Multiview Cross-supervision for Semantic Segmentation

Yuan Yao
University of Minnesota
yaoxx340@umn.edu

Hyun Soo Park
University of Minnesota
hspark@umn.edu

Abstract

This paper presents a semi-supervised learning framework for a customized semantic segmentation task using multiview image streams. A key challenge of the customized task lies in the limited accessibility of the labeled data due to the requirement of prohibitive manual annotation effort. We hypothesize that it is possible to leverage multiview image streams that are linked through the underlying 3D geometry, which can provide an additional supervisionary signal to train a segmentation model. We formulate a new cross-supervision method using a shape belief transfer—the segmentation belief in one image is used to predict that of the other image through epipolar geometry analogous to shape-from-silhouette. The shape belief transfer provides the upper and lower bounds of the segmentation for the unlabeled data where its gap approaches asymptotically to zero as the number of the labeled views increases. We integrate this theory to design a novel network that is agnostic to camera calibration, network model, and semantic category and bypasses the intermediate process of suboptimal 3D reconstruction. We validate this network by recognizing a customized semantic category per pixel from realworld visual data including non-human species and a subject of interest in social videos where attaining large-scale annotation data is infeasible.

1. Introduction

In aid of large-scale visual data, convolutional neural networks (CNN) have been transforming the level of understanding of pixels, which allows deep reasoning about their spatial extent and semantic meaning (e.g., human, bicycle, and horse) [12, 13, 24, 43]. Looking ahead, these models are expected to solve various unprecedented visual tasks customized for our personal data (e.g., recognizing pixels of my daughter among her classmates from a collection of her school play photos). However, such task customization is fundamentally limited by the ability to access the training labels for the personal data. Existing semantic segmentation approaches are mostly built upon the per-pixel semantic label manually annotated by thousands of the crowd workers such as MS COCO [41] that constitutes 2.5 millions of segmentation instances. Unfortunately, attaining such large annotations for the customized segmentation task is often infeasible, which introduces a large bias in the trained model because the required number of the training data is known to be equivalent to that of the perceptrons [68].

Figure 1. We design a semi-supervised learning to train a semantic segmentation model using multiview cross-supervision based on shape transfer. This enables customizing a semantic segmentation task, e.g., a b-boy dancer segmentation from social cameras.

In the meantime, as a small form factor of cameras accelerates a seamless integration into our daily lives, now many scenes are recorded by multiple cameras (e.g., Amazon Cloud Cam and Nest Cam), and they will permeate more and deeper. Notably, there is an emerging trend of social videos [3, 5, 19, 48]—a collection of videos that record an activity of interest (e.g., political rally, concert, and wedding) from social members at the same time[1]. These cameras readily produce terascale multiview image streams, which opens up a new opportunity to address the annotation challenge for a customized task. In this paper, we formulate a new multiview theory for semi-supervised semantic segmentation to train a CNN from the limited number of the labeled data ($<$15%) by leveraging the multiview image streams.

A key property of multiview images is that they are linked through the underlying 3D geometry, which can be beneficial for training a segmentation model. However, the representations used for 3D reconstruction and CNNs often mismatch, i.e., vector vs. raster representations, which

[1]There exist multiple online repositories such as Rashomon Project [55] and CrowdSync cellphone app [52] that host the social videos.

makes a tight integration of 3D geometric knowledge into the process of the network training challenging. Existing methods take either a) the approach that alternates between offline 3D reconstruction and training [7, 8, 54, 61]; or b) the approach that predicts the 3D geometry from a single view image with additional depth supervision [27, 60, 65]. The main limitation of these approaches is that their performance is in principle bounded by the reconstruction quality, which is often suboptimal.

Instead, we present a new multiview learning theory for a customized semantic segmentation task that integrates 3D geometry into the process of segmentation model training, which bypasses the intermediate reconstruction. We introduce a *shape belief transfer*—the segmentation belief in one image is used to predict that of the other image through epipolar geometry. We formulate this shape belief transfer as an inverse problem of shape-from-silhouette [17, 23, 35, 36] that reconstructs a 3D object volume (visual hull) from the foreground segmentation of multiview images [33, 45]. The shape belief transfer is a composition of two belief transfers: (a) 3D shape reconstruction by triangulating the segmentation probability in multiview source images; and (b) 2D projection of the reconstructed 3D shape onto a target view to approximate its segmentation probability. We derive that these two transfers can be combined in a differentiable fashion, and therefore, the end-to-end training is possible. This allows relating the segmentation across views where the unlabeled data can be cross-supervised by the labeled data.

A new theory of the shape belief transfer is derived, which provides the upper and lower bounds of the segmentation for the unlabeled data where its gap approaches asymptotically to zero as the number of the labeled views increases. We further show that the shape belief transfer can be implemented by incorporating stereo rectification that transforms the operation of 2D projection into max-pooling operation to gain significant computational efficiency. Based on the theory, we design a triplet network that takes as input multiview image streams with the limited number of the labeled data and outputs a semantic segmentation model that can reliably predict on the unlabeled data as shown in Figure 1. The network is trained by minimizing the geometric inconsistency of multiview segmentation, resulting in multiview cross-supervision.

This framework is flexible: (1) segmentations can be customized as it does not require a pre-trained model, i.e., we train a segmentation model from scratch with manual annotations for each sequence; (2) it can be built on any semantic segmentation design such as DeepLab [10], SegNet [4], and Mask R-CNN [24] that generates a distribution (heatmap) for each object class; (3) it can apply to general multi-camera systems (e.g., different multi-camera rigs, number of cameras, and intrinsic parameters). We validate this network by recognizing a customized semantic category per pixel from realworld visual data including non-human species and a subject of interest in social videos where attaining large-scale annotation data is infeasible. Also it quantitatively outperforms the the existing models without cross-view supervision and the model trained with annotations and shape prior in terms of accuracy and precision.

2. Related Work

This work lies in the intersection of semantic segmentation and multiview self-supervision, which enables learning from a small set of the labeled data possible. We briefly review these two area of study.

Semantic Segmentation Semantic segmentation has been notorious for its computational complexity [21] caused by spatial interactions between pixels. A seminal work by Long et al. [42] has shown that such complex relationship can be effectively learned by a CNN (i.e., fully convolutional network) that encodes high level visual semantics. Albeit impressive, due to the limited network capacity and low resolution, the segmentation results misses object boundary details. Many subsequent studies have integrated a conditional random field or Markov random field [2, 11, 49] that can jointly optimize the object boundary and region. Another approach is to leverage devolutional layers similar to variational autoencoder to reconstruct full resolution segmentation [25, 47]. Such advancement produces a variety of applications such as graphics [1, 38, 50, 57], autonomous driving [31, 32], and first-person vision [18]. When multiple images are used, co-segmentation is possible, i.e., segmenting common objects. Most approaches often leverage individual segmentation to correlate their visual features [28, 39, 46]. Notably, multiview co-segmentation has been studied by using multiview stereo [16, 29]. Unlike these methods, our approach train a semantic segmentation networks using multiview geometry without reconstructing 3D objects, which is not sensitive to the stereo matching error.

Multiview Self-supervision Learning a view invariant representation is a long-standing goal in visual recognition research, which requires to predict underlying 3D structure from a single view image. Geometrically, it is an ill-posed problem while two data driven approaches have made promising progress. (1) Direct 3D-2D supervision: for a few representative objects such as furniture [40], vehicles [62], and human body [44], their 3D models (e.g., CAD, point cloud, and mesh) exist where the 3D-2D relationship can be directly regressed. The 3D models can produce a large image dataset by projecting onto all possible virtual viewpoints where the object's pose and shape can be learned from 3D-2D pairs. This 3D model projection can be generalized to scenes measured by RGBD

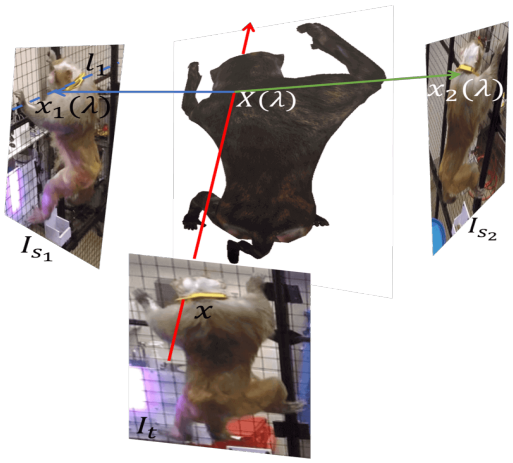

Figure 2. The 3D points and the corresponding 2D points on their epipolar lines can be parameterized with λ.

data [8, 22, 34, 56, 61] and graphically generated photo-realistic scenes [13, 51] where visual semantics associated with 3D shape can be encoded. (2) Indirect supervision via non-rigid graph matching: to some extent, it is possible to infer the common shape and appearance from a set of single view image instances without 3D supervision. For instance, tables have a common shape expressed by four legs and planar top. Such holistic spatial relationship can be unveiled by casting it as a graph matching problem where local shape rigidity and appearance models can describe the relationship between nodes and edges [6, 9, 14, 37, 58, 64]. Further, leveraging a underlying geometric constraint between instances (e.g., cyclic consistency [66, 67], volumetric projection [15, 59, 60], and kinematic chain [53, 58, 63]) can extend the validity of graph matching. These existing approaches require many correspondences between domains that are established by manual annotations. In contrast, our approach will leverage self-supervision via multiview geometry to adapt to a novel scene with minimal manual efforts.

3. Multiview Cross-view Supervision

We present a semi-supervised learning framework for training a semantic segmentation model by leveraging multiview images streams where $\eta = \frac{|\mathcal{D}_L|}{|\mathcal{D}_L|+|\mathcal{D}_U|} \ll 1$ where $\mathcal{D}_L$ and $\mathcal{D}_U$ are the labeled and unlabeled data, respectively. We formulate a novel theory of rasterized multiview geometry that enforces the geometric consistency by minimizing the reprojection error of a 3D visual hull, resulting in a differentiable loss function to train a neural network. Note that we will focus on binary segmentation for a proof of concept while the multiview theory can be applied to multi-way segmentation. Also the the framework is agnostic to the design of segmentation networks where state-of-the-art models [10, 42] can be used with a trivial modification similar to MONET [26].

Consider a network model that takes an input image $\mathcal{I}$ and outputs the class probabilities for each pixel, i.e., $\phi(\mathcal{I}; \mathbf{w}) \in [0,1]^{W\times H\times C}$ where W and H are the width and height of the output distribution, respectively, and C is the number of object classes. We consider binary segmentation, i.e., $|C| = 2$.

The network is parametrized by the weight $\mathbf{w}$ learned by minimizing the following loss:

$$\underset{\mathbf{w}}{\text{minimize}} \;\; \mathcal{L}_L + \lambda_s \mathcal{L}_S + \lambda_p \mathcal{L}_P, \tag{1}$$

where $\mathcal{L}_L$, $\mathcal{L}_S$, and $\mathcal{L}_P$ are the losses for labeled supervision, multiview cross-view supervision, and bootstrapping prior, and λ_s and λ_p are the weights that control their importance.

For the labeled data, we use the sum of pixelwise cross entropy to measure the segmentation loss:

$$\mathcal{L}_L = -\sum_{j\in\mathcal{D}_L}\sum_{\mathbf{x}\in X} y_j(\mathbf{x}) \log \phi(\mathcal{I}_j)|_{\mathbf{x}}\,, \tag{2}$$

where $y_j(\mathbf{x}) \in \{0,1\}$ is the ground truth semantic label of the j^{th} labeled data at pixel location $\mathbf{x}$, and X is the domain of $\mathbf{x}$.

3.1. Shape Transfer

Inspired by the image based shape from silhouette [45], we study the segmentation transfer through a 3D shape. Consider a point $\mathbf{x} \in \mathbb{R}^2$ in the target image $\mathcal{I}_t$. Without loss of generality, the camera projection matrix of the target image is set to $\mathbf{P} = \mathbf{K}\begin{bmatrix}\mathbf{I}_3 & \mathbf{0}\end{bmatrix}$ where $\mathbf{K}$ is the intrinsic parameter. The point in an image is equivalent to a 3D ray $\mathbf{L}_{\mathbf{x}} \propto \mathbf{K}^{-1}\widetilde{\mathbf{x}}$ emitted from the target camera. A 3D point along the ray can be represented as $\mathbf{X}(\lambda) = \lambda\mathbf{L}_{\mathbf{x}}$ where any scalar depth $\lambda > 0$.

A series of projections of $\mathbf{X}(\lambda)$ onto a source image, $\mathcal{I}_{s_1}$ form the epipolar line $\mathbf{l}_1 = \mathbf{F}_1\widetilde{\mathbf{x}}$ where $\mathbf{F}_1$ is the fundamental matrix between the target and source image. This indicates the point on the epipolar line can be parametrized by λ as shown in Figure 2, i.e., $\mathbf{x}_1(\lambda) \in \mathbf{l}_1$[2]. Likewise a point $\mathbf{x}_i$ in the i^{th} source image $\mathcal{I}_i$ can be described accordingly.

The image based shape-from-silhouette computes a binary map $z_t : \mathbb{R}^2 \to \{0,1\}$ that determines the pixel being foreground if one, and zero otherwise. This binary map can be approximated by the logical operations between the binary maps from the n source images $(z_{s_1}, \cdots, z_{s_n})$:

$$\hat{z}_t(\mathbf{x}) = \begin{cases} 1 & \text{if } \exists\, \lambda > 0 \text{ s.t. } \bigwedge_i z_{s_i}(\mathbf{x}_i(\lambda)) = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

[2]We use an abuse of notation: $\mathbf{x} \in \mathbf{l}$ is equivalent to $\widetilde{\mathbf{x}}^{\mathsf{T}}\mathbf{l} = 0$, i.e., the point $\mathbf{x}$ belongs to the line $\mathbf{l}$

The geometric interpretation of Equation (3) is that the foreground map for $\mathbf{x}$ is computed by sweeping across all 3D points along the ray $\mathbf{L_x}$ to see if the ray intersects with the 3D volumetric shape defined by the foreground maps from n views. A key property of this foreground approximation $\hat{z}_t(\mathbf{x})$ from n views that it is always inclusive of the true $z_t(\mathbf{x})$, i.e., $\{\mathbf{x}|z_t(\mathbf{x})=1\} \subseteq \{\mathbf{x}|\hat{z}_t(\mathbf{x})=1\}$.

The implication of the approximation of Equation (3) is significant for the semi-supervised learning that includes the unlabeled data because it is possible to transfer the recognition belief between views through the underlying 3D shape where the label for the unlabeled data can be approximated. Inspired by this insight, we formulate a rasterized version of Equation (3) to train a semantic segmentation network.

Let $P_i : \mathbb{R}^2 \rightarrow [0,1]$ be the foreground probability distribution of the i^{th} source image, i.e., $P_i(\mathbf{x}) = \phi(\mathcal{I}_i;\mathbf{w})|_{\mathbf{x}}$. Using the probability distribution, it is possible to compute the probability over the ray $\mathbf{L_x}$ by projecting the ray onto the i^{th} image:

$$\xi_i(\lambda;\mathbf{L_x}) = P_i(\mathbf{x}_i(\lambda)) \text{ where } \mathbf{x}_i(\lambda) \in \mathbf{F}_i\widetilde{\mathbf{x}}, \quad (4)$$

where $\xi_i(\lambda;\mathbf{L_x})$ is the probability over the ray parametrized by the depth λ.

From Equation (4), the probability of a target image $P_t : \mathbb{R}^2 \rightarrow [0,1]$ can be approximated by a 3D line max-pooling over joint probability over n views:

$$\hat{P}_t(\mathbf{x}) = \sup_{\lambda>0} \prod_{i=1}^{n} \xi_i(\lambda;\mathbf{L_x}), \quad (5)$$

where $\hat{P}_t(\mathbf{x})$ is the foreground probability transferred from n views. Equation (5) is equivalent to Equation (3) where it takes the probability of a 3D point most likely being in the volumetric shape.

Note that similar to $\hat{z}_t$, the $\hat{P}_t$ provides the upper bound of the P_t, i.e., $\{\mathbf{x}|P_t(\mathbf{x})>\epsilon\} \subseteq \{\mathbf{x}|\hat{P}_t(\mathbf{x})>\epsilon\}$. Therefore, direct probability matching using KL divergence [30] does not apply. Instead, we formulate a new loss D_S using one-way relative cross-entropy as follow:

$$\mathcal{L}_S = D_S(P_t||\hat{P}_t) = \sum_{\mathbf{x}\in X} (1-\hat{P}_t(\mathbf{x}))P_t(\mathbf{x}), \quad (6)$$

where X is the range of the target image coordinate. $D_S(P_t||\hat{P}_t)$ strongly penalizes the set of pixels, $\{\mathbf{x}|\hat{P}_t(\mathbf{x}) < P_t(\mathbf{x})\}$. Figure 3 shows the visualization of the cross-supervision loss.

The main benefits of Equation (6) are threefold. (1) Multiview segmentation involves two processes: 3D reconstruction of the shape with source views and 2D projection onto the target view. The requirement of 3D reconstruction introduces an additional estimation such as multiview [16,29] or single view depth prediction [27,60,65] where the accuracy of the segmentation is bounded by the reconstruction quality. Equation (6) integrates the 3D reconstruction and projection through the joint probability over the epipolar lines and supremum operation, which bypass the 3D reconstruction. (2) By minimizing Equation (6), it can provide a *pseudo*-label for the unlabeled data transferred from the labeled data. As the number of the labeled data increases, the transferred segmentation label approaches to the true label of the unlabeled data [33,45], which allows multiview self-superivsion, i.e., the semantic segmentation of labeled data can supervise the that of the unlabeled data. (3) Not only for the unlabeled data, but also it can correct the geometrically inconsistent segmentation label for the labeled data. This is a significant departure from the existing semantic segmentation that cannot recover erroneous segmentation label, which often arises from per-pixel manual annotations.

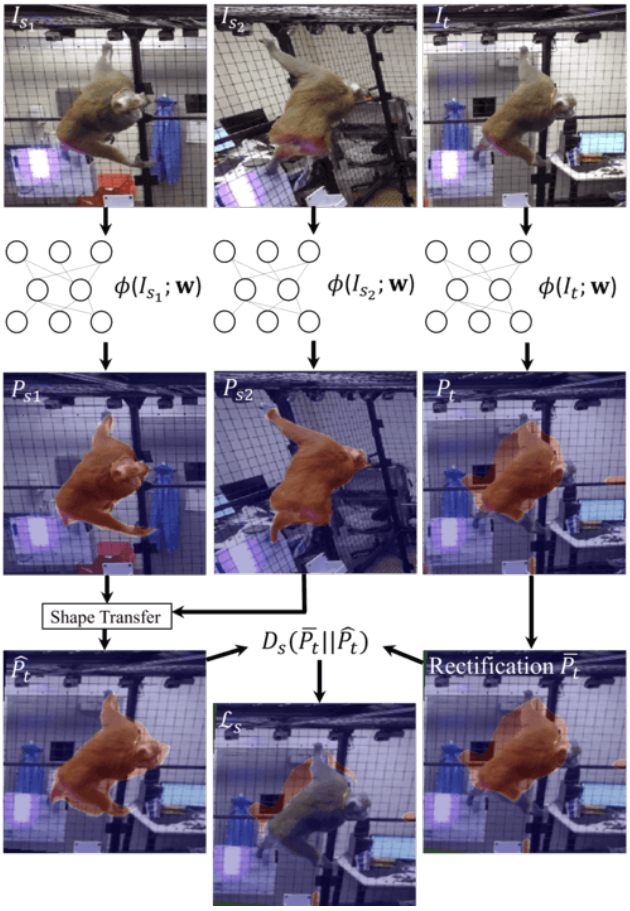

Figure 3. We design a triplet network that allows multiview cross-supervision using shape transfer. Stereo rectification is used to simplify the max-pooling operation along the epipolar line, which reduces computational complexity and sampling aliasing.

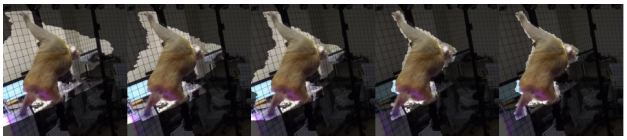

Figure 4. The upper bound of the probability for the unlabeled data becomes tighter as the number of the labeled data increases.

3.2. Degenerate Case Analysis

Equation (6) has a degenerate case: a trivial solution $P_t = 0$ is the global minimizer. Therefore, when the unlabeled data sample is used for the target view, the cross-view supervision via shape transfer based on the labeled data is not possible, i.e., $\hat{P}_U = P_U^+ > P_U$.

Theorem 1. *There exists the lower bound of the probability of the unlabeled data sample, P_U^-.*

Proof. Consider an inverse shape transfer for the unlabeled data in Equation (5), $\phi_U(\lambda; \mathbf{L_x})$, to explain the first labeled data sample P_{L_1}:

$$P_{L_1}(\mathbf{x}) = \sup_{\lambda>0} \xi_U(\lambda; \mathbf{L_x}) \prod_{i=2}^{n} \xi_{L_i}(\lambda; \mathbf{L_x}), \tag{7}$$

where P_{L_i} is the probability of the i^{th} labeled data. Since the supremum in Equation (7) is a non-decreasing function with respect to $\xi_U(\lambda; \mathbf{L_x})$, there exists $\xi_U^-(\lambda; \mathbf{L_x}) < \xi_U(\lambda; \mathbf{L_x})$ that cannot explain $P_{L_1}(\mathbf{x})$:

$$P_{L_1}(\mathbf{x}) > \sup_{\lambda>0} \xi_U^-(\lambda; \mathbf{L_x}) \prod_{i=2}^{n} \xi_{L_i}(\lambda; \mathbf{L_x}). \tag{8}$$

Therefore, there exists the lower bound of P_U. □

From Theorem 1, Equation (6) can provide both upper and lower bounds of the unlabeled data if used as the target and source views, i.e., $P_U^- < P_U \leq P_U^+$, and P_U^- asymptotically approaches to P_U^+ as the number of labeled views increases [33, 45], i.e., $\lim_{|\mathcal{D}_L|\to\infty}(P_U^+ - P_U^-) = 0$. Figure 4 shows the upper bound becomes tighter as the number of labeled data increases

We leverage this asymptotic convergence of the shape transfer to self-supervise the unlabeled data, i.e., the unlabeled data are fed into both the target and source views, which allows the gradient induced by the error in the loss function of Equation (6) can be backpropagated through the neural network to reduce the gap between P_U^+ and P_U^-.

3.3. Cross-view Supervision via Shape Transfer

In practice, embedding Equation (6) into an end-to-end neural network is not trivial because (a) a new max-pooling operation over oblique epipolar lines needs to be defined; (b) sampling interval for max-pooling along the line is arbitrary, i.e., uniform sampling does not encode geometric meaning such as depth; and (c) sampling interval across different epipolar line parameters is also arbitrary, which may introduce sampling aliasing. This leads to irregular segmentation probability distribution transfer based on the fundamental matrix.

We introduce a new operation inspired by stereo rectification, which warps the segmentation probability distribution such that the epipole is transformed to a point at infinity, i.e., the epipolar lines become parallel (horizontal). This rectification allows converting the oblique line max-pooling into regular row-wise max-pooling.

Equation (4) can be re-written by rectifying the probability distribution of the source view with respect to the target view:

$$\overline{\xi}_1(u; \mathbf{L_x}) = \overline{P}_1\left(\begin{bmatrix} u \\ v_1 \end{bmatrix}\right) \text{ where } \mathbf{K}\mathbf{R}_1\mathbf{K}^{-1}\widetilde{\mathbf{x}} \propto \begin{bmatrix} x \\ v_1 \\ 1 \end{bmatrix},$$

where $\mathbf{K}\mathbf{R}_1\mathbf{K}^{-1}\widetilde{\mathbf{x}}$ is the rectified coordinate of the target view, $\mathbf{R}_1 \in SO(3)$ is the relative rotation for the rectification. See Appendix for more details. Note that ξ is no longer a function of the depth scale λ but the x coordinate (disparity), which eliminates irregular sampling across pixels with the y coordinate v_1.

The key advantage of this rectification is that the x coordinate of the i^{th} view can still be parametrized by the same u, i.e., the coordinate is linearly transformed to from the first view to the rest views:

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

where a_i and b_i are the linear re-scaling factor and bias between the first and i^{th} views accounting for camera intrinsics and cropping parameters. $\overline{\phi}_i$ is computed by the rectified probability of the i^{th} view $\overline{P}_i$ with respect to the target view. See Appendix for more details. This simplifies the supremum operation over the 3D ray in Equation (5) to the max operation over the x coordinates:

$$\hat{P}_t(\mathbf{x}) = \max_{u\in[0,W]} \overline{\xi}_1(u; \mathbf{L_x}) \prod_{i=2}^{n} \overline{\xi}_i(a_i u + b_i; \mathbf{L_x}). \tag{9}$$

Our semi-supervised learning framework has Siamese network configuration which consists of four same segmentation models with shared weights. The first network is fully supervised which only learns the labeled data from their annotations. The triplet networks take three different images in the same frame, and they can be either labeled or unlabeled images. Figure 3 illustrates the triplet network that minimizes the cross-view supervision loss by applying stereo rectification and shape transfer where the first two images serve as source views and third image is target view. The foreground probability distributions of the first source

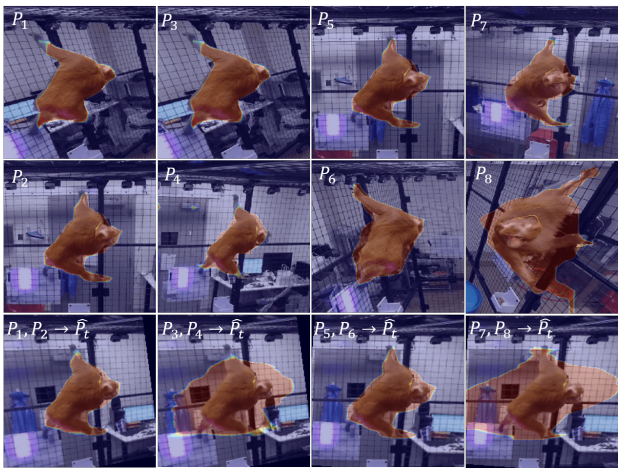

Figure 5. Different image pairs (top two rows) can be used to supervise one target view (bottom). We use such multiple triplets to supervise each other's view.

image and target view are rectified to reduce the sampling aliasing and computational complexity for computing the fundamental matrices. The foreground probability of each point in the target view transferred from two source views is calculated using Equation (9). The cross-view supervision loss computed by Equation (6) is propagated back to four networks. In the actual implementation, each image is used as source view to supervise other two images and is supervised by one image pair simultaneously. The degenerate case for unlabeled data discussed in Section 3.2 can be avoided once it is input to this triplet network. Figure 5 shows that one target view can be supervised by multiple different image pairs during the actual training.

3.4. Bootstrapping Prior

Equation (3) is often highly effective to generate a prior for 3D shape given the binary label. Inspired by multiview bootstrapping [26, 54], we approximate the 3D shape using the pre-trained neural network ϕ. Note that unlike keypoint detection, RANSAC [20] outlier rejection approaches cannot be applied because pixel correspondences are not available for semantic segmentation. We binarize the probability of the foreground segment to compute the i^{th} source binary map $z_{s_i}(\mathbf{x}) = 1$ if $P_i(\mathbf{x}) > 0.5$, and zero otherwise. Using all source binary maps, a pseudo-binary map for the j^{th} unlabeled data $\hat{z}_j$ can be computed and used for the bootstrapping prior, i.e.,

$$\mathcal{L}_P = \sum_{j \in \mathcal{D}_U} \sum_{\mathbf{x} \in X} (1 - \hat{z}_j(\mathbf{x})) P_j(\mathbf{x}) \qquad (10)$$

Similar to Equation (6), $\hat{z}_j$ provides the superset of the ground truth, which requires the one-way relative cross entropy as a prior loss.

4. Result

We validate our semi-supervised semantic segmentation framework using real-world data on human and non-human species including a subject of interest in social videos with three different multi-camera systems. Monkey, dancer and social event subjects are captured by 69, 35, and 18 cameras, respectively. One monkey was crawling against the cage in the video, and the array of cameras were placed in the cage ceiling. An Indian dancer was performing solo dance captured by 69 cameras in three layers with different heights. 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.

To evaluate the flexibility, we build a model per subject without a pre-trained model. The DeepLab v3 [10] network is used to build the fully supervised and semi-supervised triplet network. Our segmentation network takes an input image (200×200), and outputs two distribution heatmaps for foreground and background (200×200). In the training, we use the batch size 5 for fully supervised network and 3 for triplet network, learning rate 10^{-5}, batch norm epsilon 10^{-5}, and batch norm decay 0.9997. We use an ADAM optimizer of TensorFlow with nVidia GTX 1080.

We randomly sample 16, 16, and 14 cameras from monkey, dance and social event datasets and manually annotate the 20 frames in half sampled cameras. We conduct multiple experiments using different number of labeled views for bootstrapping prior and cross-supervision from two to half number of the sampled cameras. We compare our approach with two different baseline algorithms. For all algorithms, we evaluate the performance on the unlabeled data. (1) **No augmentation**: we use the manually annotated images to train the network. (2) **Prior augmentation**: the prior of unlabeled data is generated the way discussed in Section 3.4. We train the network using both annotations and prior.

4.1. Quantitative Result

We evaluate our approach based on two metrics: mean IoU and mean pixel accuracy. Figure 7 shows mean IoU and mean pixel accuracy performance on monkey, dance, and social event subjects using different number of labeled view, and no pre-trained model is used. Table 1 reports the numerical results from Figure 7. All the figures display the same trend that the accuracy increases as the number of labeled views increases.

Our cross-supervision (red) model exhibits accurate segmentation for all subjects, which outperforms 2 baselines. Cross-supervision model and the model trained with prior augmentation both perform better than the model trained with annotation only on the monkey dataset while the model trained with prior augmentation and model trained with annotation have similar performance on the dance dataset. We observe that the most labeled views selected for the mon-

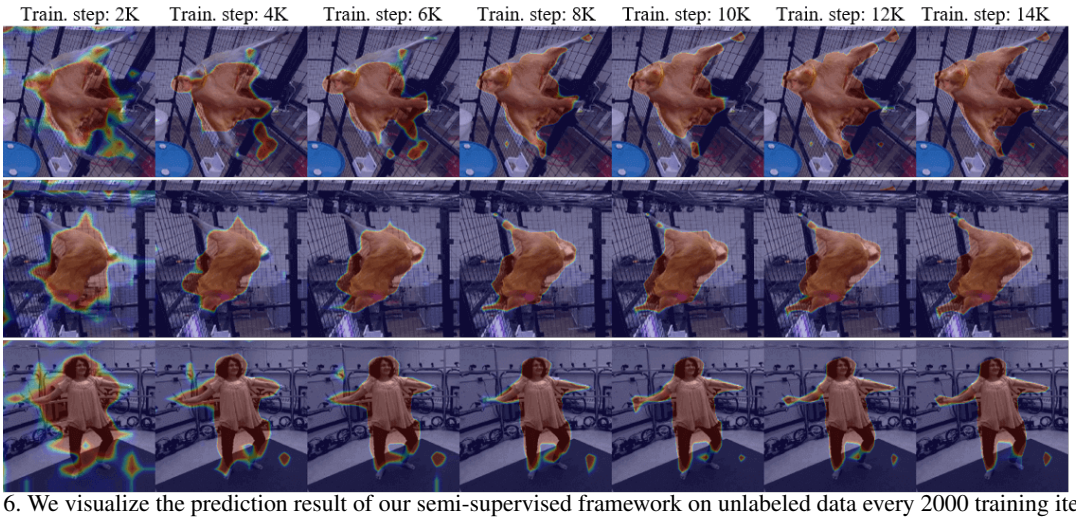

Figure 6. We visualize the prediction result of our semi-supervised framework on unlabeled data every 2000 training iterations.

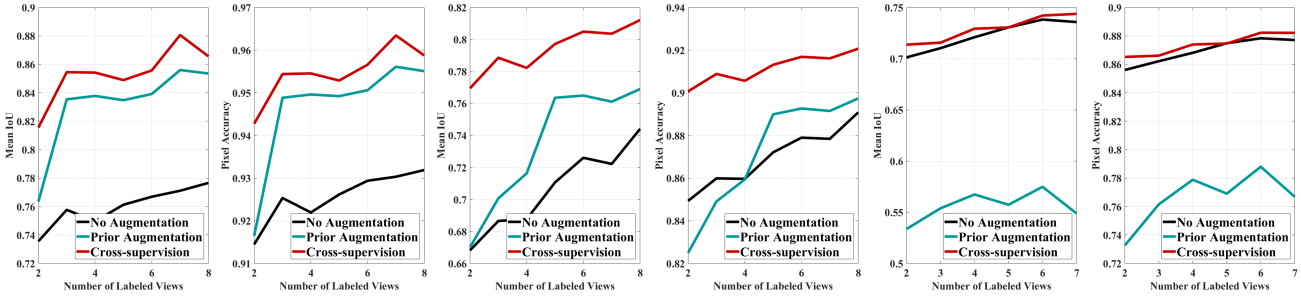

Figure 7. The mean IoU and pixel accuracy of semi-supervised model tested on the unlabeled monkey, dance, and social event dataset.

Model	Monkey (IoU)			Dance (IoU)			Social (IoU)			Monkey (Pixel Acc.)			Dance (Pixel Acc.)			Social (Pixel Acc.)		
	2	5	8	2	5	8	2	5	7	2	5	8	2	5	8	2	5	7
No aug	0.735	0.761	0.776	0.668	0.710	0.744	0.701	0.730	0.735	0.914	0.926	0.931	0.849	0.872	0.890	0.856	0.874	0.877
Prior	0.763	0.834	0.853	0.670	0.763	0.769	0.533	0.557	0.548	0.916	0.949	0.955	0.824	0.889	0.897	0.732	0.769	0.766
Cross	0.815	0.848	0.865	0.769	0.797	0.812	0.713	0.730	0.743	0.942	0.952	0.958	0.900	0.913	0.920	0.865	0.874	0.882

Table 1. Mean IoU and pixel accuracy result on different datasets with different number of labeled views

key dataset can generate a tight upper bound for the monkey in the target views. The labeled cameras sampled for dance video have very close distances and similar angles; therefore, multiple upper bounds constructed for the dancer are very loose. However, as we observe in Figure 7, if the weight on boostrapping prior is set to be lower, the cross-supervision is able to correct the prior.

4.2. Qualitative Result

The qualitative comparison can be found in Figure 8. This figure shows the prediction results on the unlabeled data using three models. Our cross-supervision method is able to correct the segmentation errors in the two baselines by leveraging multiview images jointly. This becomes more evident on the boundaries or protruding body parts, e.g., monkey's paws and tails, human's legs and hands. Figure 6 shows how the semi-supervised network progresses on the unlabeled data during the training. We can see that as the training iteration increases, the prediction becomes tighter around the subject while the protruding body parts can be predicted more accurately.

5. Discussion

We present a new semi-supervised framework to train a semantic segmentation network by leveraging multi-view image streams. The key innovation is a method of shape belief transfer—using segmentation belief in one image to predict that of the other image through epipolar geometry analogous to shape-from-silhouette. The shape belief transfer provides the upper and lower bounds of the segmentation for the unlabeled data. We introduce a triplet network which is used to embed computing of transferred shape. We also use multi-view image streams to bootstrap the unlabeled data for training data augmentation.

Figure 8. We qualitatively compare our semi-supervised framework with 2 baseline algorithms on dance, monkeys, and social event.

References

[1] Y. Aksoy, T.-H. Oh, S. Paris, M. Pollefeys, and W. Matusik. Semantic soft segmentation. *SIGGRAPH*, 2018.

[2] G. B. an Jianbo Shi and L. Torresani. Semantic segmentation with boundary neural fields. In *CVPR*, 2016.

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

[4] V. Badrinarayanan, A. Handa, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv preprint arXiv:1505.07293*, 2015.

[5] T. D. Basha, Y. Moses, and S. Avidan. Photo sequencing. In *ECCV*, 2012.

[6] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *CVPR*, 2005.

[7] 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.

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

[9] T. S. Caetano, J. J. McAuley, L. Cheng, Q. V. Le, and A. J. Smola. Learning graph matching. *TPAMI*, 2009.

[10] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), April 2018.

[11] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic segmentation with object clique potential. In *ICLR*, 2015.

[12] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[13] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, 2017.

[14] M. Cho, K. Alahari, and J. Ponce. Learning graphs to match. In *ICCV*, 2013.

[15] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3D-r2n2: A unified approach for single and multi-view 3D object reconstruction. In *ECCV*, 2016.

[16] A. Djelouah, J.-S. Franco, and E. Boyer. Multi-view object segmentation in space and time. In *ICCV*, 2013.

[17] A. Djelouah, J.-S. Franco, E. Boyer, F. Le Clerc, and P. Pérez. Sparse Multi-View Consistency for Object Segmentation. *TPAMI*, 2015.

[18] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *ICCV*, 2011.

[19] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interaction: A first-person perspective. In *CVPR*, 2012.

[20] 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.

[21] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *ICCV*, 2009.

[22] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[23] L. Guan, S. Sinha, J.-S. Franco, and M. Pollefeys. Visual hull construction in the presence of partial occlusion. In *3D Data Processing, Visualization, and Transmission*, 2006.

[24] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask R-CNN. In *ICCV*, 2017.

[25] S. Hong, H. Noh, and B. Han. Decoupled deep neural network for semi-supervised semantic segmentation. In *NIPS*, 2015.

[26] Y. Jafarian, Y. Yao, and H. S. Park. MONET: Multiview semi-supervised keypoint via epipolar divergence. In *arXiv*, 2018.

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

[28] G. Kim and E. P. Xing. On multiple foreground cosegmentation. In *CVPR*, 2012.

[29] A. Kowdle, S. N. Sinha, and R. Szeliski. Multiple view object cosegmentation using appearance and stereo cues. In *ECCV*, 2012.

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

[31] A. Kundu, Y. Li, F. Dellaert, F. Li, and J. M. Rehg. Joint semantic segmentation and 3d reconstruction from monocular video. In *ECCV*, 2014.

[32] A. Kundu, V. Vineet, and V. Koltun. Feature space optimization for semantic video segmentation. In *CVPR*, 2016.

[33] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *IJCV*, 2000.

[34] L. Ladicky, J. Shi, and M. Pollefeys. Pulling things out of perspective. In *CVPR*, 2014.

[35] A. Laurentini. The visual hull concept for silhouette-based image understanding. *TPAMI*, 1994.

[36] A. Laurentini. How many 2D silhouettes does it take to reconstruct a 3D object? *CVIU*, 1997.

[37] M. Leordeanu, M. Hebert, and R. Sukthankar. Beyond local appearance: Category recognition from pairwise interactions of simple features. In *CVPR*, 2007.

[38] W. Li, O. H. Jafari, and C. Rother. Deep object co-segmentation. In *arXiv:1804.06423*, 2018.

[39] W. Li, O. H. Jafari, and C. Rother. Deep object co-segmentation. In *arXiv*, 2018.

[40] J. J. Lim, A. Khosla, and A. Torralba. FPM: Fine pose parts-based model with 3D cad models. In *ECCV*, 2014.

[41] 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.

[42] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[43] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *TPAMI*, 2017.

[44] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 2015.

[45] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan. Image-based visual hulls. *SIGGRAPH*, 2000.

[46] P. Mukherjee, B. Lall, and S. Lattupally. Object cosegmentation using deep siamese network. In *arXiv*, 2018.

[47] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015.

[48] H. S. Park, E. Jain, and Y. Shiekh. 3D social saliency from head-mounted cameras. In *NIPS*, 2012.

[49] X. Qi, J. Shi, S. Liu, R. Liao, and J. Jia. Semantic segmentation with object clique potential. In *ICCV*, 2015.

[50] W. Ren, J. Pan, X. Cao, and M.-H. Yang. Video deblurring via semantic segmentation and pixel-wise non-linear kernel. In *ICCV*, 2017.

[51] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016.

[52] http://www.crowdsyncapp.com/.

[53] A. Shaji, A. Varol, L. Torresani, and P. Fua. Simultaneous point matching and 3D deformable surface reconstruction. In *CVPR*, 2010.

[54] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.

[55] http://rieff.ieor.berkeley.edu/rashomon/.

[56] H. Su, C. Qi, K. Mo, and L. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *CVPR*, 2017.

[57] N. J. B. W. I. S. K. Tae-Hyun Oh, Kyungdon Joo and S. B. Kang. Personalized cinemagraphs using semantic understanding and collaborative learning. In *ICCV*, 2017.

[58] L. Torresani, V. Kolmogorov, and C. Rother. Feature correspondence via graph matching: Models and global optimization. In *ECCV*, 2008.

[59] S. Tulsiani, S. Gupta, D. Fouhey, A. A. Efros, and J. Malik. Factoring shape, pose, and layout from the 2D image of a 3D scene. In *CVPR*, 2018.

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

[61] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki. SfM-Net: Learning of structure and motion from video. In *arXiv:1704.07804*, 2017.

[62] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3D object detection in the wild. In *WACV*, 2014.

[63] J. Yan and M. Pollefeys. Automatic kinematic chain building from feature trajectories of articulated objects. In *CVPR*, 2006.

[64] F. Zhou and F. D. la Torre. Deformable graph matching. In *CVPR*, 2013.

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

[66] T. Zhou, P. Krähenbühl, M. Aubry, Q. Huang, and A. A. Efros. Learning dense correspondence via 3D-guided cycle consistency. In *CVPR*, 2016.

[67] T. Zhou, Y. J. Lee, S. X. Yu, and A. A. Efros. Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In *ICCV*, 2015.

[68] X. Zhu, C. Vondrick, C. C. Fowlkes, and D. Ramanan. Do we need more training data? *IJCV*, 2015.

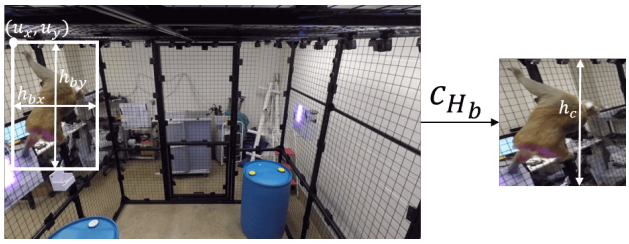

Figure 9. A cropped image is an input to the network where the output is the segmentation distribution with the same size. To rectify the segmentation distribution (heatmap), a series of image transformations need to be applied

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 Figure 9: 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:

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

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.

Figures 10–17 shows the prediction results on the unlabeled data using three models. Figures 18–25 shows how the semi-supervised network progresses on the unlabeled data during the training. Figure 27 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 10. Qualitative result of multiview segmentation.

Figure 11. Qualitative result of multiview segmentation.

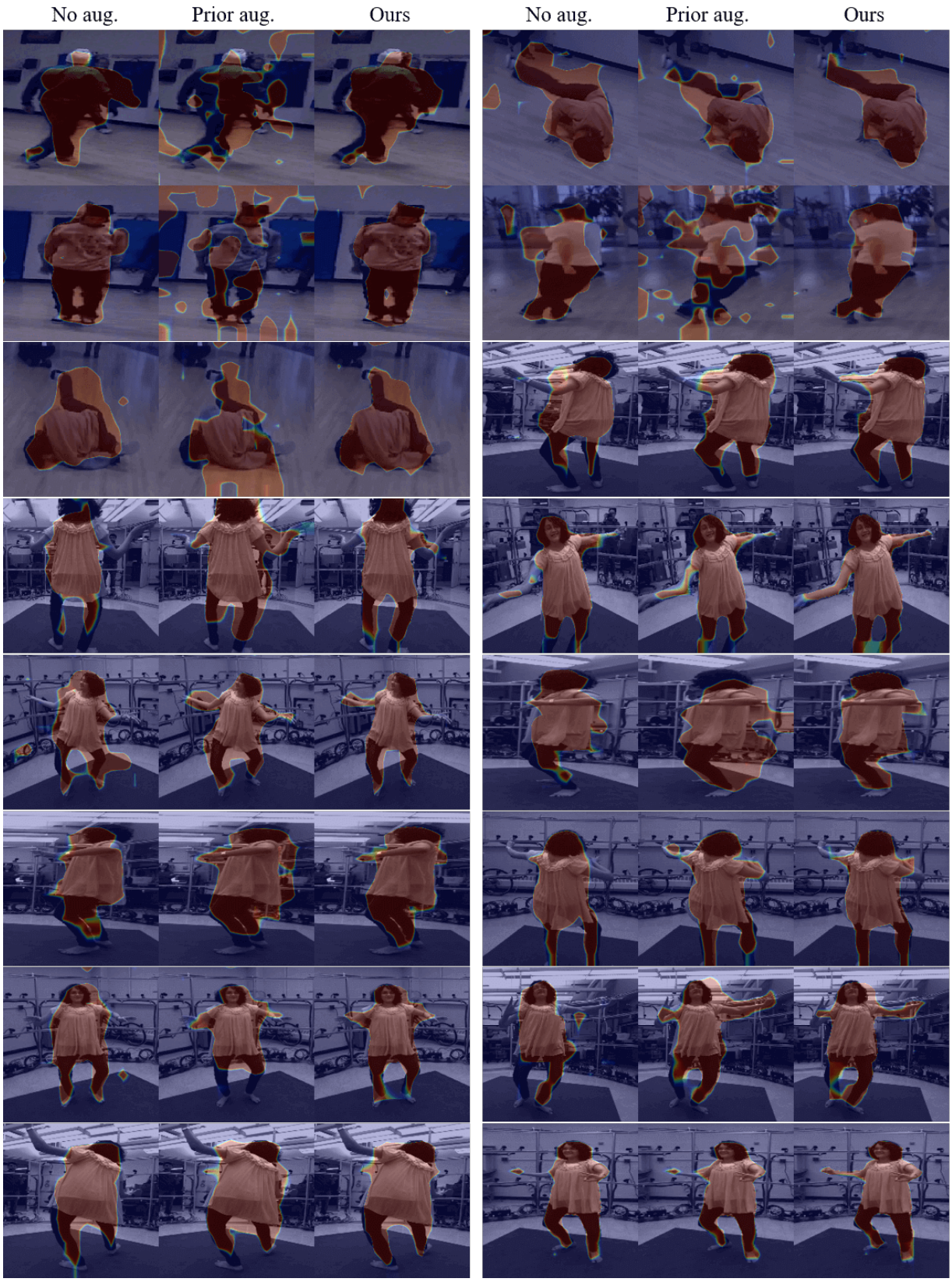

Figure 12. Qualitative result of multiview segmentation.

Figure 13. Qualitative result of multiview segmentation.

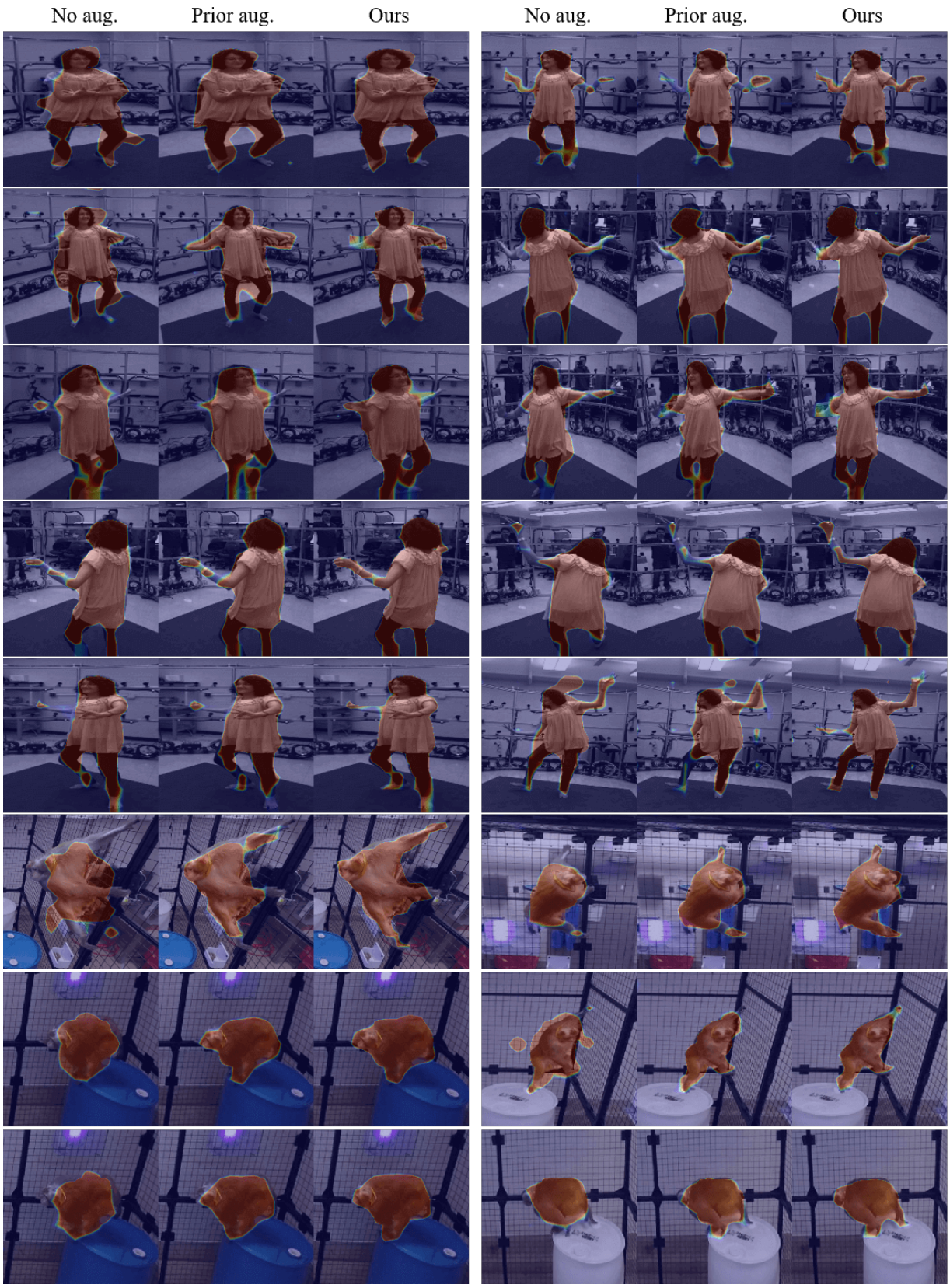

Figure 14. Qualitative result of multiview segmentation.

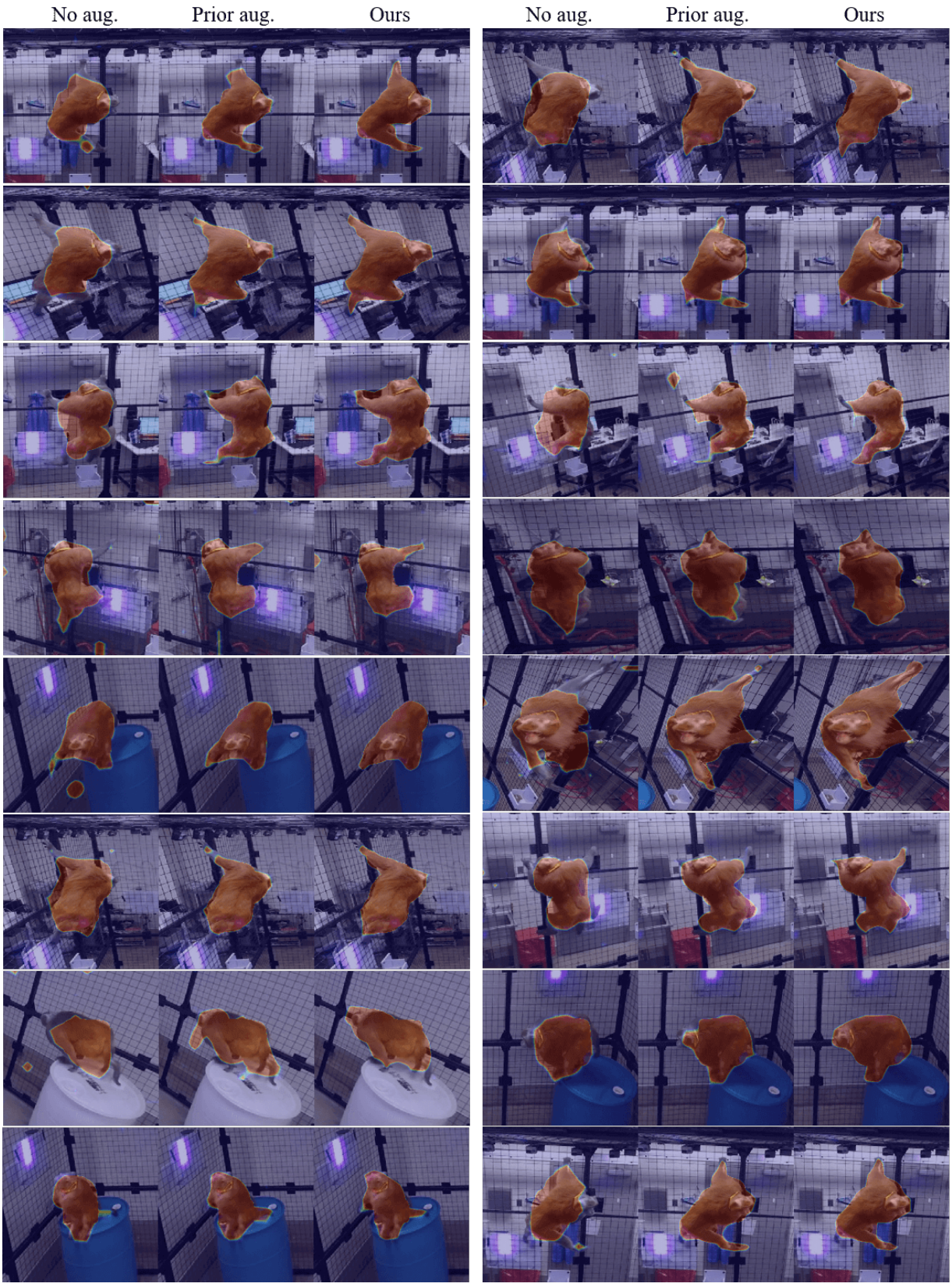

Figure 15. Qualitative result of multiview segmentation.

Figure 16. Qualitative result of multiview segmentation.

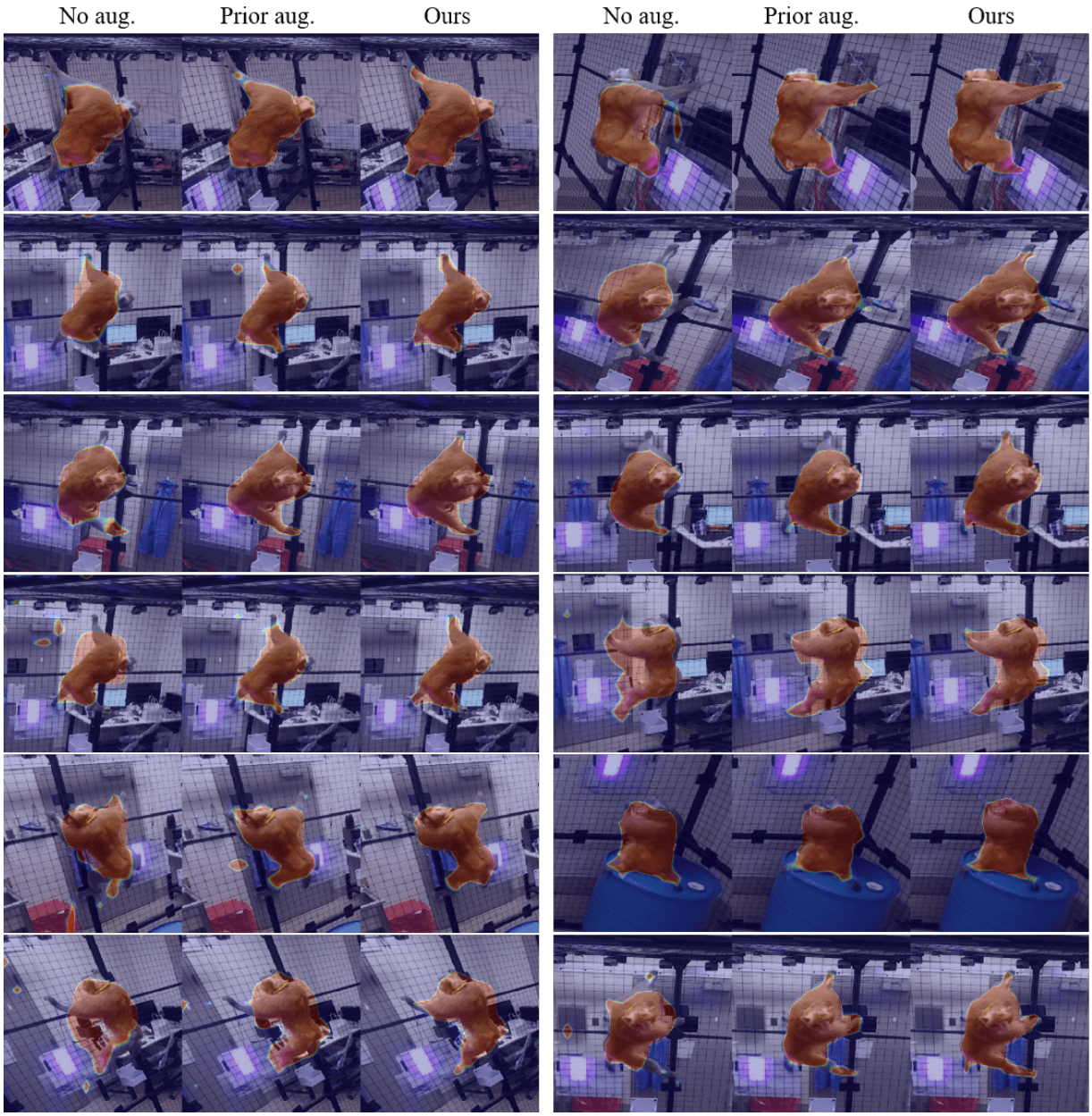

Figure 17. Qualitative result of multiview segmentation.

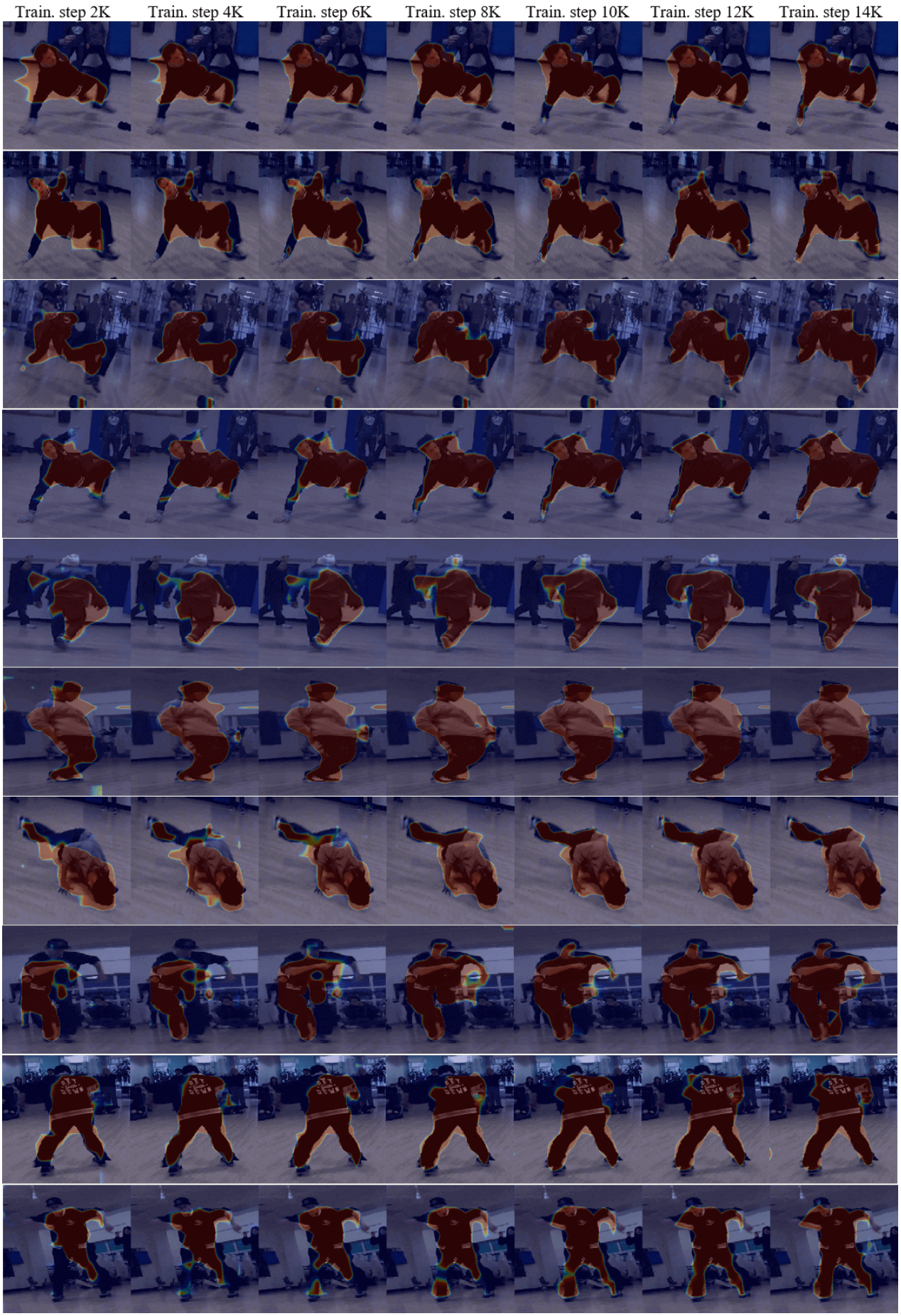

Figure 18. Qualitative result of multiview segmentation.

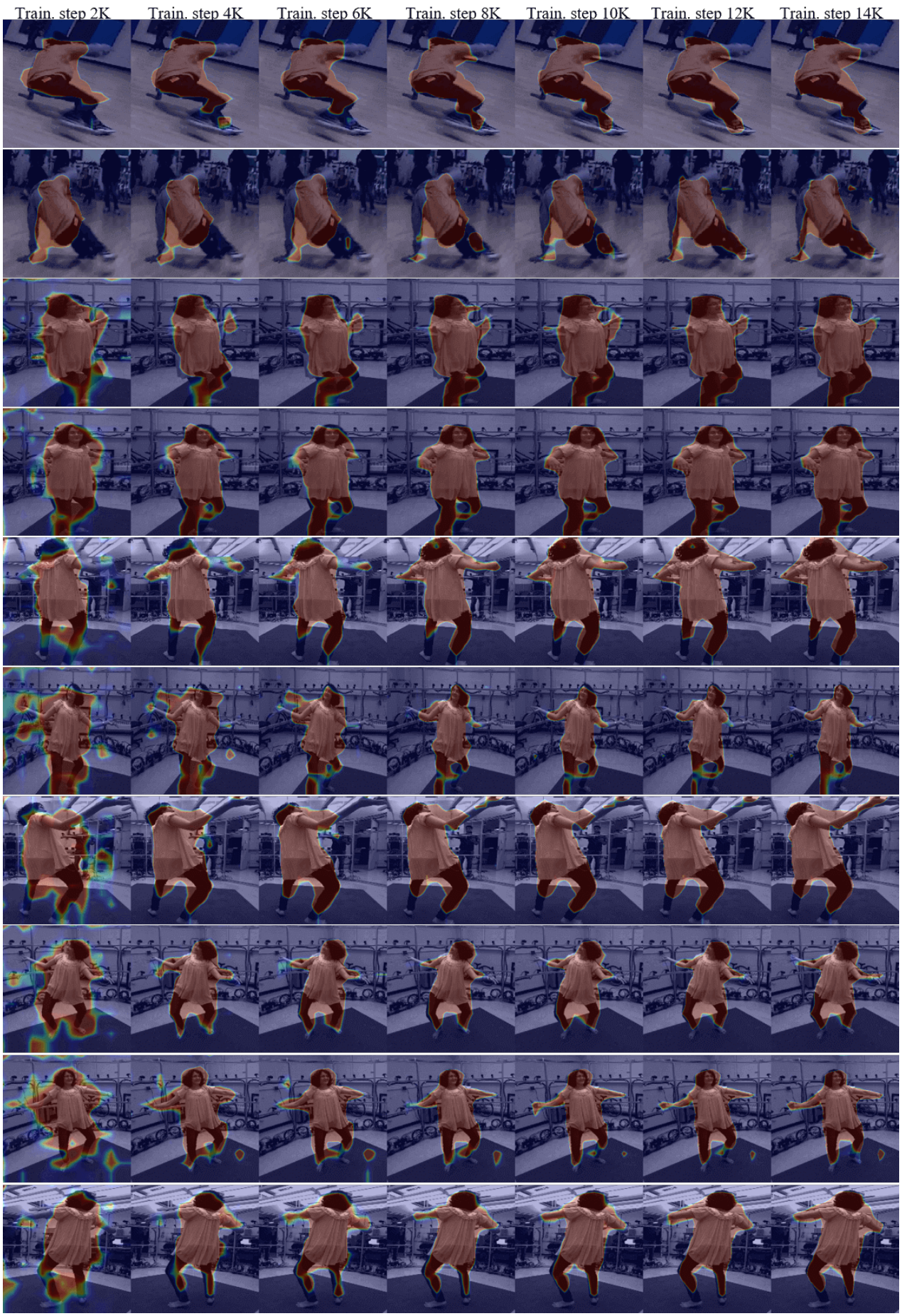

Figure 19. Qualitative result of multiview segmentation.

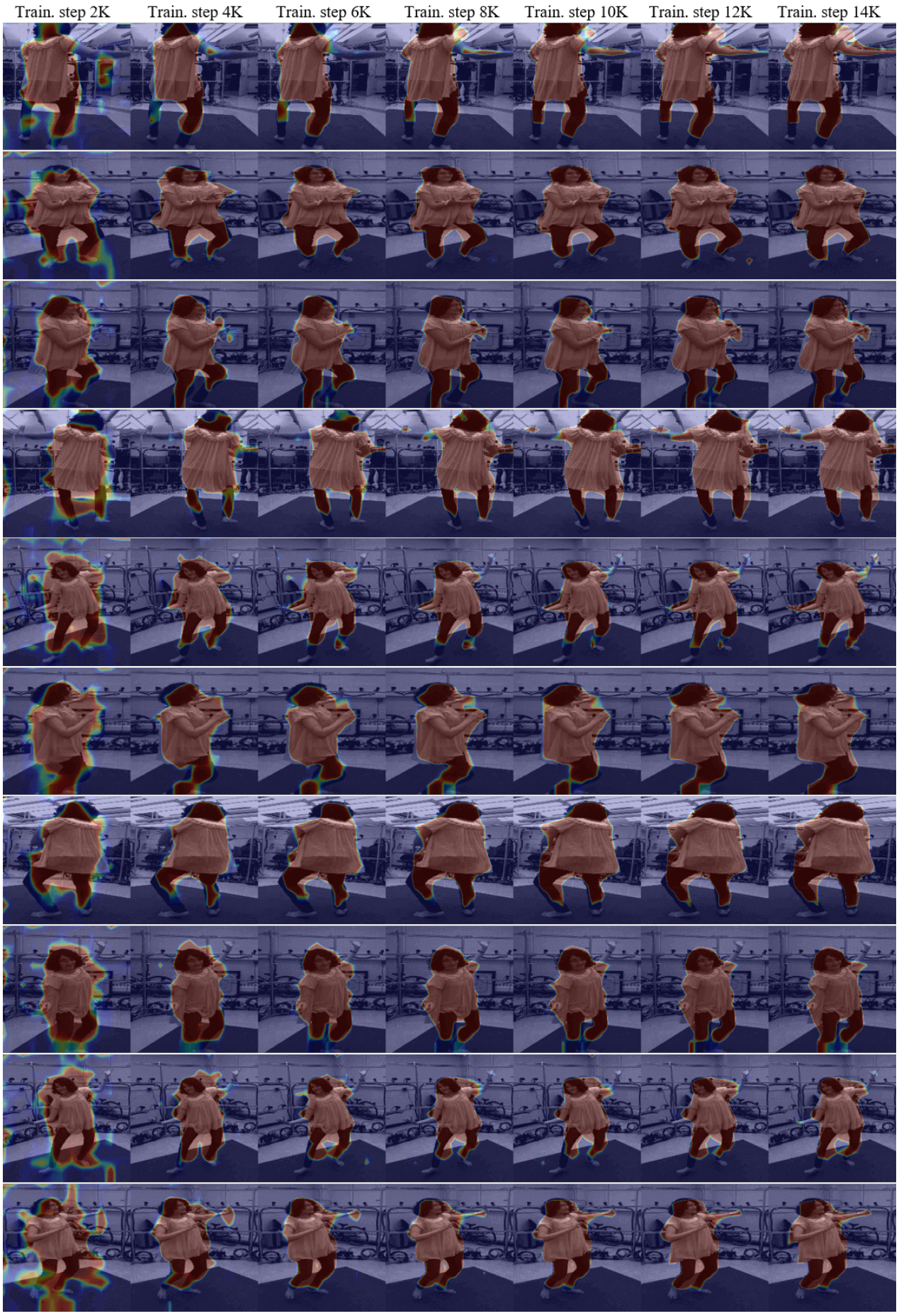

Figure 20. Qualitative result of multiview segmentation.

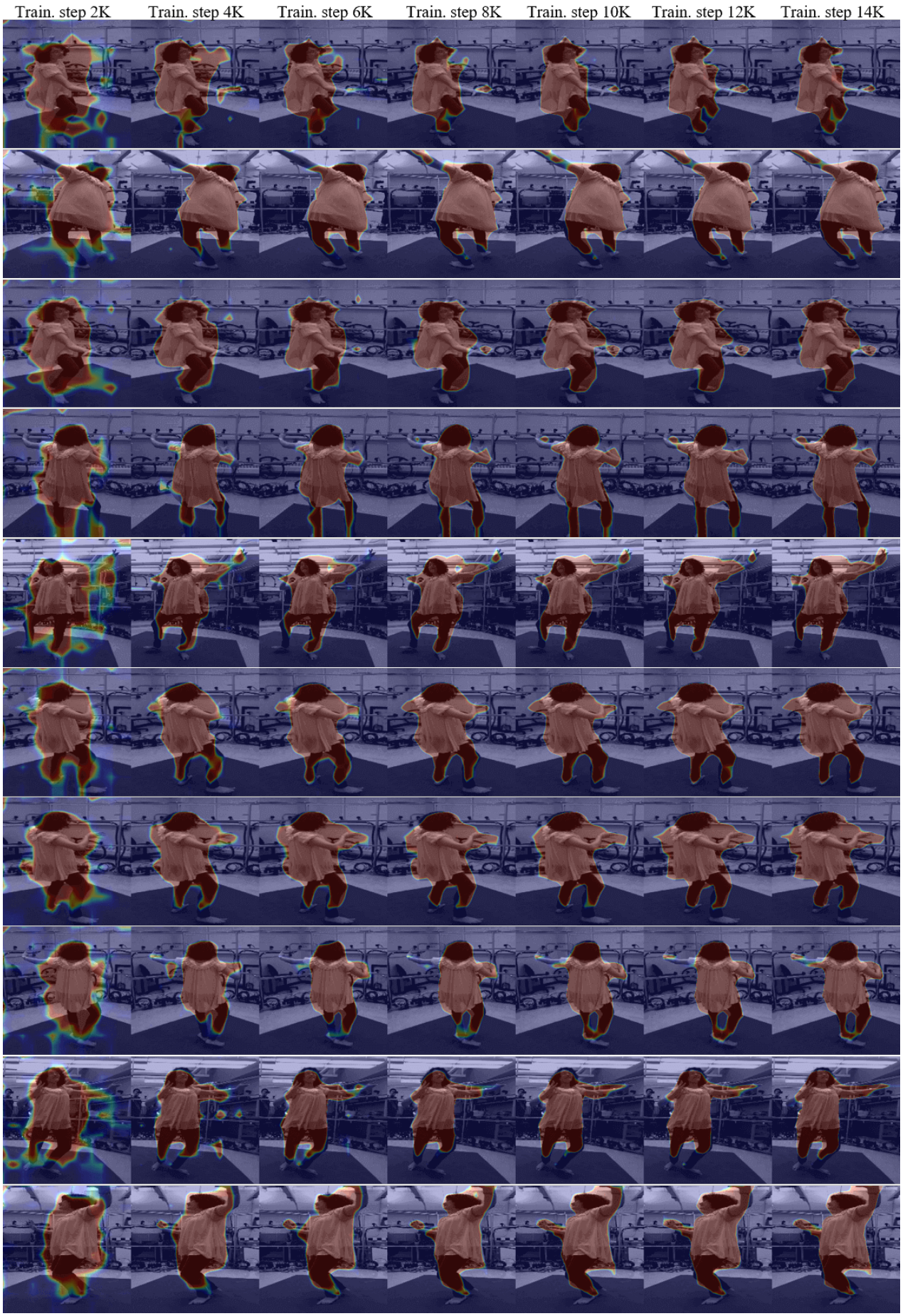

Figure 21. Qualitative result of multiview segmentation.

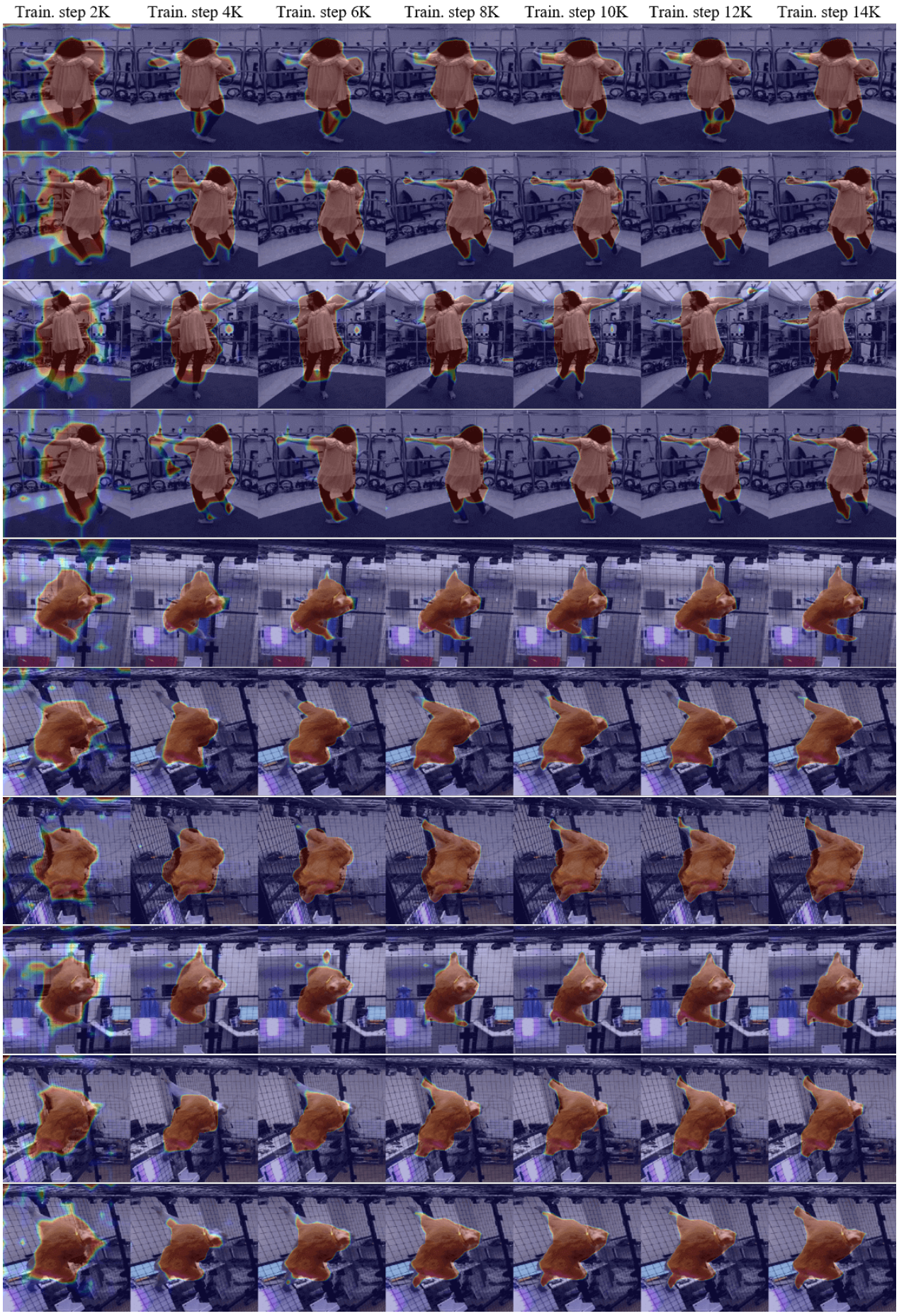

Figure 22. Qualitative result of multiview segmentation.

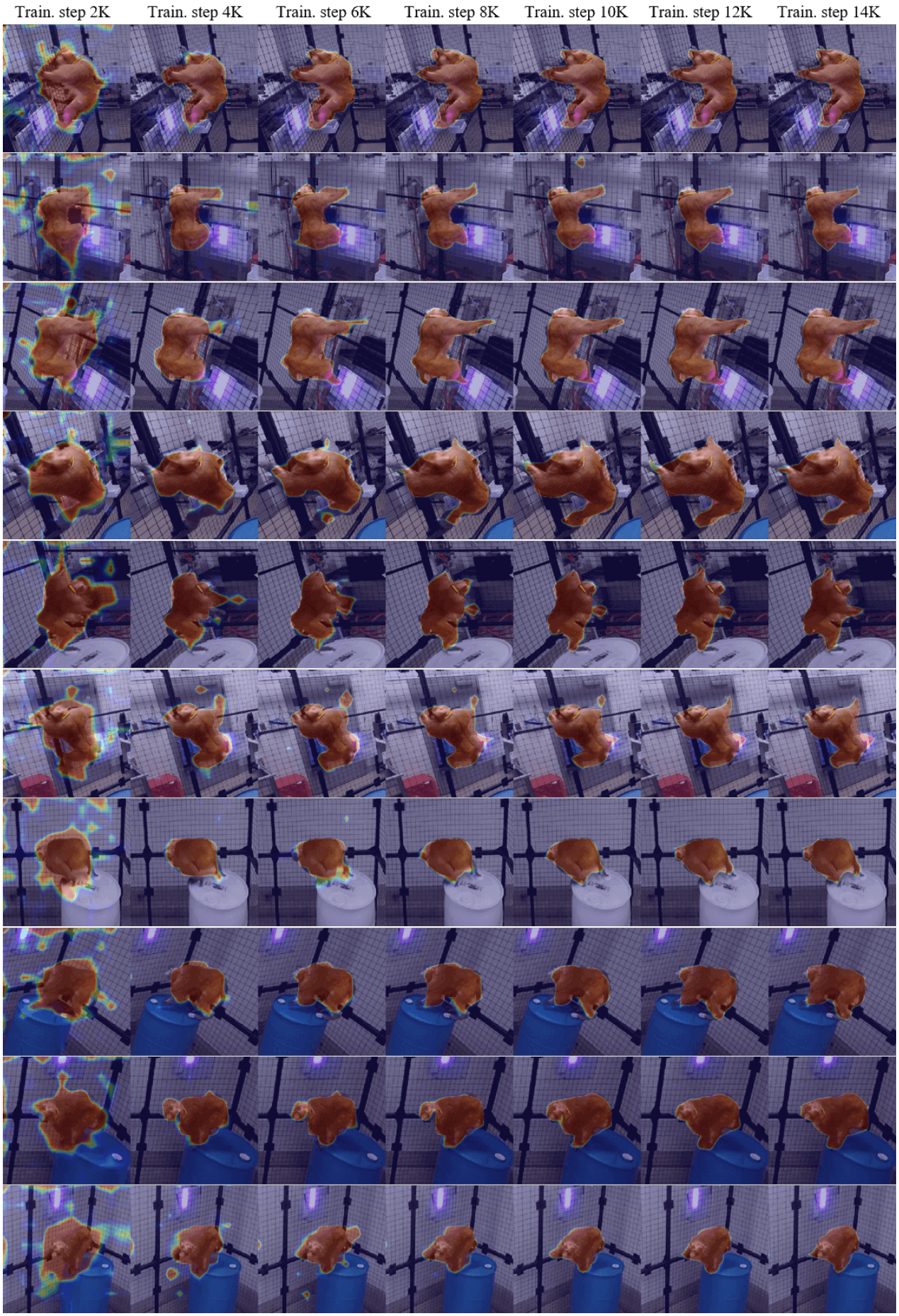

Figure 23. Qualitative result of multiview segmentation.

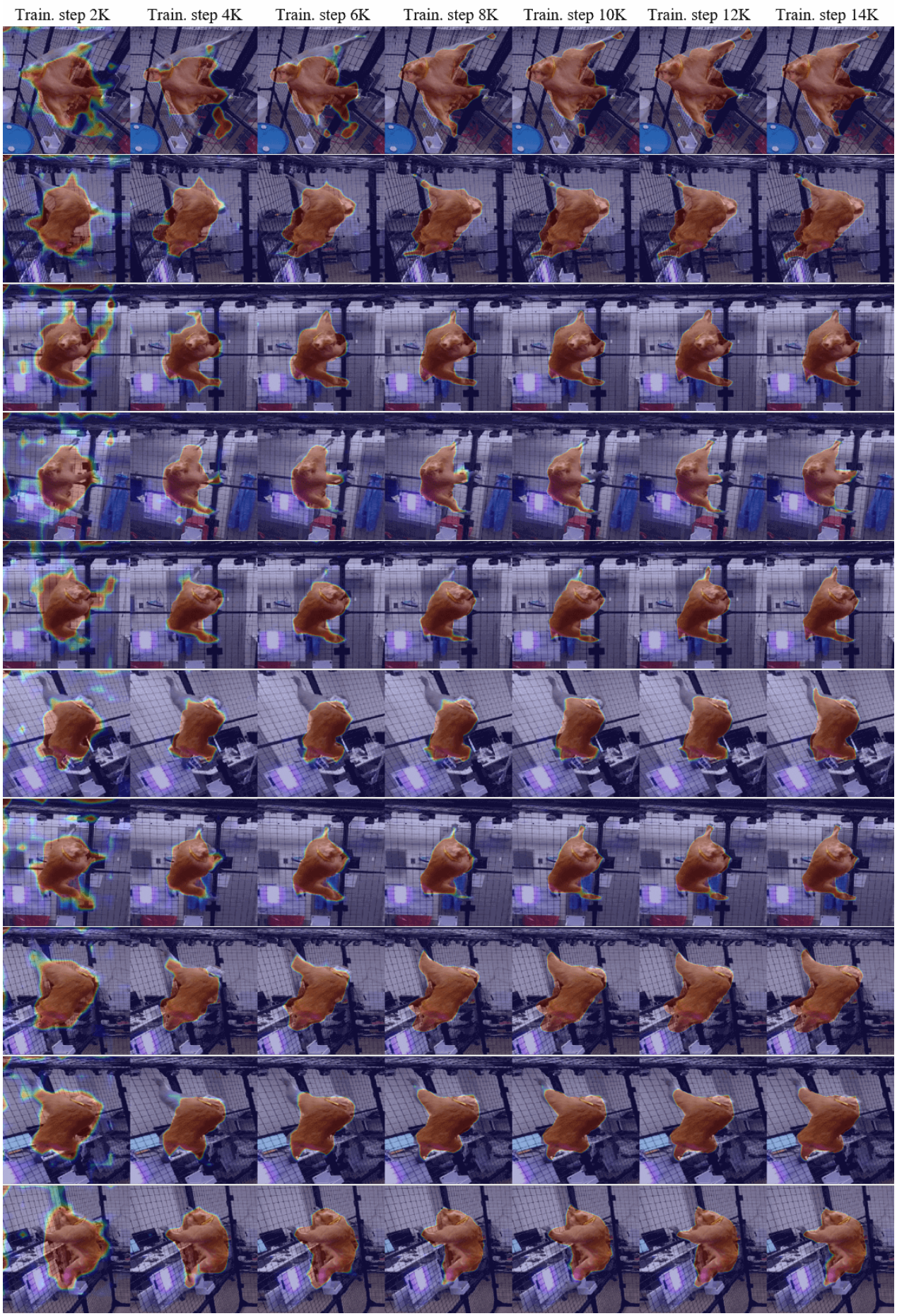

Figure 24. Qualitative result of multiview segmentation.

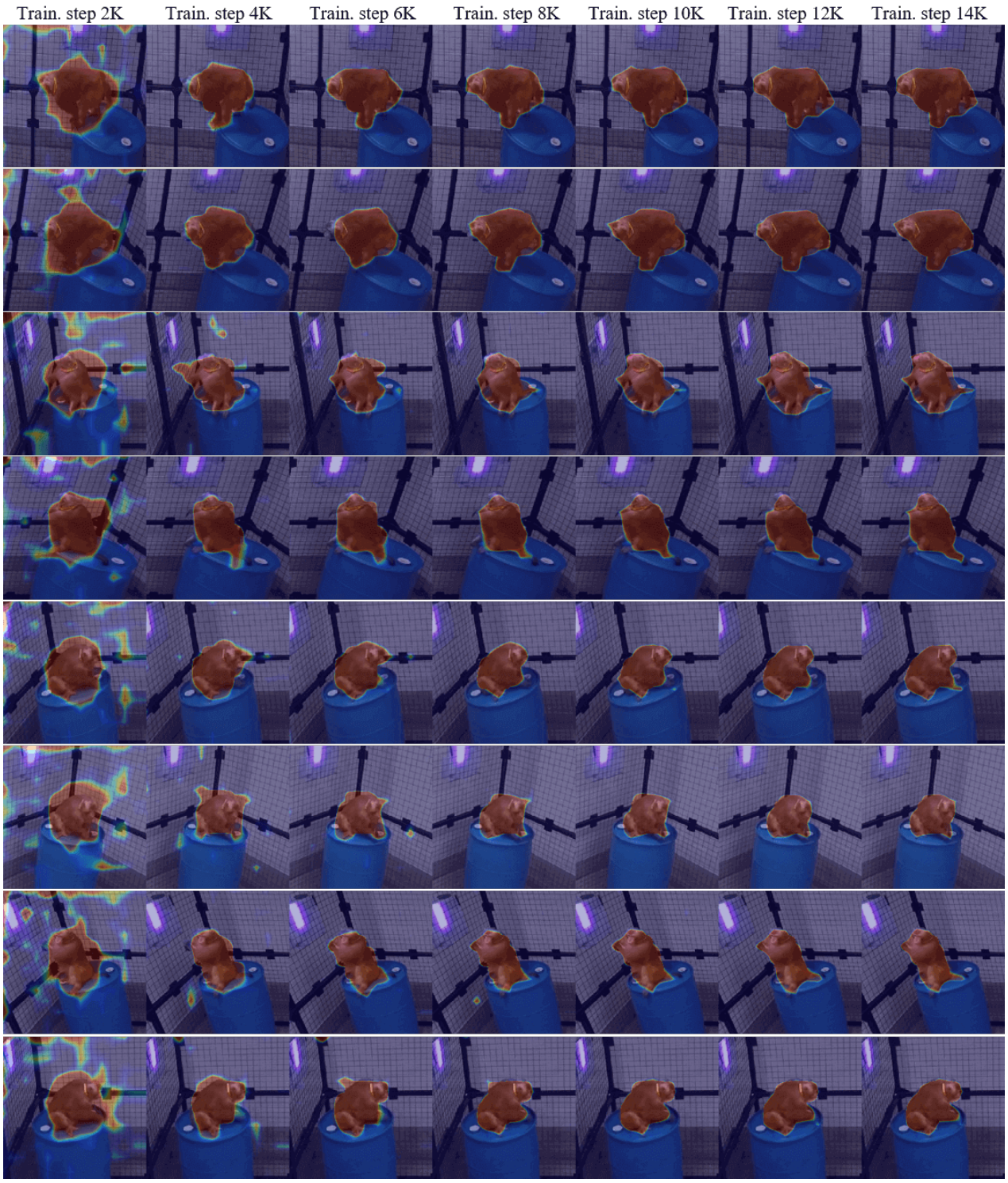

Figure 25. Qualitative result of multiview segmentation.

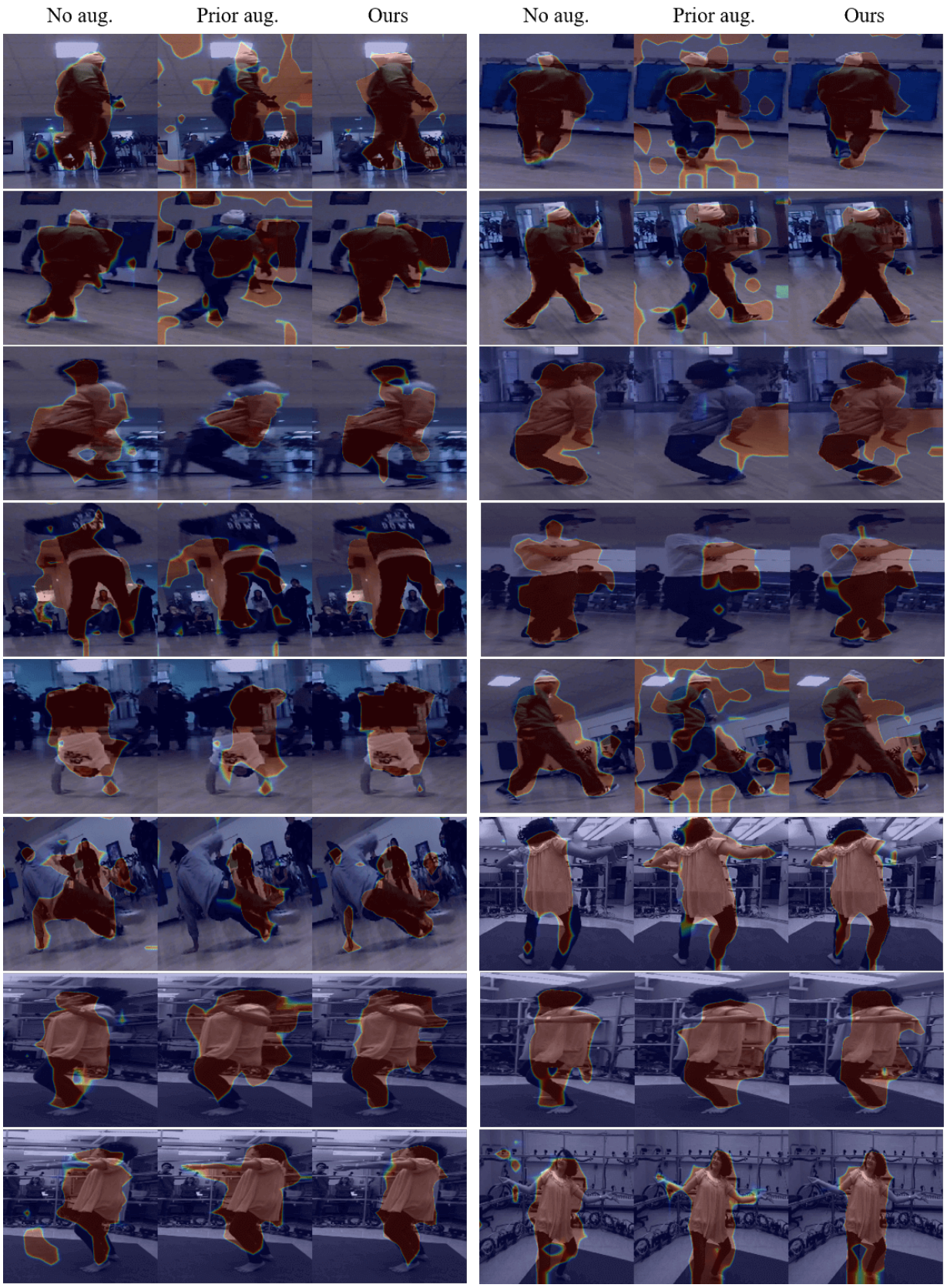

Figure 26. Qualitative result of multiview segmentation.

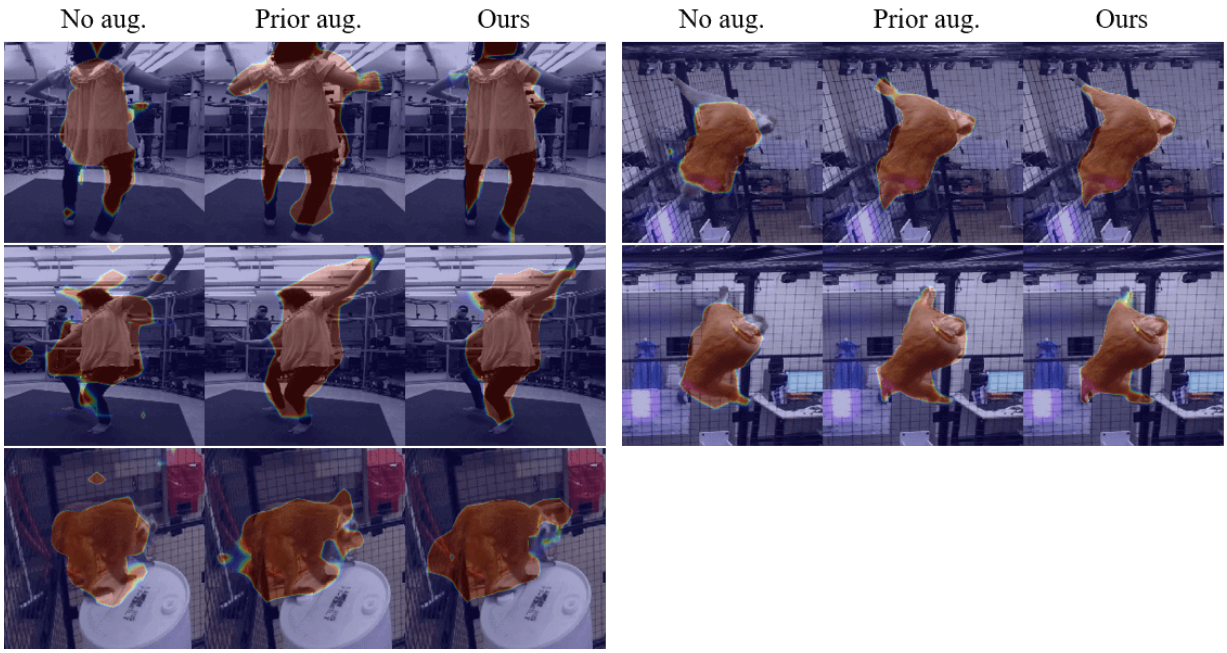

Figure 27. Failure cases.