Supplementary Material

Figure 8: A cropped image is an input to the network where the output is the segmentation distribution with the same size. To rectify the segmentation distribution (heatmap), a series of image transformations need to be applied

A. Cropped Image Correction and Stereo Rectification

We warp the segmentation 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:

$$ {}^{\overline{h}}\mathbf{H}_h = \left({}^{\overline{c}}\mathbf{H}_{\overline{b}}\right) \mathbf{H}_r \left({}^{c}\mathbf{H}_b\right)^{-1}. \tag{11} $$

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

Given an image $\mathcal{I}$, we crop the image based on the bounding box as shown in Fig. 8: 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:

$$ {}^{c}\mathbf{H}_b = \begin{bmatrix} s_x & 0 & -s_x u_x \\ 0 & s_y & -s_y u_y \\ 0 & 0 & 1 \end{bmatrix} \tag{12} $$

where $s_x = h_c/h_{bx}$ and $s_y = h_c/h_{by}$. It corrects the aspect ratio factor. $h_c = 200$ is the width and height of the cropped image, which is the input to the network. The network output have the same resolution as the input. The rectified transformations $\left({}^{\overline{c}}\mathbf{H}_{\overline{b}}\right)$ can be defined in a similar way.

Given the cropping factors, we derive v_i and the rescaling factor of a_i and b_i in the following Equation in Section 3.3:

$$ \overline{\xi}_i(a_i u + b_i; \mathbf{L_x}) = \overline{P}_i \left(\begin{bmatrix} a_i u + b_i \\ v_i \end{bmatrix} \right), $$

where v_i is the y coordinate of the rectified image that corresponds to (u_1, v_1). v_i can be computed by transforming (u_1, v_1) to the i^{th} rectified coordinate:

$$ \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \left({}^{\overline{c_i}}\mathbf{H}_{\overline{b_i}}\right) {}^{r_i}\mathbf{H}_o \left({}^{r_1}\mathbf{H}_o\right)^{-1} \left({}^{c_1}\mathbf{H}_{b_1}\right)^{-1} \begin{bmatrix} u_1 \\ v_1 \\ 1 \end{bmatrix}, $$

where ${}^{r_i}\mathbf{H}_o$ is the homography that rectifies the original target view with respect to the i^{th} camera. The point in the first cropped and rectified image (u_1, v_1) is transformed to (u_i, v_i).

For a_i and b_i,

$$ a_i = \frac{W_1 \cos\theta}{W_i} \tag{13} $$

$$ b_i = -\frac{W_1 u_o^1 \cos\theta}{W_i} + u_o^i, \tag{14} $$

where $\theta = \cos^{-1} \frac{trace(\mathbf{R}_o^{1\to i})-1}{2}$. W_i is the distance (baseline) between the i^{th} camera and target camera. u_o^i is the point in target view rectified with respect to i^{th} view. $\mathbf{R}_o^{1\to i}$ is the difference between the rotations of target view rectified with respect to the first view and i^{th} view.

B. Qualitative Result

We validate our semi-supervised semantic segmentation framework using three real-world datasets: monkey, dancer and the subjects in social videos. In the social event videos, a group of dancers were performing Hip-hop dance, and they were surround by the audiences holding hand-held cameras. An Indian dancer was performing solo dance captured by 69 cameras in three layers with different heights. One monkey was crawling against the cage in the video, and the array of cameras were placed in the cage ceiling.

Figs. 9–16 shows the prediction results on the unlabeled data using three models. Figs. 17–24 shows how the semi-supervised network progresses on the unlabeled data during the training. Fig. 26 shows some failure cases of our segmentation framework.

One possible reason of the failures on social data is that since the people who hand-held the cameras were walking around when they captured the videos, the synchronization is not accurate enough; therefore the camera rotation and location data is very noisy, which causes the shape belief transfer incorrect. Other reasons for failures can be that there are no enough source image pairs which are able to construct tight upper bound for subjects. Or the weight on prior is too small to affect the predictions.

Figure 9: Qualitative result of multiview segmentation.

Figure 10: Qualitative result of multiview segmentation.

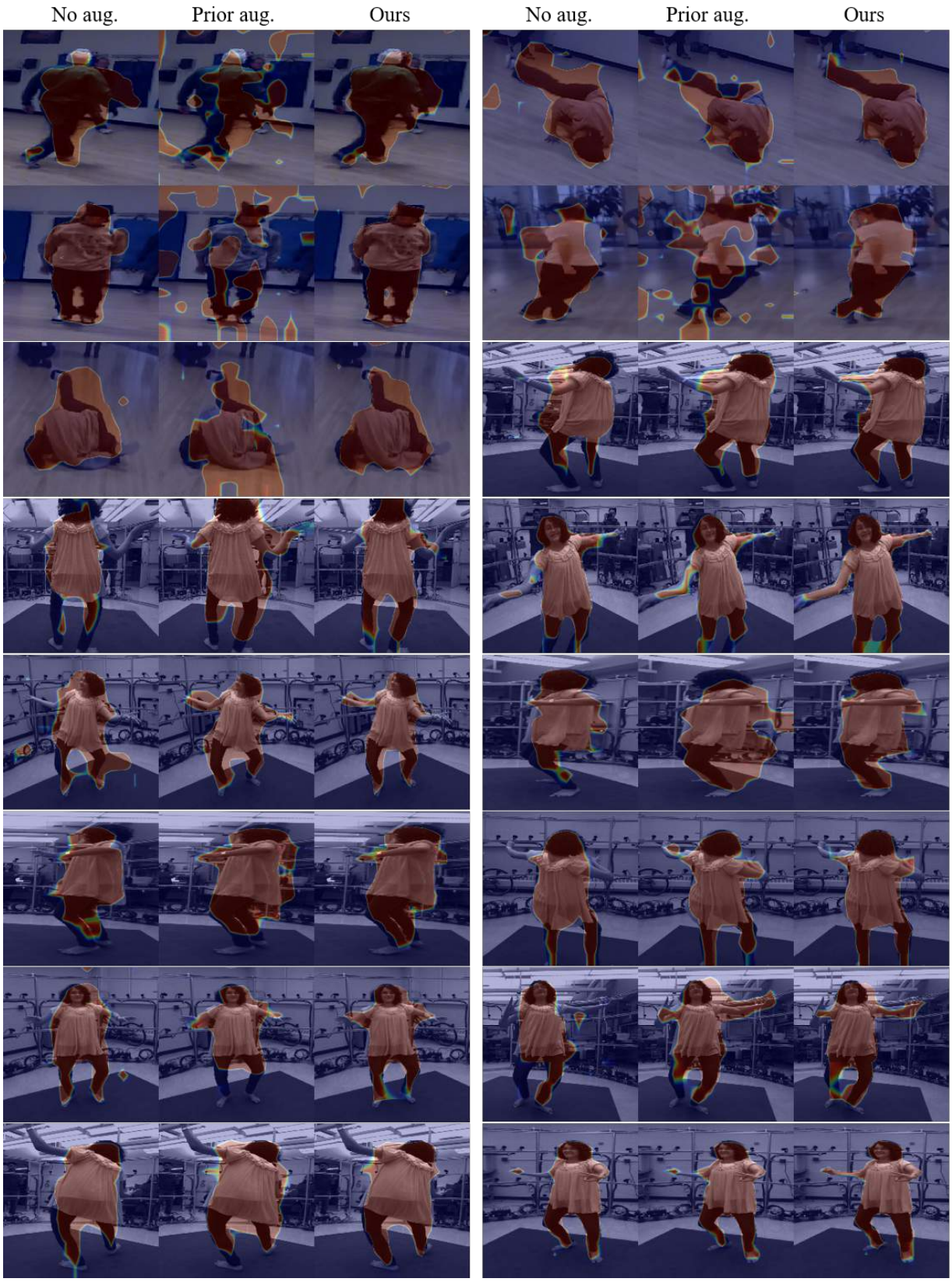

Figure 11: Qualitative result of multiview segmentation.

Figure 12: Qualitative result of multiview segmentation.

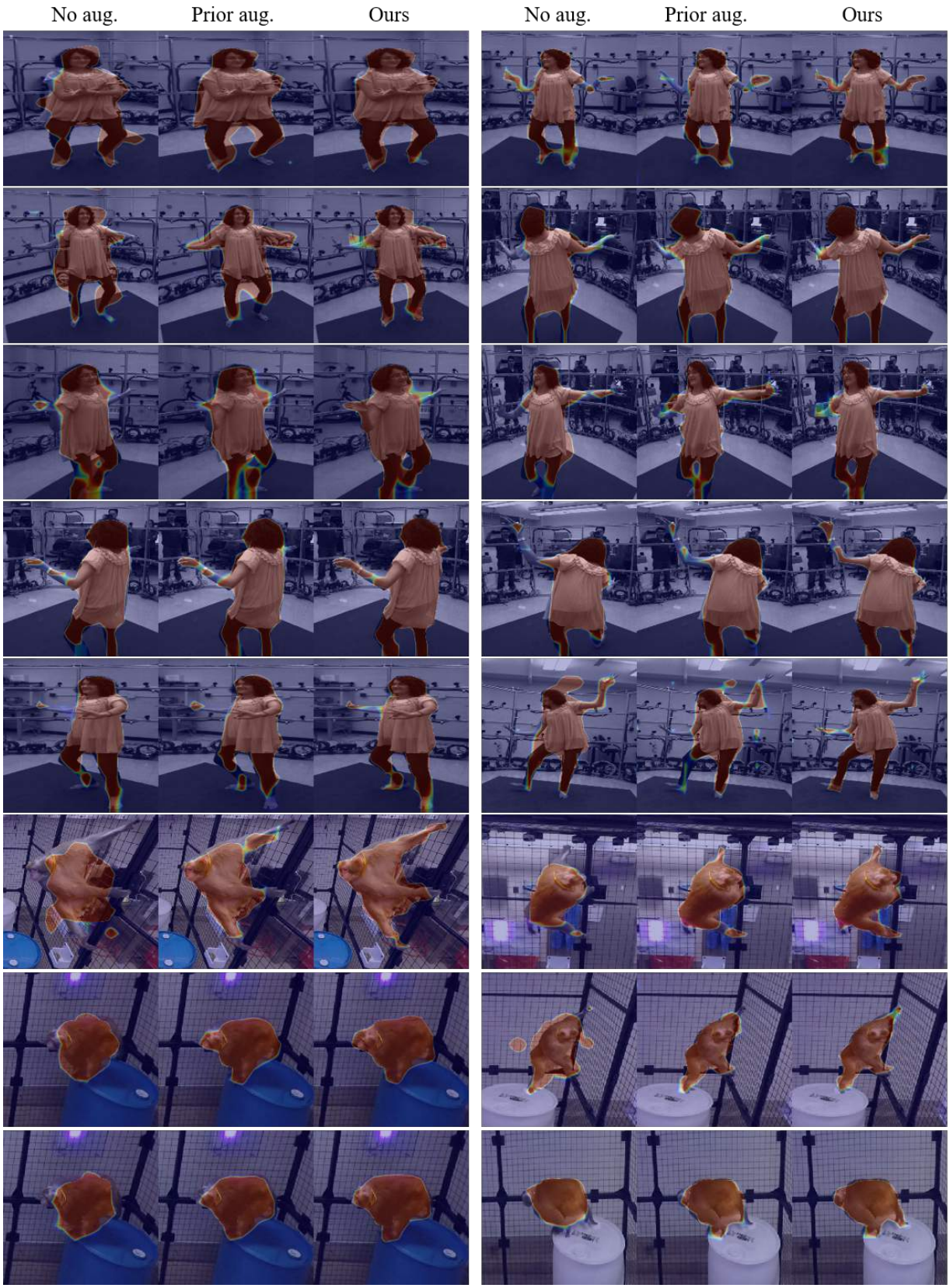

Figure 13: Qualitative result of multiview segmentation.

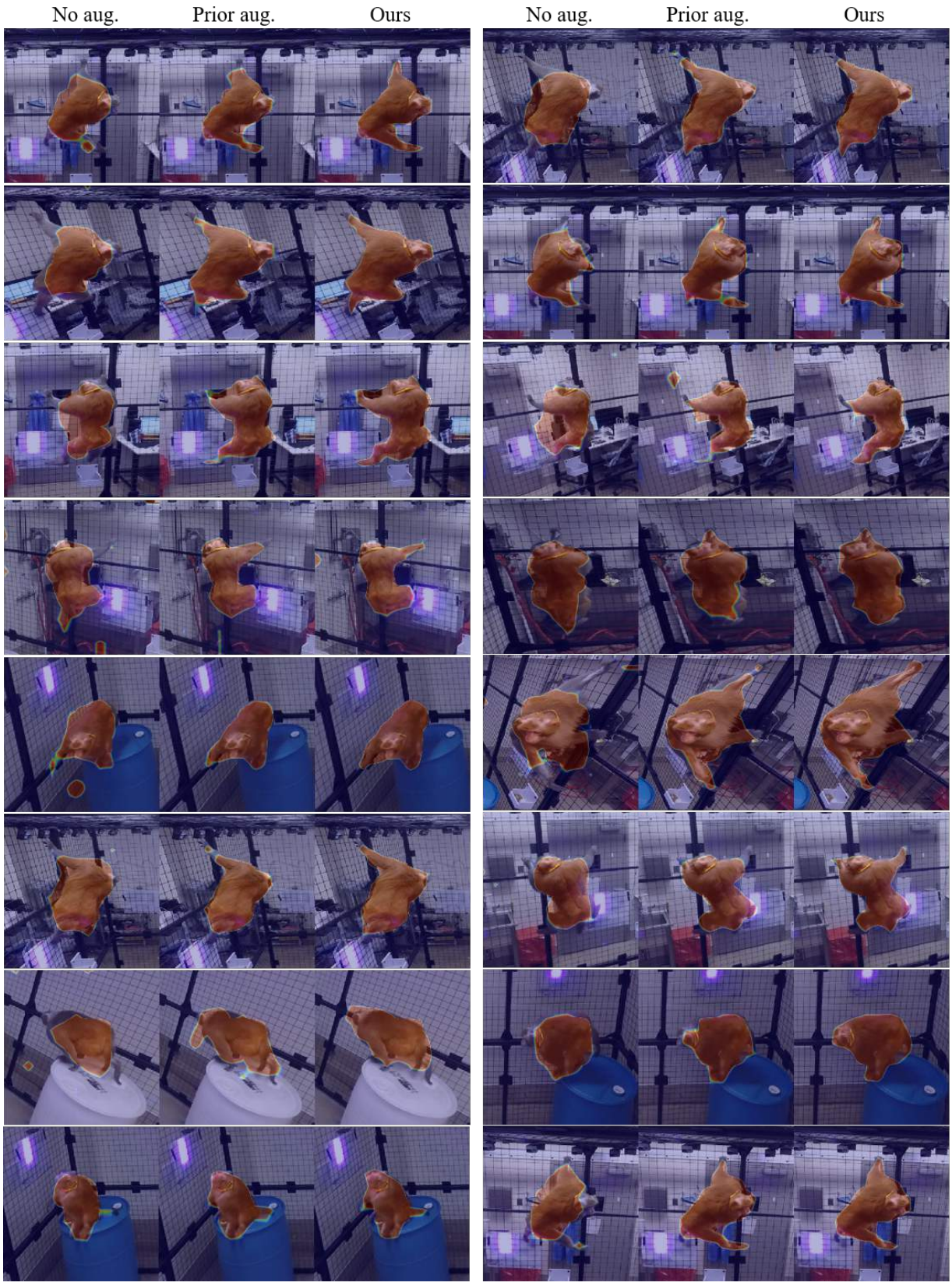

Figure 14: Qualitative result of multiview segmentation.

Figure 15: Qualitative result of multiview segmentation.

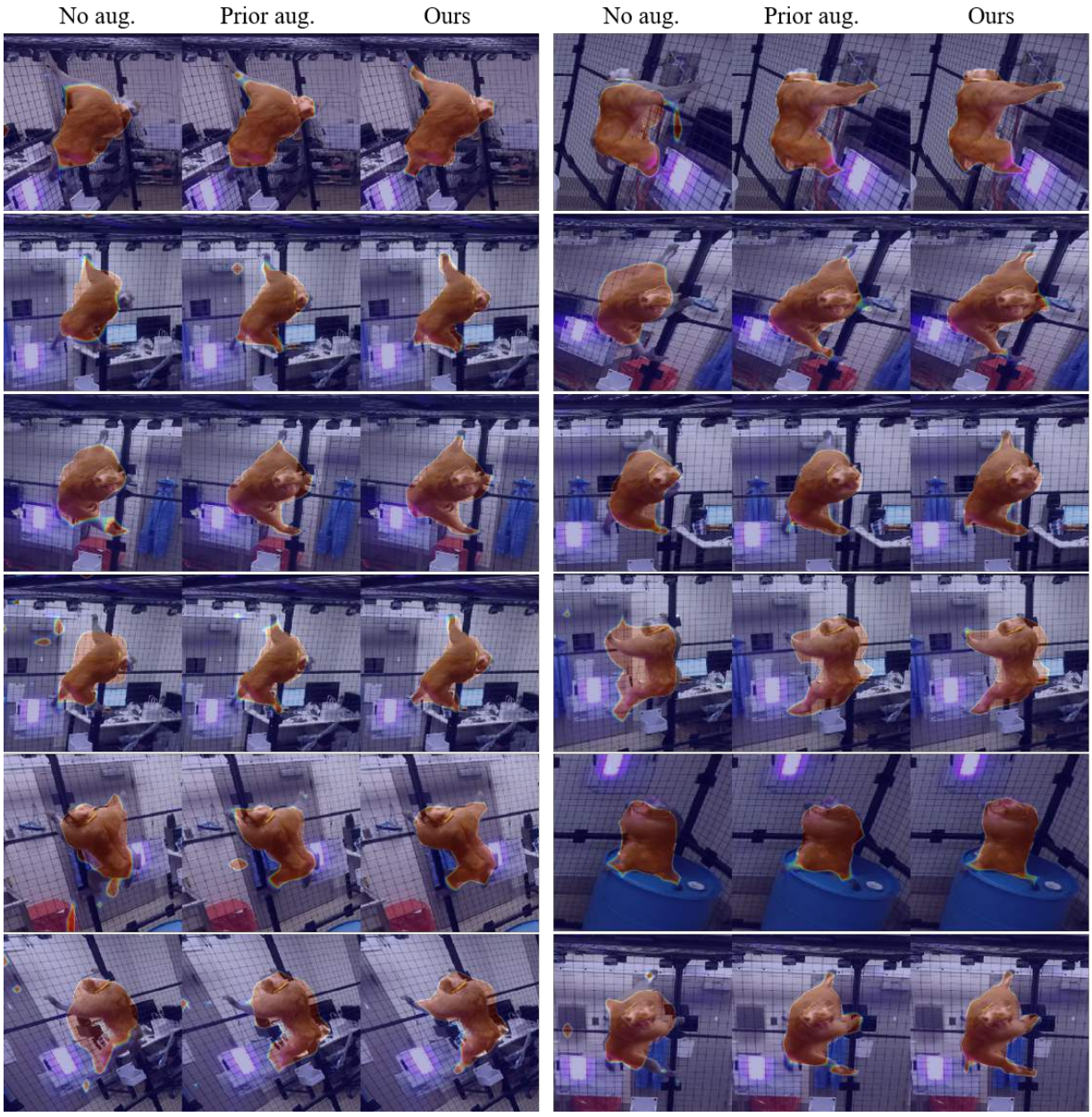

Figure 16: Qualitative result of multiview segmentation.

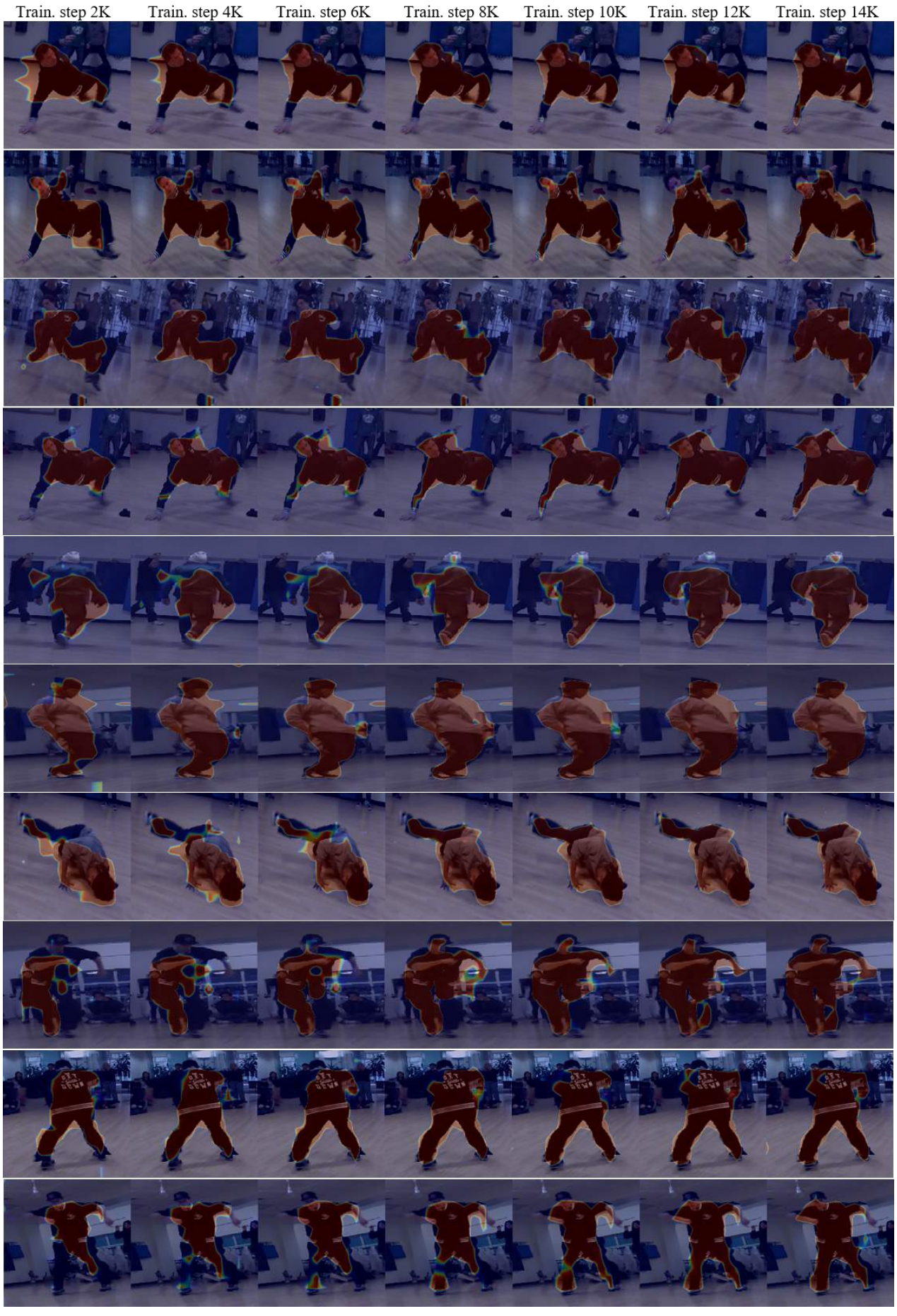

Figure 17: Qualitative result of multiview segmentation.

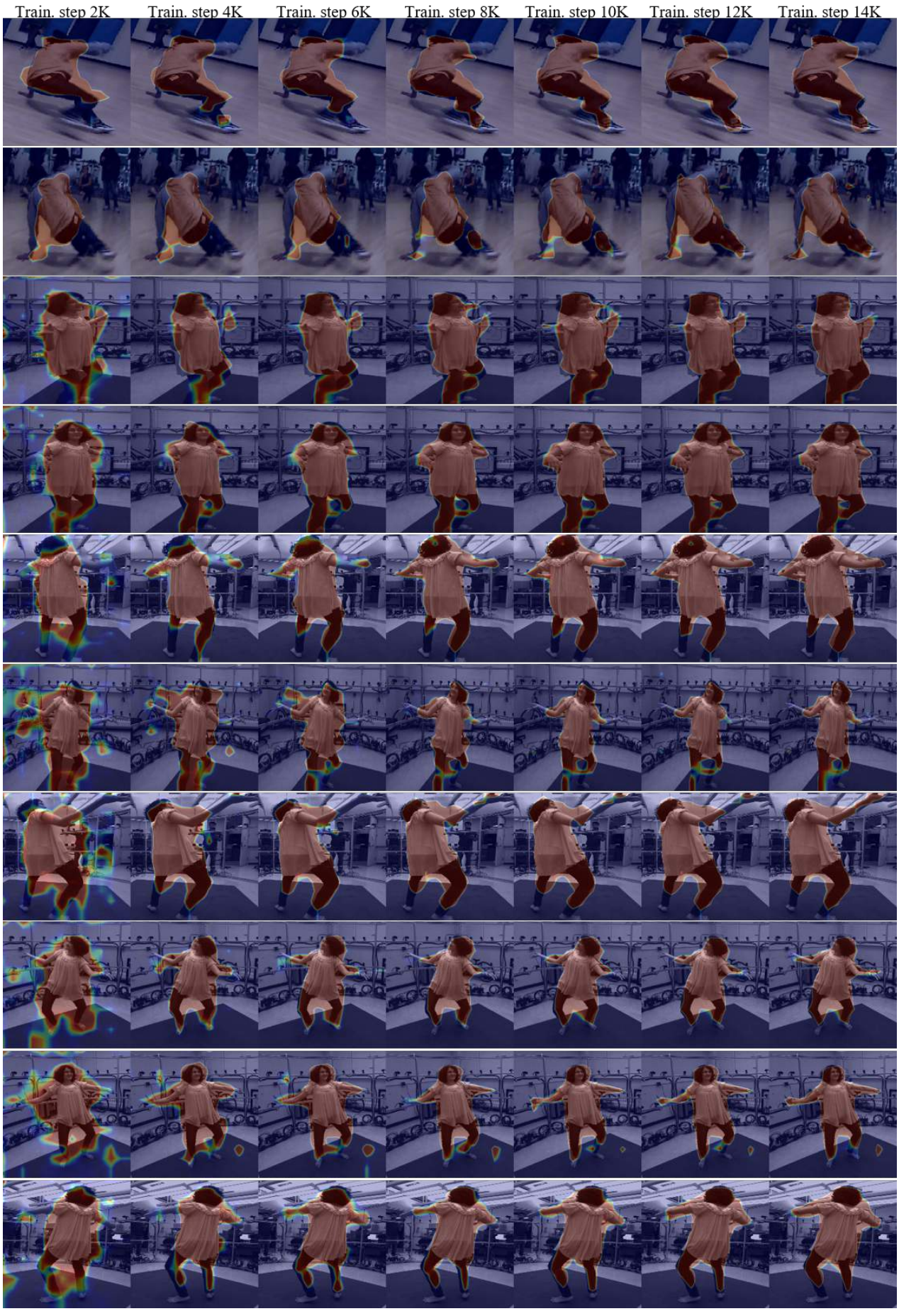

Figure 18: Qualitative result of multiview segmentation.

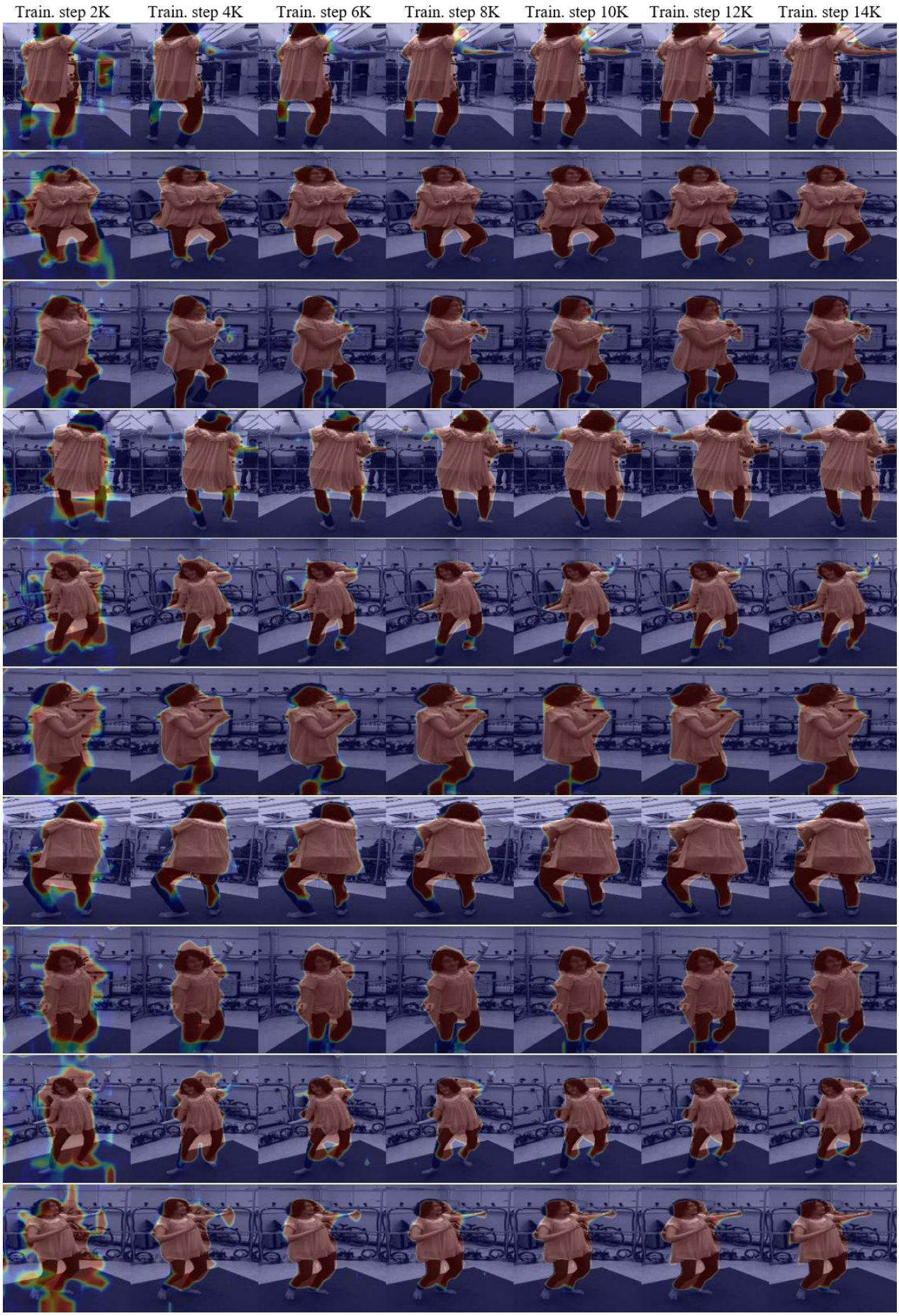

Figure 19: Qualitative result of multiview segmentation.

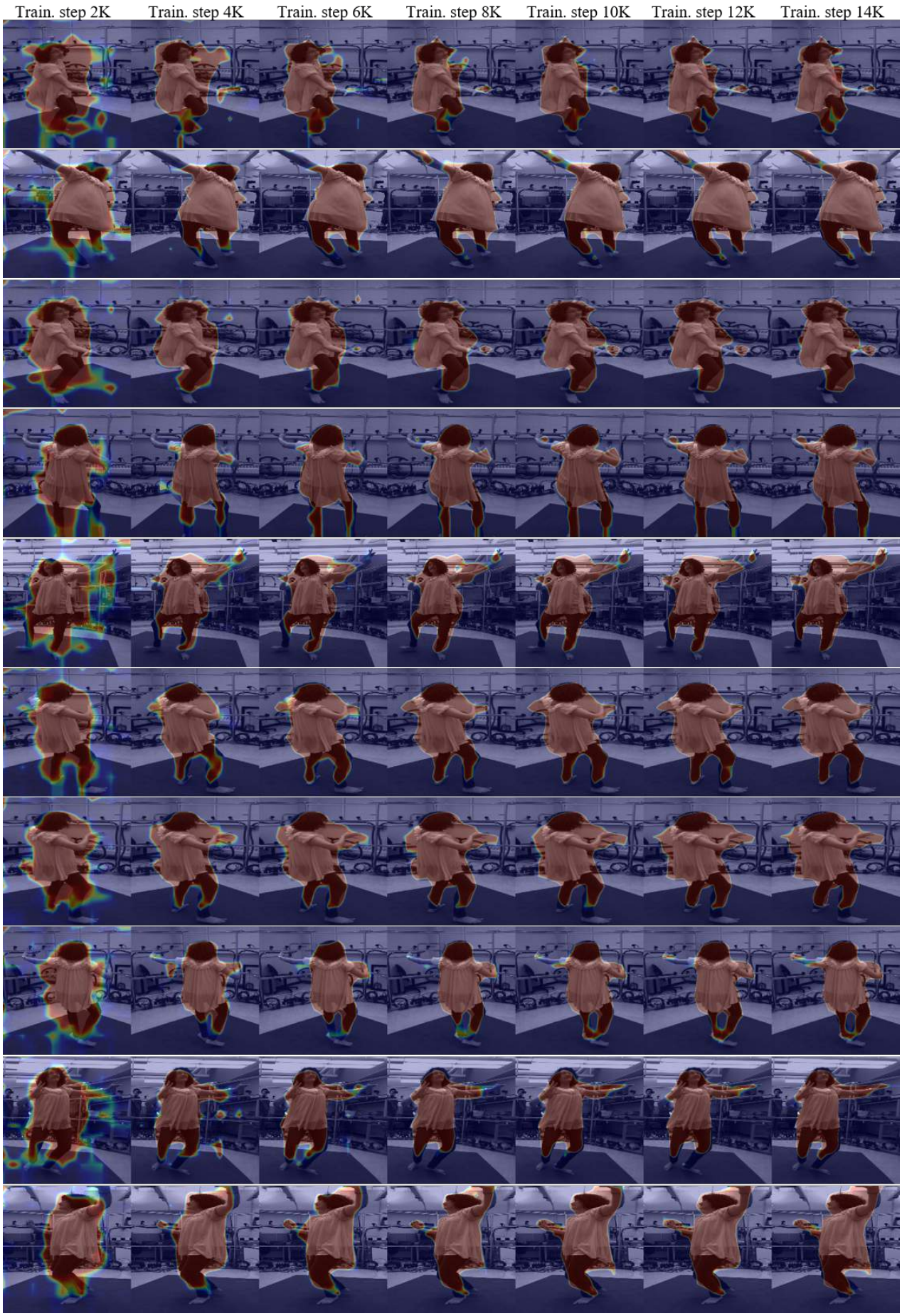

Figure 20: Qualitative result of multiview segmentation.

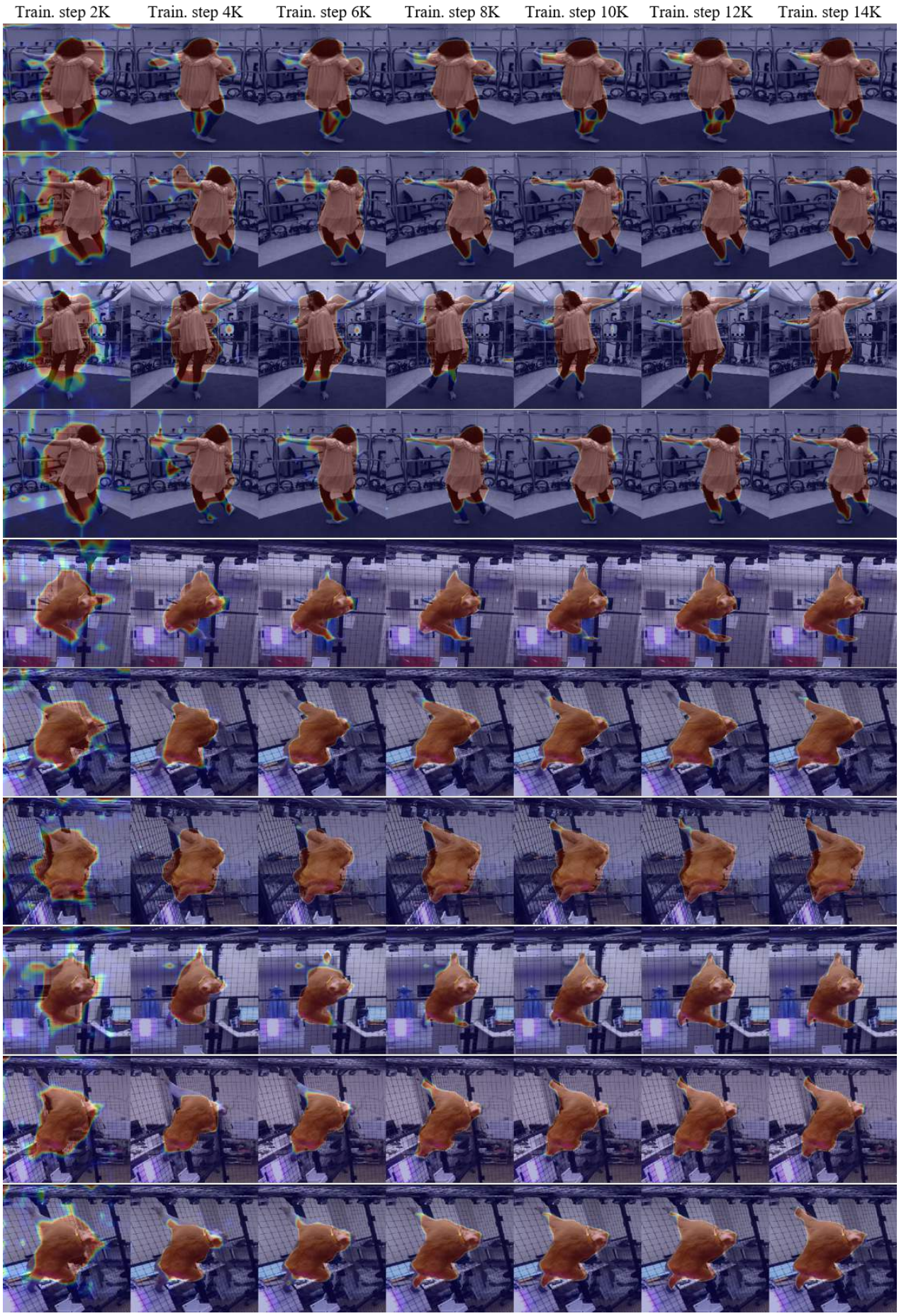

Figure 21: Qualitative result of multiview segmentation.

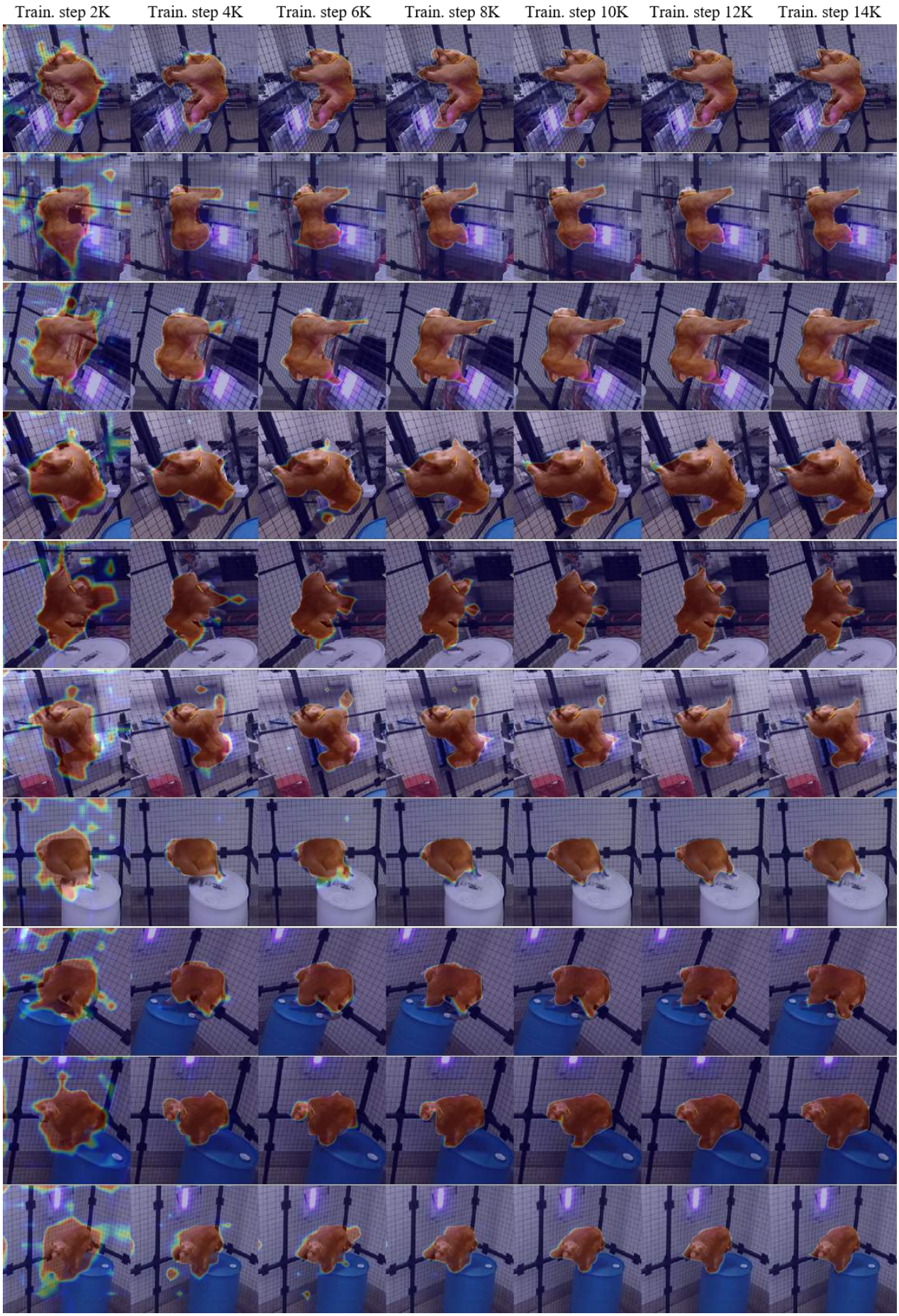

Figure 22: Qualitative result of multiview segmentation.

Figure 23: Qualitative result of multiview segmentation.

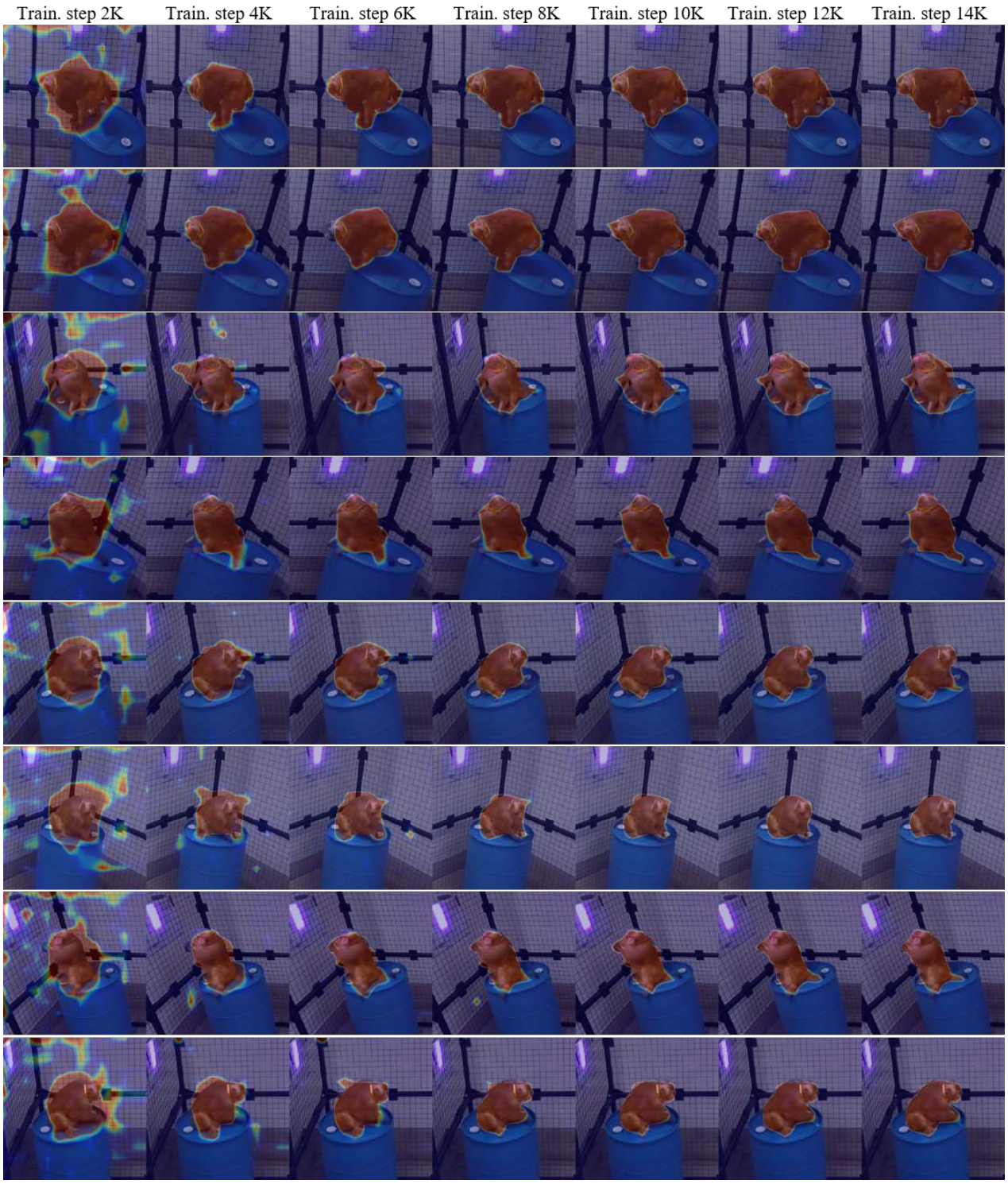

Figure 24: Qualitative result of multiview segmentation.

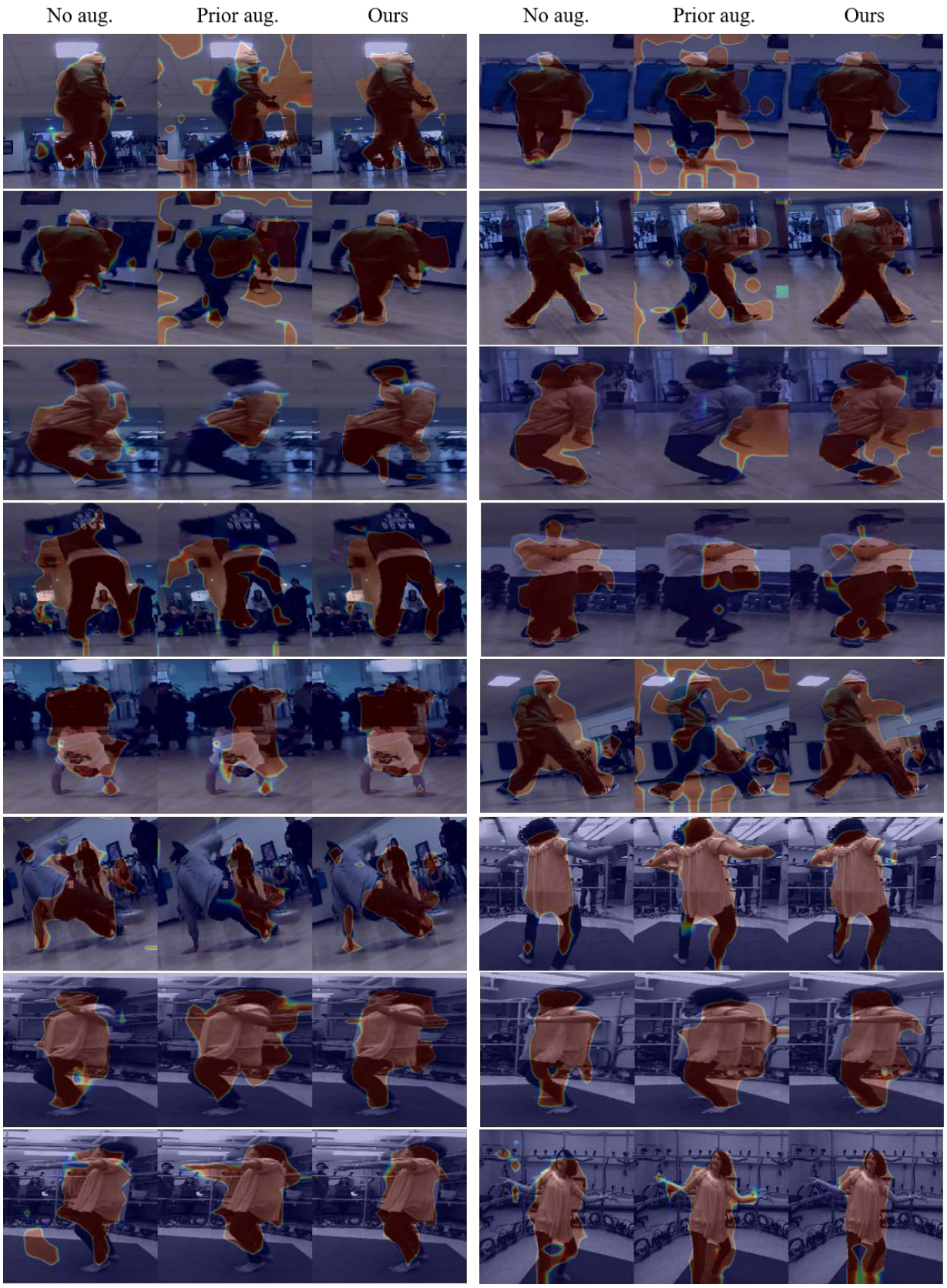

Figure 25: Qualitative result of multiview segmentation.

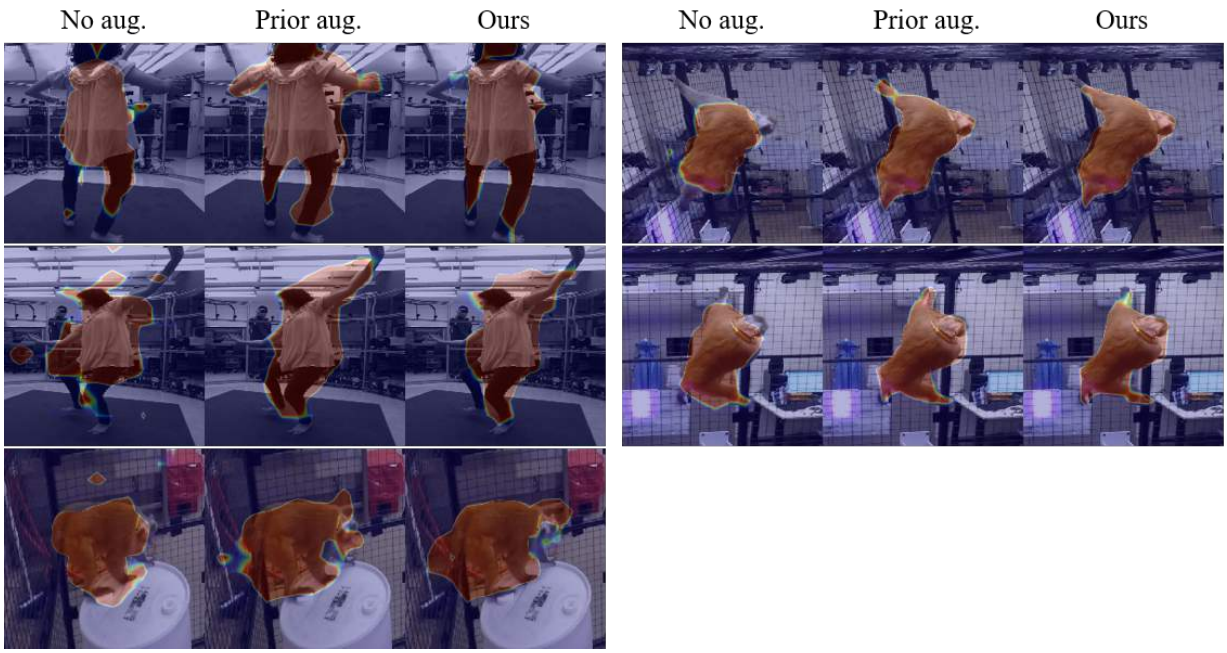

Figure 26: Failure cases.