Camera Selection for Occlusion-less Surgery Recording via Training with an Egocentric Camera

YUKI SAITO1, RYO HACHIUMA1,(Member, IEEE) HIDEO SAITO1,(Member, IEEE) HIROKI KAJITA2, YOSHIFUMI TAKATSUME2, AND TETSU HAYASHIDA.

1Keio University, 4-1-1 Hiyoshi, Kohoku-ku, Kanagawa, 223-8521, Japan
2Keio University School of Medicine, 35 Shinanomachi, Shinjuku-ku, Tokyo 160-8582, Japan
{yusa19971015, ryo-hachiuma, hs, jmbrx767, tsme}@keio.jp, tetsu@z7.keio.jp

Corresponding author: Yuki Saito (e-mail: yusa19971015@keio.jp).

This work was supported by MHLW Health, Labour and Welfare Sciences Research Grants Research on Medical ICT and Artificial Intelligence Program Grant Number 20AC1004, the MIC/SCOPE #201603003, and ROIS NII Open Collaborative Research 2020-20S0404.

ABSTRACT Recording surgery is an important technique for education and the evaluation of medical treatments. However, capturing targets such as the surgical field, surgical tools, and the surgeon’s hands, is almost impossible since these targets are heavily occluded by the surgeon’s head and body during a surgery. We used a recording system in which multiple cameras are installed in the surgical lump, supposing at least one camera would capture the target without occlusion. As this system records multiple video sequences, we address the task to select a best view camera automatically. Recently, learning-based approaches in a fully supervised manner have been proposed for this task, but these previous approaches completely rely on manual annotation of the training data. In this paper, we focus on the eye tracker mounted on the surgeon’s head, which can capture the recording targets without occlusion. Employing this first-person-view video synchronized with multiple videos of the surgical lump, we propose a novel camera selection approach using a self-supervised learning framework. In experiments, we created a dataset composed of four different breast surgery. Our extended experiments showed that our approach successfully switched to the best camera view without manual annotation and achieved competitive accuracy compared with conventional supervised methods. Also, our approach yielded effective visual representations comparable to state-of-the-art self-supervised learning frameworks.

INDEX TERMS Surgery Recording, Camera Selection, Deep Neural Networks, Variational Auto Encoder, Ego-centric vision

I. INTRODUCTION Recording surgery with a camera is an indispensable task for various reasons, such as education, sharing surgery techniques, and the evaluation of medical treatments [1]–[4]. The recording should contain essential information about the surgery, such as surgical fields, the doctor’s hand movement, and medical tools. In the surgical environment, it is almost impossible to capture these targets continuously using typical cameras, since there are several medical doctors and nurses around the surgical field and these targets are often occluded by them.

Shimizu et al. [1] proposed a novel surgical recording system using a surgical lump mounted right above the surgical field. A surgical lump in an operating room has multiple light bulbs illuminating the surgical field from various directions and reduces the shadows caused by doctor. So, by installing multiple cameras on this surgical lump, Shimizu et al. proposed that at least one camera would capture the surgical field as long as the surgical field was illuminated.

In this system, multiple video sequences are recorded for a single surgery. Shimizu et al. also proposed a method to automatically select the camera with the best-view to produce a single video without occlusion. They assumed that the best-view camera would capture the scene based on the maximum area size of the surgical field. They employed image segmentation [11] which was trained by manually annotated
images for every surgery. Then, they applied optimization using the Dijkstra algorithm to avoid the cameras frequently switching. Also recently, the camera selection approach with deep neural networks (DNNs) has been proposed. Hachiuma et al. [4] proposed a fully supervised Convolutional Neural Network (CNN) to predict the best-view camera. Different from Shimizu et al., they considered that the motion or the pose of the doctor’s hands and surgical tools may sometimes be important for determining the best-view camera. So, they selected the camera based on annotated labels by humans, not simply calculating the area size of the surgical field.

However, both conventional approaches are of a supervised manner and completely rely on training data which is richly annotated by manual operations. So, if the surgery recording is long and the training data becomes very large, the manual annotation for the large dataset may require much more time and a significant workforce. Also, manual annotation for surgeries sometimes needs the perspectives of medical experts. This means that it may sometimes be difficult for some users working out of the medical field to annotate by themselves.

In this paper, we focus on a first-person-view video captured from an eye tracker mounted on a surgeon’s head during a surgery. The doctor’s perspective is the best position for capturing the surgical targets, and the video is guaranteed to capture the targets without occlusion for as long as the surgeon is gazing at the surgical field. However, these recording systems suffered from device limitations and strong motion blur caused by the movement of the doctor’s head [2], [16]. As such, the recordings were not high-resolution and therefore, not of sufficient video quality. In contrast, multiple cameras on the surgical lump can stably capture the targets without motion blur, but are sometimes limited by human’s occlusion and there is a need to select the best-view camera at every moment. Figure 1 shows the surgery scenes recorded by multiple cameras on the surgical lump and the surgeon’s first-person-view camera. We can see that the first-person-view video constantly captures the surgical targets, whereas some of the multiple cameras on the surgical lump are heavily occluded by the head. We consider that by combing the information from both sensors, we can compensate for the drawbacks of both sensors. Here, we propose a novel approach to selecting the best-view camera mounted on a surgical lump by choosing the camera whose view is most similar to the first-person-view image at that moment. To avoid complicated manual annotation for multiple videos of the surgical lump, we used only first-person-view videos.

We tackled this image retrieval task using DNNs. We employed deep variational auto encoder (VAE) [7] to extract the features of the first-person-view at the training phase. VAE can represent the input data as compact and effective stochastic embeddings in a self-supervised manner. We also adopted various data augmentations in the training phase. We considered there are strong modality distinctions between
the doctor’s first-person-view images and the surgical lump images. Since we only use first-person-view images for training data, these augmentations encourage image modalities captured from the first-person-view camera to be close to the ones captured from the surgical lump. Moreover, we propose a contrastive loss function to fill the gap between the images captured by different sensors. Finally, by comparing the features between the first-person images and multiple camera images at the test phase, we predict the best-view camera.

Different from previous supervised approaches [1], [4], our approach is based on a self-supervised manner and does not need manual annotation for multiple cameras on the surgical lump to create the training data. Also, since we only employ first- and third-person videos for training, the network architecture is more simple and reduces calculation costs compared with the previous learning-based approach [4]. Moreover, our approach performs simple but effective feature extraction compared with the state-of-the-art contrastive learning frameworks, which require large batch sizes [38], [48] and specialized network architectures [39], [40], [52], [53]. To the best of our knowledge, this is the first attempt to address the task of selecting the best-view camera for surgery recordings in a self-supervised manner.

In experiments, as there is no public dataset available for surgery recordings via multiple cameras and first-person-view camera, we recorded our dataset using the system proposed by Shimizu et al. [1] and eye tracker mounted on the surgeon’s head. We recorded four different breast surgeries at our university’s school of medicine. Both multiple videos on the surgical lump and first-person-view video were synchronized with each other. We trained and validated our network with this dataset, and evaluated our method with other supervised and self-supervised approaches qualitatively and quantitatively. The results show that our approach can successfully select the best-view camera for capturing the surgical target appropriately.

In summary, there are two main contributions in this paper: (1) this is the first work to address the camera selection task for surgery recordings without occlusion by using first-person-view video in a self-supervised manner, which enables annotation-free learning. (2) we conducted an extensive evaluation of the proposed method using a self-created dataset which includes a variety of breast surgeries recorded with multiple cameras and a first-person-view camera. Our approach is capable of selecting the best-view camera compared with other comparable supervised and self-supervised approaches.

This paper is composed of 6 chapters. Chapter I is the Introduction, Chapter II is Related Works, Chapter III is on our Proposed Method, Chapter IV is Experiment Details, Chapter V presents the Results, and Chapter VI is the Conclusion.

### II. RELATED WORK

#### A. AUTOMATIC CAMERA SELECTION FROM MULTIPLE CAMERAS

Recording systems with multiple cameras are used in various situations such as sports games [12], [13], office environments [14], [15], and public places [24]. Since these systems obtain a huge amount of video sequences, only the necessary information must be selected by automatic viewpoint switching or video summarizing techniques.

For the surgery scenes, Shimizu et al. [1] firstly proposed a recording system with multiple cameras mounted on a surgical lump. They designed the system with the assumption that at least one camera would capture the surgical targets without occlusion. They also proposed a camera selection algorithm based on egocentric image segmentation trained by manual annotation and Dijkstra optimization. In another perspective, Hachiuma et al. [4] proposed a fully supervised deep neural network to predict the best-view camera based on expert supervision. They applied manual annotation by an expert for training video sequences beforehand and trained their network to predict a ground truth camera label directly. However, both approaches need time-consuming manual annotation and involve high costs to make the training data. Therefore, we employed the first-person-view video instead of the complicated annotation beforehand.

#### B. JOINT UNDERSTANDING FIRST/THIRD PERSON VIEW DOMAIN

The egocentric (first-person) view and the exocentric (third-person) view are drastically different in nature and relating them has been a challenging problem in computer vision. The relationship has been explored in many tasks such as human identification [20], [21], semantic segmentation [22], action recognition [18], [19] and cross-view image synthesis [17], [23]. However, these prior works required paired datasets of first/third person views that are simultaneously recorded to train CNNs. Recently, Li et al. [55] proposed a feature learning approach to learn egocentric video representations from third-person video datasets for pre-training on egocentric video understanding. They considered latent embeddings in third-person-view videos approximated egocentric-specific properties and leveraged only third-person video datasets augmented with pseudo-labels to learn egocentric video representations.

In this paper, we address the image retrieval task for surgery recordings with the first-person-view camera and multiple cameras on a surgical lump. Similar to Li et al. [55], our key insight is that a doctor’s ego-video would display various exocentric cues and help to classify the non-occluded frames among third-person-view videos. Therefore, we only employed the first-person-view video, instead of using both first- and third-person videos to train our network. To bridge the domain gaps between multiple videos and the first-person-view videos, we adopted various data augmentations to produce pseudo-exocentric images derived from only first-person-view videos.
C. SELF-SUPERVISED VISUAL REPRESENTATION LEARNING

To learn effective visual representation without human supervision, a number of learning approaches have been proposed for downstream tasks such as classification and detection. They are generally divided into two types: generative and discriminative.

Generative approaches obtain deep visual features by learning to generate or model pixels in the input space. For example, variational autoencoding models [7], [25], [26] learn disentangled visual representation in a low-dimensional latent space with conventional auto-encoding networks. Also, generative adversarial networks [27]–[32] learn to map from simple latent variables to arbitrarily complex data distributions with generator and discriminator networks.

Discriminative approaches calculate models to perform pretext tasks where both the inputs and labels are obtained from an unlabeled dataset [33]–[35]. Recently, breakthroughs in contrastive learning, such as Deep InfoMax [37], MoCo [39], and SimCLR [38], revealed the potential of discriminative models for representation. Contrastive learning aims at grouping similar samples closer and diverse samples far from each other with encoder networks and contrastive loss functions. They have shown great promise and achieved state-of-the-art results in various benchmark scores.

In this paper, we aimed to extract deep visual features of the first-person-view video with a self-supervised framework for the camera selection task. We used VAE [7], which enables simple and effective feature extraction through pixel-level generation. Our experiment in Chapter V showed that our proposed VAE achieved competitive accuracy compared with state-of-the-art self-supervised learning frameworks.

D. TRANSFER LEARNING

Transfer learning, i.e., fine-tuning CNN models pretrained from natural image datasets to train new tasks, has been studied in many computer vision tasks such as image classification [41], [42], object detection [43], [44], and image segmentation [45]. Pretraining with a large amount of annotated datasets contributed to shortening the training duration and improving the accuracy of models with a limited number of unlabeled data in the fine-tuning. In self-supervised learning, transfer learning is applied to boost the image reconstruction [7], [27], [46], [47] and the feature extraction ability [37]–[42]. Features learned using these methods demonstrated state-of-the-art accuracy in image classification and object detection compared with models trained from scratch.

For our network training, we employed ImageNet [9] pretrained model for the initialized weight of ResNet-18 [5] in VAE. Although there is a large gap between the image modalities of ImageNet and our surgical datasets, these pretrained models successfully enhanced the feature extraction ability of our VAE.

III. METHOD

A. PROBLEM FORMULATION

Figure 2 shows the overview of our proposed method. Our objective is to select the images from multiple cameras mounted on the surgical lump that best match those from the first-person-view. For our network training, we only employed the first-person-view image sequences \( V = \{ V_1, V_2, \ldots, V_T \} \), where \( T \) denotes the number of frames in a sequence. This video is captured from a camera on the surgeon’s head and is fully synchronized with the surgical lump videos. We employ VAE based network architecture for extracting the features of the first-person images. For our network training, in order to compensate for the difference in image modalities by different sensors, we adopt self-augmented learning with contrastive loss. For the test phase, we input both surgical lump images \( I_t = \{ I_1^t, I_2^t, \ldots, I_N^t \} \) and the first-person-view image \( V_t \) into our network, where \( N \) denotes the number of cameras installed on the surgical lump. Then, by comparing the latent vectors of the surgical lump cameras \( z_{\text{cam}}^t \) and that of the first-person-view camera \( z_{\text{ego}} \) with Euclidean distance, we predicted the camera labels as follows:

\[
y_t = \arg \min_{1 \leq i \leq N} \| z_{\text{cam}}^t - z_{\text{ego}} \|_2
\]

B. NETWORK ARCHITECTURE

Figure 3 shows the detailed architecture of our proposed network. We employed a fairly standard variational autoencoder [7] using CNNs for both encoder and decoder. For the encoder, we employ ResNet-18 [5] as feature extractor. A variational component in the bottleneck of VAE is composed of two fully connected layers, which calculate the mean and standard vectors. By using these two vectors, the input image \( x \) is encoded into latent space \( z = Encoder(x) \sim q(z|x) \). Then, the latent vector \( z \) is decoded to an image which has the same resolution as the input image, \( \tilde{x} = Decoder(z) \sim p(x|z) \). For the decoder, we employ the inverse architecture of ResNet-18 using residual blocks, which is composed of convolutional layers and nearest neighbor interpolations with the shortcut connection.

C. LOSS FUNCTION

Firstly, we employed the VAE loss proposed in [7] composed of two parts: the reconstruction loss \( L_{\text{rec}} \) and KL divergence loss \( L_{KL} \). The reconstruction loss expressed in Equation 2 is the negative expected log-likelihood of the observations in \( x \). Also, KL divergence loss \( L_{KL} \) expressed in Equation 3 is the distance between the variational approximation \( q(z|x) \) and the gaussian distribution \( p(z) \):

\[
L_{\text{rec}} = -E_{q(z|x)} [\log p(x|z)], \quad (2)
\]

\[
L_{KL} = D_{KL}[q(z|x)||p(z)]. \quad (3)
\]

In addition to these standard components, we introduce the notion of self-supervised contrastive learning. In the training phase, we only use the first-person-view videos as training data. However, in the testing phase, video sequences...
**FIGURE 2.** The network architecture of our proposed method

**Encoder**

**Decoder**

**FIGURE 3.** The detailed architecture of our proposed variational autoencoder. The top figure indicates the encoder network, which is the same architecture as Resnet-18 [5], the bottom figure indicates the decoder network, which is the inverse architecture of Resnet-18.
from the surgical lump are also input into this network, which have significantly different modalities from the first-person-view video at various points, such as the illumination environment, camera poses, and visible regions. In order to make our network robust against this inconsistent data, we incorporate a random composition of data augmentation. Our aim is to generate imitated images of the surgical lump by only using first-person-view images. Therefore, we adopted several types of augmentations: one type is spatial/geometric transformation, such as center cropping, perspective transformation, and rotation, and another type is appearance transformation, such as color distortion (brightness, contrast, saturation) and gaussian blur. We avoid augmentations which cause strong perturbations such as random cropping, random cutout, and hue jittering. Then, both non-augmented images and augmented images are fed into the network and converted into mean vectors and standard deviation vectors. To obtain the consistent latent embedding with different sensors, we added contrastive loss for positive samples [10] using the Euclidean distance between mean vectors \( \mu \), \( \mu_{\text{aug}} \) expressed in Equation 4.

Thus, our final loss function is given by Equation 5, where \( \alpha \), \( \beta \), \( \gamma \) indicates the weighting parameters between each term.

\[
L_{\text{contrastive}} = \frac{1}{2} (\mu - \mu_{\text{aug}})^T (\mu - \mu_{\text{aug}}) \\
L_{\text{loss}} = \alpha L_{\text{rec}} + \beta L_{\text{KL}} + \gamma L_{\text{contrastive}}
\]  

(4)  

(5)

IV. EXPERIMENTS

A. DATASET

As there is no public dataset available for surgery recordings with multiple cameras and a first-person-view camera, we created our own dataset. The surgeries were recorded using the surgical lump proposed in Shimizu et al. [1] and a Tobii Pro3. Multiple videos captured from the surgical lump were recorded at 30 frames per second (FPS) and the first-person-view video was recorded at 25 FPS. These two videos were manually synchronized with each other. Also, the first-person-view frames were center-cropped using the surgeon’s gaze points recorded in Tobii Pro3, and both surgical lump frames and Tobii frames were resized into resolution \( 224 \times 224 \). Furthermore, first-person-view videos sometimes contain frames that did not capture the surgical targets, like the preparation scene (wearing a surgical gown, washing hands, and replacing the medical equipment). Since our approach only needs frames which contain the surgical target, we removed these unnecessary frames manually. In Table 1, we summarized the number of images used for training and test set after removing unnecessary frames.

We recorded four different breast surgeries at Keio University School of Medicine. Each surgery was about one hour long. The video recording of the patients was approved by the Keio University School of Medicine Ethics Committee, and written informed consent was obtained from all patients.

B. NETWORK TRAINING

We trained our model using a batch size of 64 and optimized using Adam [8] with a learning rate of \( 1.0 \times 10^{-4} \). The model converged after 30 epochs, which takes about 20 hours on a GeForce Quadro GV100. For Equation 5, we employed \( \alpha = 1 \), \( \beta = 1 \), \( \gamma = 10 \). The weights of ResNet-18 was initialized with the pretrained ImageNet [9].

For our data augmentation explained in Figure 4, we summarize the hyper-parameters of our implementation below. Each augmentation was applied with a 0.5 chance.

- **Center crop**: Randomly apply center cropping with a size from 180 to 200, then resize to resolution \( 224 \times 224 \).
- **Perspective transformation**: Randomly sample 4 points with a distortion scale of 0.5 and apply perspective transformation. Bicubic interpolation was employed for the transformation.
- **Rotate**: Randomly rotate the image in the range from -180 to 180 degrees.
- **Color jittering**: Apply brightness, contrast and saturation control randomly. Each factor was chosen uniformly in the range \([0.9, 1.65],[1.0, 1.35],[1.0 1.4]\), where a large disparity with 1.0 means potentially large changes.
- **Gaussian blur**: Apply a Gaussian filter using a random

---

**TABLE 1.** The number of frames for the training set and the test set in our dataset after removing unnecessary frames.

|        | train | test  |
|--------|-------|-------|
| surgery1 | 45,577 | 12,204 |
| surgery2 | 49,654 | 11,310 |
| surgery3 | 49,651 | 13,805 |
| surgery4 | 55,388 | 14,687 |
| total    | 200,270 | 52,006 |

---
radius from 0.5 to 1.5.

C. EVALUATION DETAIL

For test sequences, our goal is to select the best-view camera which contains essential information of the surgery. To achieve our objective, we selected the ground truth camera labels by a single expert as done in Hachiuma et al. [4]. However, they annotated only one label among five cameras at any moment, even though there were often several best-view cameras which have almost the same view. To provide a more appropriate evaluation, we annotated several ground truth frames, which allows for selecting quite similar frames. We summarized the chance rate of each test sequence in Table 2 as a reference. Since approaches of camera switching predict only one camera label and our ground truth had multiple annotated labels, we only used the precision score for quantitative evaluation metrics, with a higher score being better.

For Chapter V-A and V-B, we compared our approach with two prior works of camera switching. Also, we set one baseline approach using our VAE.

- Shimizu et al. [1]: A supervised algorithm to select camera indices with maximum area size of surgical field. At first, we calculated the surgical area size using image segmentation [11] and then optimized the sequence camera indices with maximum area sizes using the Dijkstra algorithm. The image segmentation was trained using image masks of the surgical fields, manually produced by a single annotator. The model was trained and tested using only the raw frames of surgical lump videos.

- Hachiuma et al. [4]: A fully supervised camera selection network which predicts the best-view camera directly. The network was trained with ground truth camera labels manually annotated by a single expert. The model was trained and tested using only the raw frames of surgical lump videos.

- Hachiuma et al. [4] w/o FC: A fully supervised camera selection network without an output layer. We extracted intermediate features with 64 dimensions just before the final fully-connected layer (FC). By comparing these features of surgical lump cameras and the first-person-view camera, we predicted the camera labels as in Equation 1. The model was trained with the raw frames of the surgical lump videos and tested with both the surgical lump videos and the first-person-view videos.

- Ours w/o aug: Our camera selection network trained without data augmentation and its corresponding contrastive loss. The model was trained as the primitive VAE [7] using only the raw frames of the first-person-view videos. Then, test was conducted with the raw frames of the surgical lump videos and the first-person-view videos.

Note that different from our approaches, Shimizu et al. and Hachiuma et al. were based on supervised learning, which requires rich manual annotations such as image segmentation and ground-truth labeling for training data. Our proposed method and our baseline approach only use the surgeon's ego-centric videos. We evaluated our method with these four approaches qualitatively and quantitatively.

Also, for Chapter V-F, we compared our VAE’s feature extraction ability with the following state-of-the-art self-supervised learning networks for image classification.

- CPC-v2 [49]: A self-supervised learning algorithm to learn representation by predicting the future segments in latent space with an autoregressive model. The network was composed of a base encoder and the PixelCNN decoder. The model was trained with a path size of $32 \times 32$ and the InfoNCE loss function [50].

- PIRL [53]: A contrastive learning algorithm to learn visual representation using pretext task solving jigsaw puzzles [34]. The network was composed of a base encoder and memory bank. An image was divided into a $3 \times 3$ grid and we extracted a patch of $64 \times 64$ from each of these grids. The model was trained with the NCE loss function.

- SimCLR [38]: A contrastive learning algorithm that maximizes the agreement between differently augmented views of the same data example in the latent space. The network was composed of a base encoder and a multilayer perceptron (MLP) with one hidden layer. The model was trained with the NT-Xent loss function with a temperature of 0.5.

- MoCo [39]: A contrastive learning algorithm training via momentum contrast, which substantially increases the amount of negative samples. The network was composed of query and key encoders with MLP headers. The model was trained with a cross-entropy loss function.

- PCL [52]: A self-supervised learning algorithm that unifies contrastive learning and k-means clustering. We employed the number of clusters $K = \{5000, 10000, 20000\}$. The network was composed of a base encoder and memory bank. The model was trained with the ProtoNCE loss function.

- SwAV [51]: A self-supervised learning algorithm to predict the cluster assignment of a view from the representation of another view, instead of comparing features directly as in contrastive learning. The assignment problem was solved with the Sinkhorn-Knopp algorithm [54]. The network was composed of a base encoder, an MLP header, and prototype vectors. Multi-crop (2x224 + 6x96) was applied to the training data.

- BYOL [40]: A self-supervised learning algorithm training with only positive samples. It contains an online network and the target network, with each network composed of a base encoder and two MLP headers. The model was trained with a mean squared error loss function.

Each network was trained with a batch size of 64 and
TABLE 2. The chance rate of the the test sequences in our dataset.

| sequence | chance rate |
|----------|-------------|
| surgery1 | 0.667       |
| surgery2 | 0.761       |
| surgery3 | 0.841       |
| surgery4 | 0.669       |
| Average  | 0.734       |

used optimized using Adam [8]. For the encoders, we used ResNet-18 [5] and we initialized the weight with the pretrained ImageNet [9]. Each model was trained with data augmentation, as described in Chapter IV-B. The MLP headers outputted latent vectors with 64 dimensions, and we predicted the camera labels as in Equation 1.

V. RESULTS

A. SEQUENCE-OUT EVALUATION

In this Sequence-Out setting, we employed 200,270 frames for the training set and 52,006 frames for the test set, as shown in Table 1. We also used 20% of the training data for the validation set. We used all four sequences for the training data. In other words, the surgery type was known at the test time, but the sequence was unknown. Although the surgery type used at the training and test stages was the same, the frame conditions were quite different. So, selecting the camera in such a setting was non-trivial.

The top of Table 3 shows the quantitative result of this setting. Our self-supervised approach with contrastive loss resulted in higher precision than our baseline approach. However, other supervised approaches (Shimizu et al., Hachiuma et al.) outperformed our proposed method. Since our unsupervised approach did not use any information about the surgical lump cameras beforehand, it can be seen that our approach had a significant disadvantage compared with the previous methods. Also, in terms of feature extraction, Hachiuma et al. w/o FC achieved almost the same precision score as Hachiuma et al. The image representations encoded with the supervised method could successfully distinguish occlusion-less frames and occluded frames by themselves.

Figure 5 (left) shows the qualitative results of our proposed method. Also, Figure 6 (left) shows the predicted camera frames in each method and Figure 7 shows the predicted camera frames and their corresponding first-person-view frames in all sequences. Our method can avoid selecting the highly occluded camera like other supervised approaches. However, it tends to fail to discriminate between the frames which contain slight occlusion. For example, in Figure 6, in the left most-column, our VAE selected a partially occluded frame, whereas other supervised approaches successfully selected the non-occluded camera. In addition, we summarized incorrect prediction examples of our method in Figure 8. Since our approach uses the first-person-view frames to predict the camera labels, it sometimes suffered from ill-conditioned first-person frames, such as inappropriately cropped frames, frames which contained over exposure, and frames with strong motion blur.

Furthermore, we evaluated the transfer learning performance of our VAE encoder and present the results in Table 4. We performed hyperparameter tuning of our dataset in Sequence-Out setting and selected the best hyperparameters in the validation set. Our method with fine-tuning significantly improved the precision compared with the pretrained Resnet-18 on ImageNet without fine-tuning, whereas our VAE without augmentation that we trained on our dataset underperformed compared to the pretrained Resnet-18 without fine-tuning. Also, for both our VAE with and without augmentation, the ImageNet pretrained models outperformed the randomly initialized models with He et al. [6]. The pretraining of our encoder on ImageNet contributed to the improvement of its feature extraction ability, and the fine-tuning of our dataset was valid only for our proposed VAE with augmentation.

B. SURGERY-OUT EVALUATION

To test the robustness of the proposed approach in a more challenging setting, we performed a Surgery-Out evaluation. In this experiment, we used three sequences among four videos for training data and used the other one sequence for test data. Since the surgery type used at the test time was absolutely unknown at the training time, this was a more difficult setting. We also used 20% of the training data as a validation set.

The bottom of Table 3 shows the quantitative result. Despite the difficulties of this setting, our proposed approach had competitive precision score compared with Shimizu et al. and Hachiuma et al. Also, our image representations in the latent space performed better prediction than the features trained with the supervised methods (Hachiuma et al. w/o FC). Considering that our approach did not use any images from the surgical lump and any manually annotated labels for the training, our approach had significant advantages over previous supervised approaches [1], [4] in terms of the cost of making the training data. While the precision of those supervised approaches in Surgery-Out setting is much lower than the ones in Sequence-Out setting, our self-supervised approach is less susceptible to this strict setting. Our method, which was not based on surgical camera information, naturally improved the robustness in this more challenging setting.

Also, Figures 5, 6 (right) show the qualitative results of our method. Different from Sequence-Out setting, the supervised approaches sometimes predict partially occluded frames incorrectly. However, our method can select the camera without occlusion appropriately. From a qualitative perspective, our approach can select the frames without occlusion as well or better than other supervised methods, even though our method does not need any surgical lump frames for training.

C. ANALYSIS OF LATENT EMBEDDINGS OF VAE

We also analyzed the effect of altering the latent vector size of our proposed VAE. Figure 9 shows the relationship...
FIGURE 5. Qualitative results of our method in Sequence-Out setting (left) and Surgery-Out setting (right). The predicted camera is highlighted in green and the ground truth cameras are highlighted in red. The column direction indicates time axis and the row direction indicates the camera indices (Cam. 0 ∼ 4).

FIGURE 6. The predicted cameras using different methods in Sequence-Out setting (left) and Surgery-Out setting (right). The ground truth cameras are highlighted in red. The column direction indicates time axis and the row direction indicates the methods.
between latent vector size and the precision score for all test sequences trained in Sequence-Out setting. The horizontal axis indicates the latent vector size and the vertical axis indicates the precision score. Overall, the precision score was highest when the latent vector size was 512. Also, the precision tends to increase higher as the latent vector dimension increases larger. It can be seen that the larger latent feature spaces enabled the network to provide more complicated and expressive representations of the first-person-view video, which allowed reasonable similarities between the different sensor data.

D. ANALYSIS OF DATA AUGMENTATION

For our method, we employed five different types of augmentation, expressed in Figure 4, which altered the modalities of the first-person-view frames to those of frames captured by the surgical lump. To investigate the impacts of these types of augmentation more specifically, we validated the effect of each type of augmentation independently. We trained these independently augmented models using the same conditions explained in Chapter IV-B. We also employed latent embeddings with 64 dimensions.

Table 5 shows the precision score of our four test sequences in Sequence-Out setting. From top rows to bottom rows, the results with single augmentation, pairs of each augmentation, combination of all augmentation are summarized. Appearance transformation (color jittering, Gaussian blur) is seen to have a more dominant effect than other geometric augmentation (center cropping, perspective transformation, rotation) overall. In particular, color jittering was the most effective augmentation in our models. Figure 10 shows color histograms of the surgical lump’s frames and first-person-view frames (with/without color jittering) from our dataset. By applying color jittering for first-person-view frames, the intensities were more similar to the surgical lump frames than the non-color distorted first-person-view frames. Since there was a significant difference of illumination conditions between the first-person-view video and the surgical lump’s videos, color distortion was a critical component in effectively learning inconsistent modalities captured by different sensors.

Also, the average precisions tended to improve as the number of augmentation increased. For the surgery1 and surgery3, scores tended to increase with the increase in the augmentations, whereas the scores tended to decrease for surgery4. When composing augmentations, the more diverse types of augmentation strongly change the modality of the original images. Therefore, the contrastive prediction task, i.e., minimizing Equation 4, becomes harder, but the quality of the latent representation is considered dramatically improved.

E. ANALYSIS OF LOSS PARAMETERS

For our network training, we employed the conventional VAE loss and contrastive loss with the balancing parameters $\alpha$, $\beta$, and $\gamma$. We analyzed the effect of these parameters by alternating the values of $\alpha$, $\beta$, and $\gamma$. We conducted the experiments under the same conditions explained in Chapter IV-B. We also used the latent embeddings with 64 dimensions.

Table 6 shows the precision score of our four test sequences in Surgery-Out setting. In particular, color jittering was the most dominant effect than other geometric augmentation (center cropping, perspective transformation, rotation) overall. In particular, color jittering was the most effective augmentation in our models. Figure 10 shows color histograms of the surgical lump’s frames and first-person-view frames (with/without color jittering) from our dataset. By applying color jittering for first-person-view frames, the intensities were more similar to the surgical lump frames than the non-color distorted first-person-view frames. Since there was a significant difference of illumination conditions between the first-person-view video and the surgical lump’s videos, color distortion was a critical component in effectively learning inconsistent modalities captured by different sensors.

Also, the average precisions tended to improve as the number of augmentation increased. For the surgery1 and surgery3, scores tended to increase with the increase in the augmentations, whereas the scores tended to decrease for surgery4. When composing augmentations, the more diverse types of augmentation strongly change the modality of the original images. Therefore, the contrastive prediction task, i.e., minimizing Equation 4, becomes harder, but the quality of the latent representation is considered dramatically improved.
FIGURE 7. The predicted cameras by our method in four test sequences of Sequence-Out setting. The ground truth cameras are highlighted in red. The column direction indicates time axis and the row direction indicates surgery type (surgery 1 ∼ 4).

sequences in Sequence-Out setting. The average precision had the best score when the loss parameter was $\alpha = 1$, $\beta = 1$, and $\gamma = 10$. We can see in the Table that the precision tended to be higher as we increased the value of $\gamma$ with fixed $\alpha$ and $\beta$ values. Our contrastive loss term $L_{\text{contrastive}}$ significantly contributed to filling the domain gap between the first-person-views and the surgical lump views for our model.

Also, as we increased the value of $\beta$ instead of introducing the contrastive loss term, the precision increased. These $\beta$-VAE [26] models made the latent distributions of the original first-person-view images and the augmented images strongly follow the same Gaussian distribution. Therefore, enhancing the effect of the KL divergence term $L_{KL}$ is considered to have almost the same effect of introducing our contrastive loss term $L_{\text{contrastive}}$.

F. COMPARISON WITH SELF-SUPERVISED NETWORKS

We compared our VAE’s feature extraction ability with seven state-of-the-art self-supervised learning frameworks: CPC-v2 [49], PIRL [53], SimCLR [38], MoCo [39], PCL [52], SwAV [51], and BYOL [40]. Table 7 shows the precision score of our four test sequences in Sequence-Out setting.

Our proposed VAE achieved the best score on average and had competitive accuracy with these contrastive learning approaches. Since these conventional self-supervised approaches were not originally designed to fill the domain gaps between first- and third-person-view videos, our method is considered slightly better at feature extraction.

It can be seen that the pixel-level image generation in our VAE had much computational cost than these contrastive learning approaches in terms of feature extraction. However, our network can be trained only with positive feature samples
and the simple l2-norm contrastive loss, which does not require large batch sizes [38], [48] and specialized network architectures like a memory bank [39], [53].

Also, our method adopted the l2-norm objective \( L_{\text{contrastive}} \) in the framework of vanilla VAE [7], different from employing fixed randomly initialized network [40]. In general, encoders trained with l2-norm objectives admit collapsed solutions, e.g., outputting the same vector for all images. By incorporating this simple contrastive loss term into the VAE model, which is a much more complicated task than conventional contrastive learning approaches, our method successfully filled domain gaps between the first- and third-person views and obtained better image representations.

### G. ANALYSIS OF RECONSTRUCTION ABILITY OF VAE

We evaluated the effects of imaging reconstructions on the feature extraction ability of our VAE. Figure 11 shows the relationship between the per-pixel reconstruction loss and the precision score for our test sequences in Sequence-Out setting. The left shows the results with the first-person-view images, and the right shows the results with the surgical lump camera images. The horizontal axis indicates the per-pixel reconstruction errors, and the vertical axis indicates the precision score. Each plot shows the model extracted after the training for 1~50 epochs. For each sequence, the precision score tended to increase as the reconstruction loss decreased along with the progress of the training. It can be seen that enhancing the reconstruction ability of our VAE boosted the feature extraction ability of our encoder and produced better image representations in the latent space.

Figure 12 shows the reconstructed images of our VAE for the first-person-view images and the surgical lump camera images in Sequence-Out and Surgery-Out settings. Also, we summarize the average per-pixel reconstruction errors of our VAE in Table 8. Since our VAE was trained only with first-person-view videos, the reconstruction ability of the surgical lump camera images tended to be inferior to that of our first-person-view images. Furthermore, the reconstructed images of our proposed VAE tended to be blurred and brighter than the reconstructed images of our VAE without augmentation. The results in Table 8 show that our augmentation, explained in Figure 4, is considered to have improved the reconstruction ability of our VAE.

### VI. CONCLUSIONS

In this paper, we addressed the task of selecting the best-view camera from multiple videos of surgery via first-person-view videos. Our approach is based on self-supervised learning using VAE and constructing the latent embeddings which enable the effective comparison of the features of images captured from different sensors. Our proposed method does not need any manual annotation for multiple video sequences.
TABLE 7. The quantitative results of the feature extraction ability of self-supervised learning frameworks in Sequence-Out setting. Precision was employed as the evaluation metric, so a higher value is better. We used the latent vectors with 64 dimensions.

| Method   | Architecture | Param (M) | surgery1 | surgery2 | surgery3 | surgery4 | Average |
|----------|--------------|-----------|----------|----------|----------|----------|---------|
| CPC-v2 [49] | ResNet-18    | 11        | 0.803    | 0.875    | 0.876    | 0.797    | 0.838   |
| PIRL [53]  | ResNet-18    | 11        | 0.798    | 0.836    | 0.854    | 0.759    | 0.812   |
| SimCLR [58] | ResNet-18    | 11        | 0.911    | 0.567    | 0.973    | 0.924    | 0.844   |
| MoCo [39]  | ResNet-18    | 11        | 0.746    | 0.760    | 0.882    | 0.927    | 0.829   |
| PCL [52]   | ResNet-18    | 11        | 0.764    | 0.894    | 0.891    | 0.789    | 0.835   |
| SwAV [51]  | ResNet-18    | 12        | 0.896    | 0.794    | 0.914    | 0.755    | 0.840   |
| BYOL [40]  | ResNet-18    | 11        | 0.872    | 0.681    | 0.928    | 0.687    | 0.792   |
| Ours       | ResNet-18    | 24        | 0.816    | 0.838    | 0.883    | 0.842    | 0.845   |

TABLE 8. Per-pixel reconstruction errors of the first-person-view images and the surgical lump camera images with our VAE for the test sequences. The mean squared error was employed as the evaluation metric, so a lower value is better. We used the latent vectors with 64 dimensions.

| Method                        | Tobii reconstruction error | Camera reconstruction error |
|-------------------------------|---------------------------|-----------------------------|
|                               | surgery1 | surgery2 | surgery3 | surgery4 | Average | surgery1 | surgery2 | surgery3 | surgery4 | Average |
| Ours w/o aug. (Sequence-Out) | 0.0135   | 0.0133   | 0.0166   | 0.0246   | 0.0170   | 0.0218   | 0.0207   | 0.0324   | 0.0247   | 0.0249  |
| Ours (Sequence-Out)           | **0.0078** | **0.0061** | **0.0117** | **0.0146** | **0.0101** | **0.0117** | **0.0110** | **0.0201** | **0.0163** | **0.0148** |
| Ours w/o aug. (Surgery-Out)   | 0.0173   | 0.0142   | 0.0164   | 0.0262   | 0.0185   | 0.0243   | 0.0246   | 0.0349   | 0.0254   | 0.0273  |
| Ours (Surgery-Out)            | 0.0086   | 0.0084   | 0.0121   | **0.0146** | 0.0109   | 0.0125   | 0.0122   | 0.0208   | **0.0161** | 0.0154  |

FIGURE 8. Examples of incorrect predictions by our method. The predicted camera is highlighted in green and the ground truth cameras are highlighted in red.

FIGURE 9. The relationship between latent vector size and precision score in our proposed VAE. We validated with Sequence-Out setting for all test sequences.

and reduces the cost of creating training data. Also, our VAE does not require large batch sizes and specialized architectures to obtain better image representations in terms of self-supervised learning. The results show that our approach has competitive accuracy in selecting the best-view camera as well as the other supervised approaches and state-of-the-art contrastive learning frameworks.

In future work, we will further validate the effective types of augmentation to more accurately match first-person-view frames to multiple camera frames. Also, we will investigate the effective learning approach by using sequential information from the first-person-view video to avoid frequent camera switching.
FIGURE 10. Histograms of pixel intensities (over all channels) for frames captured by surgical lump and first-person-view cameras. (a) ground truth camera frames of surgical lump, (b) first-person-view frames without color distortion, (c) first-person-view frames with color distortion. Vertical axes are normalized to [0-1] and all axes have the same range.

REFERENCES

[1] T. Shimizu, K. Oishi, R. Hachiuma, H. Kajita, Y. Takatsume, and H. Saito, “Surgery Recording without Occlusions by Multi-view Surgical Videos”, in International Conference on Computer Vision Theory and Applications, 2020, pp. 837–844.

[2] S. Matsumoto, K. Sekine, M. Yamazaki, T. Funahashi, T. Orita, M. Shimizu, and M. Kitano, “Digital video recording in trauma surgery using commercially available equipment”, Scandinavian journal of trauma, resuscitation and emergency medicine, vol. 21, no. 1, pp. 21–27, April 2013.

[3] A. Sadri, D. Hunt, S. Rhotaye, and A. Juma, “Video recording of surgery to improve training in plastic surgery”, Journal of Plastic, Reconstructive & Aesthetic Surgery, vol. 66, no. 4, pp. 122–123, April 2013.

[4] R. Hachiuma, T. Shimizu, H. Saito, H. Kajita, and Y. Takatsume, “Deep Selection: A Fully Supervised Camera Selection Network for Surgery Recordings”, in International Conference on Medical Image Computing and Computer Assisted Intervention, 2020, pp. 419–428.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition”, in International Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification”, in International Conference on Computer Vision, 2015, pp. 1026–1034.

[7] D.P. Kingma, and M. Welling, “Auto-encoding variational bayes”, in International Conference on Learning Representations, 2014.

[8] D.P. Kingma, and B. Jimmy, “Adam: A Method for Stochastic Optimization”, in International Conference on Learning Representations, 2015.

[9] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li and F.F Li, “ImageNet: A Large-Scale Hierarchical Image Database”, in International Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[10] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping”, in International Conference on Computer Vision and Pattern Recognition, 2006, pp. 1735–1742.

[11] C. Li, and K. Kitani, “Pixel-level hand detection in ego-centric videos”, in International Conference on Computer Vision and Pattern Recognition, 2013, pp. 3570–3577.

[12] J. Chen, K. Lu, S. Tian, and J. Little, “Learning Sports Camera Selection from Internet Videos”, in Winter Conference on Applications of Computer Vision, 2019, pp. 1682–1691.

[13] F. Daniyal, M. Taj and A. Cavallaro, “Content and task-based view selection from multiple video streams”, Multimedia tools and applications, vol. 46, no. 2-3, pp. 235–258, January, 2010.

[14] Q. Liu, Y. Rui, A. Gupta, and J. Cadiz, “Automating camera management for lecture room environments”, in International Conference on Human Factors in Computing Systems, 2001, pp. 442–449.

[15] P. Doubeik, I. Geys, T. Svoboda, and L.V. Gool, “Cinematographic rules applied to a camera network”, in Workshop on Omnidirectional Vision, Camera Networks and Non-Classical Cameras, 2004, pp. 17–29.

[16] A.G. Nair, S. Kamal, T.V. Dave, K. Mishra, H.S. Reddy, D.D. Rocca, R.C.D Rocca, A. Andron, and V. Jain, “Surgeon point-of-view record-
FIGURE 11. Relationship between the precision score and the reconstruction ability of our proposed VAE for the test sequences in Sequence-Out setting. The left figure shows the relationship between the precision score and the per-pixel reconstruction errors of the first-person-view images, and the right figure shows the relationship between the precision score and the per-pixel reconstruction errors of the surgical lump camera images. Each plot shows the model extracted after training for 1–50 epochs. We used the latent vectors with 64 dimensions.
FIGURE 12. Comparison of the reconstructed images from our dataset. The first row shows the input images from our test sequences. The second row shows the reconstructed images generated by our VAE without augmentation. The third row shows the reconstructed images generated by our proposed VAE. The predicted camera is highlighted in green, and the ground truth cameras are highlighted in red.
RYO HACHIUMA received the B.E., M.Sc.Eng., and Ph.D. degrees in information and computer science from Keio University, Japan, in 2016, 2017, and 2021, respectively. Since 2021, he has been a project Assistant Professor with the Faculty of Science and Technology, Keio University. His research interests include 3D object recognition, 3D object tracking, and simultaneous localization and mapping.

HIDEO SAITO received the Ph.D. degree in electrical engineering from Keio University, Japan, in 1992. Since 1992, he has been on the Faculty of Science and Technology, Keio University. From 1997 to 1999, he joined the Virtualized Reality Project at the Robotics Institute, Carnegie Mellon University, as a Visiting Researcher. Since 2006, he has been a Full Professor with the Department of Information and Computer Science, Keio University. His research interests include computer vision and pattern recognition, and their applications to augmented reality, virtual reality, and human–robotic interaction. His recent activities in academic conferences include being the Program Chair of ACCV 2014, the General Chair of ISMAR 2015, the Program Chair of ISMAR 2016, and the Scientific Program Chair of EuroVR2020.

HIROKI KAJITA received his M.D. degree from Keio University School of Medicine, Japan, in 2011. He has been a Project Instructor in the Department of Plastic and Reconstructive Surgery at Keio University School of Medicine since 2017. He has also been a Visiting Instructor at Keio University Faculty of Science and Technology, Department of Information and Computer Science since 2018.

TETSU HAYASHIDA received the M.D., Ph.D. degree from Keio University School of Medicine in 1998. He has been a lecturer in the Department of Surgery at Keio University School of Medicine since 2015. She specializes in surgery and drug therapy for breast cancer. He is a board certified surgeon and medical advisor of the Japanese Surgical Society. He is also a specialist, supervisor, and councilor of the Japanese Breast Cancer Society.

HIDEO SAITO received the Ph.D. degree in electrical engineering from Keio University, Japan, in 1992. Since 1992, he has been on the Faculty of Science and Technology, Keio University. From 1997 to 1999, he joined the Virtualized Reality Project at the Robotics Institute, Carnegie Mellon University, as a Visiting Researcher. Since 2006, he has been a Full Professor with the Department of Information and Computer Science, Keio University. His research interests include computer vision and pattern recognition, and their applications to augmented reality, virtual reality, and human–robotic interaction. His recent activities in academic conferences include being the Program Chair of ACCV 2014, the General Chair of ISMAR 2015, the Program Chair of ISMAR 2016, and the Scientific Program Chair of EuroVR2020.

HIROKI KAJITA received his M.D. degree from Keio University School of Medicine, Japan, in 2011. He has been a Project Instructor in the Department of Plastic and Reconstructive Surgery at Keio University School of Medicine since 2017. He has also been a Visiting Instructor at Keio University Faculty of Science and Technology, Department of Information and Computer Science since 2018.

YOSHIFUMI TAKATSUME received the Ph.D. degree from Graduate School of Agriculture, Kyoto University, Japan, in 2007. From 2007 to 2017, he worked at META Corporation Japan Ltd., and was also a co-researcher at the Department of Anatomy, Keio University School of Medicine. He has been a Project Research Associate in the Department of Anatomy, at Keio University School of Medicine since 2020.