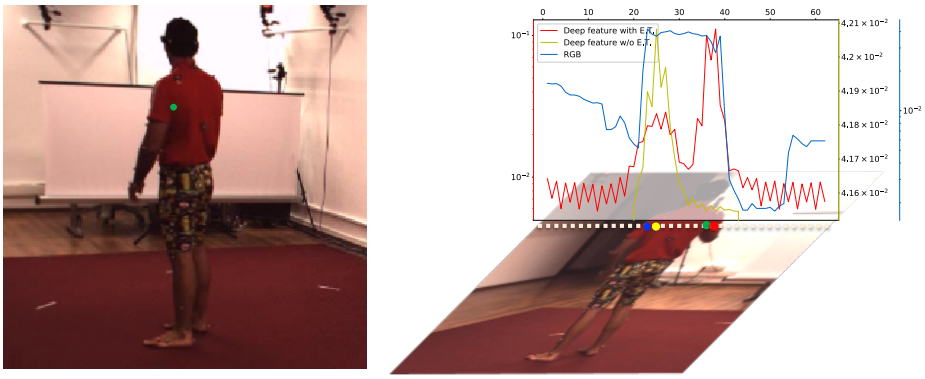
Figure 9: Visualizations of the matching results along the epipolar line in an **easy case** in Human3.6M [13]. We here use E.T. as a shorthand for epipolar transformer. The compared features are (a) deep features learned through the epipolar transformer (deep features with E.T., denoted in red), (b) deep feature learned by ResNet-50[12] without epipolar transformer (deep features w/o E.T., denoted in yellow), and (c) RGB features (denoted in blue). Green dot on the reference view is the selected joint, and the green dot on the source view is the corresponding point offered by the groundtruth.



Reference view Source view

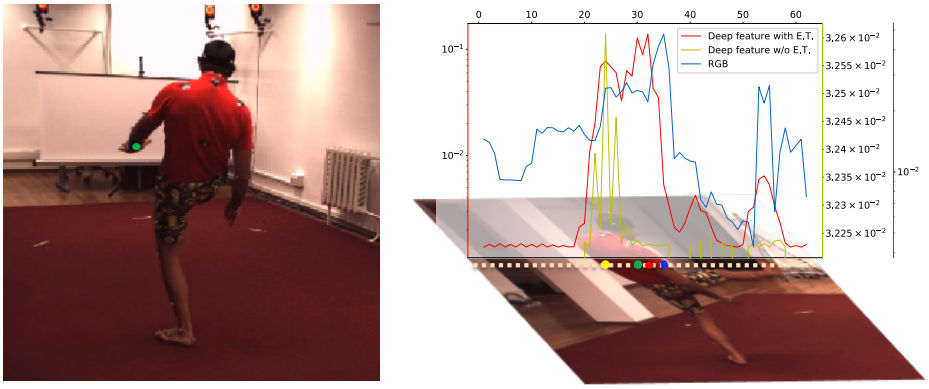
(i) Right knee selected, denoted in green.

Figure 10: Visualizations of the matching results along the epipolar line in a **more difficult case** in Human3.6M [13]. We here use E.T. as a shorthand for Epipolar Transformer. The compared features are (a) deep features learned through the epipolar transformer (deep features with E.T., denoted in red), (b) deep feature learned by ResNet-50[12] without epipolar transformer (deep features w/o E.T., denoted in yellow), and (c) RGB features (denoted in blue). Green dot on the reference view is the selected joint, and the green dot on the source view is the corresponding point offered by the groundtruth.



Reference view Source view

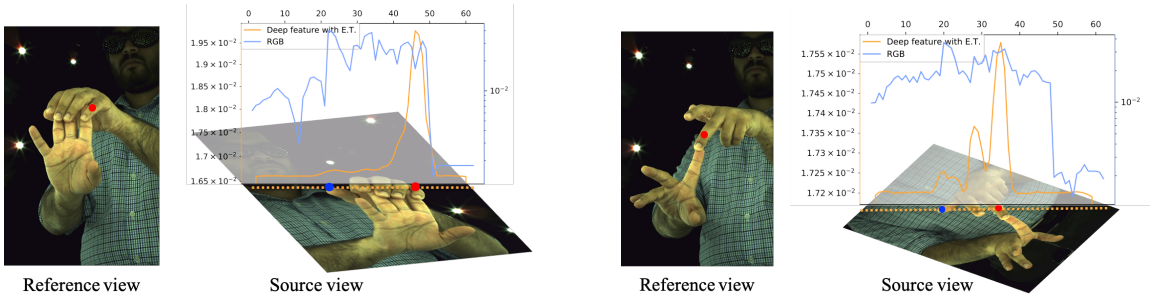
(ii) Right wrist selected, denoted in green.



Reference view Source view

(iii) Left wrist selected, denoted in green.

Figure 11: (Cont')Visualizations of the matching results along the epipolar line in **more difficult cases** in Human3.6M [13]. We here use E.T. as a shorthand for Epipolar Transformer. The compared features are (a) deep features learned through the epipolar transformer (deep features with E.T., denoted in red), (b) deep feature learned by ResNet-50[12] without epipolar transformer (deep features w/o E.T., denoted in yellow), and (c) RGB features (denoted in blue). Green dot on the reference view is the selected joint, and the green dot on the source view is the corresponding point offered by the groundtruth.



Reference view Source view Reference view Source view

Figure 12: Comparison of feature-matching results along the epipolar line. The compared features are color features and the deep features learned through the epipolar transformer (deep features with E.T.). The best matches of the epipolar transformer (red) and RGB (blue) are shown on the epipolar lines. The similarity distributions along the epipolar lines are also shown.