Invisible-to-Visible: Privacy-Aware Human Instance Segmentation using Airborne Ultrasound via Collaborative Learning Variational Autoencoder

Risako Tanigawa *,1, Yasunori Ishii †,1, Kazuki Kozuka ‡,1, and Takayoshi Yamashita §,2

1Technology Division, Panasonic Holdings Corporation
2Machine Perception and Robotics Group, Chubu University

Abstract

In action understanding in indoor, we have to recognize human pose and action considering privacy. Although camera images can be used for highly accurate human action recognition, camera images do not preserve privacy. Therefore, we propose a new task for human instance segmentation from invisible information, especially airborne ultrasound, for action recognition. To perform instance segmentation from invisible information, we first convert sound waves to reflected sound directional images (sound images). Although the sound images can roughly identify the location of a person, the detailed shape is ambiguous. To address this problem, we propose a collaborative learning variational autoencoder (CL-VAE) that simultaneously uses sound and RGB images during training. In inference, it is possible to obtain instance segmentation results only from sound images. As a result of performance verification, CL-VAE could estimate human instance segmentations more accurately than conventional variational autoencoder and some other models. Since this method can obtain human segmentations individually, it could be applied to human action recognition tasks with privacy protection.

1. Introduction

Recently, human action recognition has attracted wide attention [10, 44] because of its applications, such as automated surveillance [1], intelligent robots [23], and behavior monitoring in homes [3, 15, 16, 60]. Several sensors, such as cameras [60], electromagnetic waves [31], and wearable devices [2, 8, 47, 55, 66], have been employed to detect human activities. Among them, a camera is a human-friendly sensing device because we can intuitively understand the validity of recognition results by referring to the recorded videos. Although camera-based human action recognition has been widely investigated and achieved high precision, there are environments where it is difficult to recognize human actions with cameras, e.g., dark environments. In addition, camera images do not preserve privacy.

Privacy-preserved human action recognition has been investigated using cameras other than RGB, such as event cameras [4, 36, 65] and depth sensors [11, 30, 33, 45, 49, 56, 58, 69]. Even if we use an RGB camera, a method of recognizing human action by a shadow reflected on a white wall without directly capturing a person has been proposed [50]. Further, cameraless approaches for human action recognition have been achieved by wearable sensors and phone sensors [2, 8, 47, 55, 66]. Although these methods can detect the part of motions, detecting actions that require whole-body information are difficult. By contrast, whole-body segmentation has been achieved by electromagnetic waves, such as radars [32, 64], LiDARs [28, 37], and WiFis [25, 54, 59].
These methods can detect persons without human images captured by cameras. Although they are privacy-preserved sensing methods, it is difficult to capture the surrounding environment information.

Moreover, audio signals have been used to detect the surrounding environment information and humans without cameras [22, 29, 39, 51, 52, 62]. Predicting depth maps and segmentations from audio signals have been proposed [22, 22]. These methods can generate images from invisible physical information and can be applied for human action recognition by analyzing human segmentation images. Although sounding objects can be visualized by those methods, detecting nonsounding objects is difficult. From the human recognition perspective, people who do not make sounds, such as not talking or walking, cannot be detected.

Using airborne ultrasound echoes is a possible technique to detect nonsounding people. There are methods for detecting surrounding information by analyzing echoes [9, 13, 21, 35, 40, 41, 61]. Although these methods can estimate the position of objects, methods that specialize in human recognition and estimate human instance segmentation have not been well investigated. If the human instance segmentation can be estimated from echoes, it is possible to detect nonsounding people and it can be applied to estimate human action. In addition, if it is possible to estimate using narrow-band ultrasounds without interfering with other bands, it can be expected to be used in combination with environmental sound analysis.

In this study, we propose a method for estimating human instance segmentation from airborne ultrasound detected by multichannel microphones using a neural network. The concept of our work is illustrated in Figure 1. In the sensing section, an ultrasonic sensor at 62 kHz resonance, and 16 channels of microelectromechanical system (MEMS) microphone array are used. The sound emitted from the ultrasonic sensor is reflected by a human body and then captured by microphones. By analyzing the differences between multichannel audio signals, the reflected sound directional images (hereinafter, sound images) can be obtained. Since the sound image represents the intensity of the reflected wave at each pixel, it has ambiguous shapes that are far from the human shape. Therefore, we introduce a deep neural network to obtain human instance segmentation images from sound images. The accuracy of conventional methods, such as U-Net [48] and Mask R-CNN [19], reduces if the edges of objects between input and output are not similar, as mentioned in [18]. However, the sound images have different edge positions from the segmentation images. Thus, we propose a collaborative learning variational autoencoder (CL-VAE), which is based on variational autoencoder (VAE) [26]. The CL-VAE learns latent variables from both segmentation and sound images in the training phase and estimates human instance segmentation images from sound images in the test phase. Experiments showed that the human instance segmentation images could be generated from sound images. To the best of our knowledge, this is the first work to estimate human instance segmentation images from airborne ultrasounds.

The main contributions of this work are as follows:

- A human body sensing system using an airborne ultrasonic sensor and a microphone array for privacy-aware human instance segmentation.
- An architecture of CL-VAE, which learns latent variables from both segmentation and sound images in the training phase.
- A loss function for CL-VAE, including a reconstruction error, Kullback-Leibler (KL) divergence, and mean square error (MSE) of parameters from the encoders between segmentation and sound images.
- Creation of a dataset for evaluating CL-VAE.

2. Related works

2.1. Privacy-preserved human instance segmentation

Several RGB camera-based human instance segmentation methods have been developed [60]. Although the camera-based method has been well investigated and achieved high accuracy, privacy concerns should be considered for applications such as home surveillance.

Cameraless human instance segmentation methods have also been investigated. Wang et al. [54] proposed a method of estimating human instance segmentation images, joint heatmaps, and part affinity fields using WiFi signals. The channel state information [17] was analyzed using the method. They used three transmitting-receiving antenna pairs and thirty electromagnetic frequencies with five sequential samples. The networks mainly comprise upsampling blocks, residual convolutional blocks, U-Net, and downsampling blocks. Although this method achieved privacy-friendly fine-grained person perception with off-the-shelf WiFi antennas and regular household WiFi routers, it was not focused on environmental objects.

Alonso et al. [4] proposed an event camera based semantic segmentation method. The event camera senses information about pixels when the brightness changes, such as when the subject moves. Since event cameras only capture the changes in intensities on a pixel-by-pixel basis, they do not capture personal information more clearly than the RGB cameras. In [4], event information from event cameras was formed as 6-channel images. The first two channels were histograms of positive and negative events, whereas
the other four channels were the mean and standard deviation of normalized timestamps at each pixel for the positive and negative events. They showed that an Xception-based encoder-decoder architecture could learn semantic segmentation from the 6-channel information. Although this method achieves semantic segmentation from privacy-preserving event cameras, it is difficult to detect people who are not moving.

Irie et al. [22] proposed a method that generates segmentation images from sounds. They focused on the ability of humans to imagine the surrounding environmental scenes and developed the method. They recorded sounds using four-channel microphone arrays. To estimate segmentation images, they extracted Mel frequency cepstral coefficients and angular spectrum from sounds, which contain object types and their locations, respectively. This method can estimate both human and environmental objects only from sounds. However, in principle, it is difficult to estimate segmentation images for nonsounding people. Therefore, we focused on airborne ultrasonic sensing to achieve the detection of nonsounding people.

2.2. Airborne ultrasonic sensing

Airborne ultrasonic sensors have been used in various industries, such as the automobile [57] and manufacturing industries [7, 13, 24], to detect the distance from objects. Ultrasonic sensors emit short pulses at regular intervals. If there are objects on the propagation path, the ultrasounds are reflected. By analyzing the time differences between emitted and reflected sounds, the distance of objects can be detected [6].

In addition to the distances, the directions of objects can be detected when sounds are captured using microphone arrays [38]. Sound localization using beamforming algorithms has been developed. A common beamforming algorithm is the delay-and-sum (DAS) method [45]. The DAS method can estimate the direction of sound sources by adding array microphone signals delayed by a given amount of time. Considering a reflected position as a sound source, the positions of objects can also be estimated using ultrasonic sensors and microphone arrays. Although these methods can detect the position of objects, it is difficult to obtain the actual shape of objects from echoes of a single pulse.

To detect the shape of objects, Hwang et al. [21] performed three-dimensional shape detections using wideband ultrasound and neural networks. Although analysis of multiple frequencies can precisely detect the positions and shapes of objects, the speaker system for emitting wideband frequency sounds becomes large, and the amount of data increases due to the high sampling rate required to sense wideband ultrasound. Therefore, we consider that obtaining segmentation images from the positional information on reflected objects analyzed by narrowband frequency ultrasound increases due to the high sampling rate required to sense frequency sounds becomes large, and the amount of data.

3. Ultrasound sensing in proposed method

3.1. Ultrasound sensing system

The hardware setup is shown in Figure 2. The transmitter comprised a function generator and an ultrasonic sensor at 62 kHz resonance. The ultrasonic sensor was driven with burst waves with 20 cycles and 50 ms intervals at 62 kHz by the function generator. The receiver comprised a MEMS microphone array, an analog-to-digital converter, a field-programmable gate array (FPGA), and a PC. A 4 × 4 grid MEMS microphone array, whose microphones were mounted on a 30 mm² substrate at 3.25 mm intervals, was used. The analog signals captured by the microphones were converted to digital signals and imported to the PC via the FPGA. The distance between the microphone array and ultrasonic sensor was set to 30 mm.

3.2. Data preprocessing

The diagram of the data preprocessing is shown in Figure 3. First, a band-pass filter with a center frequency of 62 kHz and bandwidth of 10 kHz was applied for audio signals captured by the 16 microphones. The filtered audio signals were divided into blocks including direct and reflected waves. Then, we upsampled the soundwaves four times at each block because the 192 kHz sampling rate was inadequate to represent sounds of 62 kHz. Afterward, we produced sound images from reflected sounds via DAS beamforming. The beamformed signal $y$ can be defined as

$$y(t) = \sum_{m=1}^{M} x_m(t - \Delta_m), \quad (1)$$

where $t$ is the time, $M$ is the number of microphones, $x_m$ is signals received by the $m$-th microphone, and $\Delta_m$ is the time delay for the $m$-th microphone, which is determined by the speed of sound and the distance between the $m$-th microphone and observing points. We calculated beamformed
signals at the range of $\theta = -45–45$ degrees in the azimuthal direction and $\phi = -60–60$ degrees in polar direction. Then we obtained the reflected directional heat maps. To reduce the noises from reflected waves from objects, we calculated sound heat maps $H_{us}$ by subtracting a reference map $H_{ref}$ from reflected directional heat maps $H$ as

$$H_{us}(i, j) = H(i, j) - kH_{ref}(i, j), \quad (2)$$

where $(i, j)$ is the pixel of the heat maps, and $k$ is the coefficient, which is determined by

$$k = \frac{H(i_{max}, j_{max})}{H_{ref}(i_{max}, j_{max})}, \quad (3)$$

where $(i_{max}, j_{max})$ is the index of the maximum pixel of $H_{ref}$. Notably, the reference map was calculated using the data without humans, and $H_{us}$ was normalized as

$$X_{us}(i, j) = \begin{cases} 0, & H_{us}(i, j) < 0 \\ \frac{H_{us}(i, j)}{\max(H_{us})}, & H_{us}(i, j) \geq 0 \end{cases} \quad (4)$$

when it was converted to sound images $X_{us}$.

Examples of sound images are shown in Figure 4. The segmentation images in the top row were annotated from RGB images. The sound images, which represent the intensity of the reflected sound at each pixel, had ambiguous shapes at the positions corresponding to a person. Furthermore, there were artifacts in the region with no person.

4. Human instance segmentation via ultrasound

We first describe an overview of the proposed network. Second, we briefly explain VAE that is the basis of the proposed method. Then, we describe the proposed network in detail. Finally, we explain the loss function of our network.

4.1. Overview

As shown in Figure 4, the sound images do not have the shape of a person, and the edges are far different from those of the segmentation images. In conventional segmentation methods, such as U-Net [48], SegNet [5], and Mask R-CNN [19], the accuracies are reduced where the edges of the objects between input and output are not similar. Han et al. [18] stated that dividing regions from images with blurred edges was difficult. Thus, we consider performing human instance segmentation based on probabilistic models, such as VAE, rather than deterministic models. Kohl et al. [27] proposed a probabilistic U-Net combining VAE with U-Net. Since this method fed values sampled from the latent space to the last layer of U-Net, it could not handle the ambiguity of edges. Therefore, we propose CL-VAE, which introduces collaborative learning with the segmentation and sound images into VAE. In CL-VAE, both segmentation and sound images were input to an encoder, and means and variances of each image were calculated at a training phase. The loss includes the sum of the errors related to the distance between means and the distance between variances, in addition to the loss of VAE with the segmentation image. By learning with the loss, the model learned to adapt the latent variable of the sound images to that of the segmentation images. Consequently, both the segmentation and sound images are mapped to the common latent space; segmentation images can be estimated by generating images with a decoder from the latent variables sampled where the sound images are input into the model. The details of the proposed method are described below.

Figure 4. Example of sound images. The top row is the segmentation images generated by RGB images and the bottom row is the sound images.
4.2. VAE [26]

Since our network is based on VAE, we briefly explain it. VAE is a generative model and has an encoder-decoder architecture. The objective of VAE is to infer latent variables existing in datasets and generate data from the latent variables. Let $X$ be a data and $z$ be a latent vector, the joint probability $p_0(X,z)$ is defined as $p_0(X,z) = p_0(X|z)p_0(z)$, where $p_0(X|z)$ is the conditional distribution of $X$ given $z$, $p_0(z)$ is the prior distribution of the latent vector $z$, and $\theta$ is the parameter of the generative model. To infer the latent vector, the posterior $p_0(z|X)$ is calculated using Bayes's theorem: $p_0(z|X) = \frac{p_0(X|z)p_0(z)}{p_0(X)}$, where $p_0(X)$ is the marginal likelihood, given by $p_0(X) = \int p_0(X|z)p_0(z)\,dz$. The integral requires samplings from huge data. Therefore, the approximate distribution of the posterior distribution $q_0(z|X)$ is introduced in VAE. The approximate distribution $q_0(z|X)$ is learned as the encoder, and the generative model $p_0(X|z)$ is learned as the decoder.

The VAE learns the parameters of $\theta$ and $\phi$ by maximizing the marginal likelihood $p_0(X)$. The objective function $\log p_0(X)$ can be written as $\log p_0(X) = L(X, \theta, \phi) + D_{KL}(q_0(z|X) \parallel p_0(z|X))$, where $L(X, \theta, \phi)$ is the evidence lower bound (ELBO) and $D_KL$ is the KL divergence. Since the KL divergence is nonnegative, maximizing the ELBO, the reparameterization trick is introduced. The sampling of $z$ is alternatively performed by another random variable as $z = \mu + \epsilon\sigma$, where $\mu$ and $\sigma$ are the mean and variance of the posterior distribution $q_0(z|X)$, respectively, and $\epsilon \sim \mathcal{N}(0, I)$. The loss function of the VAE comprises the reconstruction error and a regularization term and can be written as $L_{VAE} = L_{RE} + D_{KL}(q_0(z|X) \parallel p_0(z|X))$, where $L_{RE}$ is the reconstruction error.

4.3. CL-VAE

The VAE can reconstruct images by obtaining the parameters of the probability distribution of the input dataset. However, sound images differ from human instance segmentation images. To overcome the problem, we considered that both the segmentation images $X_{seg}$ and sound images $X_{us}$ as input to the encoder and trained them to bring the distributions $q_\phi(z|X_{seg})$ and $q_\phi(z|X_{us})$ closer. The encoder $q_\phi(z|X_{seg}) \sim \mathcal{N}(\mu_{seg}; \sigma_{seg})$ is the distribution of the latent space to the input $X_{seg}$, represented by the means $\mu_{seg}$ and the variances $\sigma_{seg}$. The decoder $p_\phi(X_{seg}|z)$ generates plausible inputs for the image by sampling latent variables $z$ from the latent space. Hence, if the encoder is trained to map to the same point in the latent space regardless of whether segmentation or sound images are input, the images generated from the latent variables obtained by inputting the sound images become close to the segmentation images (see Figure 6). Since the encoder is represented by a Gaussian distribution with means and variances, it is possible to map to a latent space common to both images by matching the mean and variance of the segmentation and sound images.

The diagram of our network is shown in Figure 6. At the training phase, both the segmentation images $X_{seg}$ and sound images $X_{us}$ are input to the encoder. Then, each of them is encoded, and the means $\mu_{seg}, \mu_{us}$ and variances $\sigma_{seg}, \sigma_{us}$ are estimated from the posterior distributions $q_\phi(z|X_{seg})$ and $q_\phi(z|X_{us})$, respectively. The parameters $\mu_{seg}$ and $\sigma_{seg}$ are used to approximate the parameters $\mu_{us}$ and $\sigma_{us}$ by comparing their values. At the testing phase, the sound images are input to the encoder, and the segmentation images are obtained.

4.4. Loss functions

To train the distribution $q_\phi(z|X_{us})$ to become close to the distribution $q_\phi(z|X_{seg})$, we introduce MSEs of the means and variances of the distributions $q_\phi(z|X_{seg})$ and...
\( q_\theta(z|X_{\text{us}}) \). The loss is defined as

\[
L = \alpha \{ L_{RE} + D_{KL}(q_\theta(z|X_{\text{seg}}) \| p_\theta(z|X_{\text{seg}})) \} + (1 - \alpha) L_{\text{MSE}},
\]

The first term is the sum of the reconstruction loss and KL divergence, which are the same as in the conventional VAE. The second term is the MSE of means and variances. To adjust the scales of the VAE loss and the MSE loss, the coefficient \( \alpha \) is introduced. The reconstruction error \( L_{RE} \) is calculated as

\[
L_{RE} = \frac{1}{N} \sum_{n=1}^{N} (-x_{\text{seg},n} \log \hat{x}_{\text{seg},n} - (1 - x_{\text{seg},n}) \log(1 - \hat{x}_{\text{seg},n})),
\]

where \( x_{\text{seg}} \) is the value of input segmentation images, \( \hat{x}_{\text{seg}} \) is the value of reconstructed images, \( N \) is the dimension of input/output images, and \( n \) is the index of the dimension. The KL divergence \( D_{KL} \) is calculated as

\[
D_{KL}(q_\theta(z|X_{\text{seg}}) \| p_\theta(z|X_{\text{seg}})) = -\frac{1}{2} \sum_{d=1}^{D} \left( 1 + \log(\sigma_{\text{seg},d}^2) - \mu_{\text{seg},d}^2 - \sigma_{\text{seg},d}^2 \right),
\]

where \( D \) is the dimension of the latent variables \( z \), and \( d \) is the index of the dimension. The \( L_{\text{MSE}} \) is the MSE between \( \mu_{\text{seg}}, \sigma_{\text{seg}} \) and \( \mu_{\text{us}}, \sigma_{\text{us}} \) calculated as

\[
L_{\text{MSE}} = \frac{1}{D} \sum_{d=1}^{D} (\mu_{\text{us},d} - \mu_{\text{seg},d})^2 + \frac{1}{D} \sum_{d=1}^{D} (\sigma_{\text{us},d} - \sigma_{\text{seg},d})^2.
\]

5. Experiments

First, we describe the experimental setup. Then, the results of the experiments are explained.

5.1. Experimental setup

Implementation details The batch size was 128 and the initial learning rate was 0.001. We used an Adam optimizer with \( \beta_1 = 0.9, \beta_2 = 0.999 \) in training. We trained the network for 40 epochs. The coefficient of the loss function \( \alpha \) was set to 0.0001, which was experimentally determined by confirming the scales of \( L_{RE}, D_{KL}, \) and \( L_{\text{MSE}} \) in advance. Our network was implemented by PyTorch.

Datasets We created a dataset for our experiment because no datasets use airborne ultrasound to detect the human body so far. Written consent was obtained from participants in the data acquisition. We captured the ultrasounds at 192 kHz sampling from 16 channel microphones and videos at 30 frames per second (fps) from the RGB camera (a built-in camera of Let’s Note, CF-SV7, Panasonic), which was located 35 mm under the microphone array, for 10 s. The resolution was 180 × 120, and the videos were used for creating segmentation images used in a training phase. The data were extracted at 10 fps because the time interval of the ultrasonic sound generation was 20 bursts per second and the frame rate of the video was 30 fps. We produced segmentation images using Mask R-CNN [19]. We used the dataset that people, who were located from 1 to 3 m away from the sensing devices, performed continuous motions such as standing, sitting, walking, and running in scenes. Three people performed in two scenes. The total number of images was 19,982; 80% and 20% were used for training and testing, respectively.

Evaluation metrics We evaluated the performance of the model using a mean intersection-over-union (mIoU). To calculate the mIoU of the output images from the VAE-based methods, including CL-VAE, the sigmoid function was used for the last layer of the decoders, and the decoded images were binarized at a threshold of 0.5.

5.2. Experimental results

To evaluate the performance, we first describe the performance of the proposed CL-VAE and compare it with other methods. Then, we describe the comparison with the segmentation of RGB images. Finally, we explain the performance of CL-VAE against environmental changes.

5.2.1 Performance of CL-VAE

We first evaluate the performance of the proposed CL-VAE. To confirm the validity of learning with both sound and segmentation images, we compared our model with VAE, \( \beta \)-VAE [20], Joint VAE [12], U-Net, Probabilistic U-Net [27], and contrastive unpaired translation (CUT) [42]. VAE, \( \beta \)-VAE, Joint VAE were trained by segmentation images and inferenced by sound images. U-Net and Probabilistic U-Net were trained by the sound images with segmentation label images and inferenced by sound images. CUT and CL-VAE were trained by sound and segmentation images and inferenced by sound images. The image sizes were 64 × 64 in \( \beta \)-VAE and Joint VAE, 128 × 128 in Probabilistic U-Net, 256 × 256 in CUT, and 180 × 120 in VAE, U-Net, and CL-VAE. The mIoUs were calculated with images resized to 180 × 120.

Table 1 and Figure 7 illustrate the quantitative and qualitative evaluations, respectively. The mIoU of our model was higher than those of other models, except for U-Net and Probabilistic U-Net. Although the mIoUs of U-Net and Probabilistic U-Net were higher than CL-VAE, the shapes of people were ambiguous in the estimated images compared with those of CL-VAE as shown in Fig. 7. The images estimated by VAE were noisy and the shapes of people were not clearly estimated. On the other hand, the shapes of people of CL-VAE could be estimated more clearly than that of
VAE. Since the latent variables obtained from the sound and segmentation images by CL-VAE were close, it was possible to estimate using the edge information on the segmentation images even when the sound images were input. In $\beta$-VAE, only a single person segmentation was estimated under conditions of multiple people, and the shapes tended to be different from the ground truth. Although the segmentation of multiple people was estimated by Joint VAE, the shapes were different from the ground truth. Though CUT segmentation images estimated the human shapes more clearly than our model, some people were not estimated especially in multiple people images, and had different shapes from the segmentation images.

### 5.2.2 Gaps with camera-based approaches

Further, we compared our method with camera-based approaches. To confirm the instance segmentation performance from RGB images, the mIoU of segmentation images estimated by Mask R-CNN was evaluated. The ground truth images were manual annotations of 500 images selected from the test images. The manual annotation was performed using LabelMe [53]. The RGB images are required to be more sensitive to improve visibility in dark environments. In that case, the noises contained in the images increase, which affects the estimation accuracy of the segmentation. Besides, ultrasounds are unaffected by the changes in brightness. Therefore, to compare the performance with camera-based approaches, we used Mask R-CNN to estimate segmentation images from RGB images and evaluated the mIoU of these images. The manual annotations were performed using LabelMe [53].
5.2.2 Performance in a dark environment

To evaluate the performance of CL-VAE in a dark environment, we created noisy images that simulate images with high ISO sensitivity. The mIoUs of the model trained by RGB images and tested by noisy images were evaluated. Notably, the noisy images were generated as follows. First, the brightness of the image \( I \) was decreased, and the darkened image \( I_{\text{dark}} \) was generated as

\[
I_{\text{dark}} = \frac{I}{255} - 0.2, \tag{9}
\]

\[
I_{\text{dark}}(i,j) = \begin{cases} 
0, & I_{\text{dark}}(i,j) < 0 \\
I_{\text{dark}}(i,j), & I_{\text{dark}}(i,j) \geq 0.
\end{cases} \tag{10}
\]

Then, noisy image \( I_{\text{noise}} \) was created by adding Gaussian noises \( N \) with the mean 0 and variance 0.5 to the darkened image \( I_{\text{dark}} \).

\[
I_{\text{noise}}(i,j) = I_{\text{dark}}(i,j) + \kappa \sqrt{I_{\text{dark}}(i,j)} N(i,j), \tag{11}
\]

where \( \kappa \) is the coefficient to adjust the peak signal-to-noise ratio (PSNR) of the noisy images. Examples of noisy images are shown in Figure 8. Mask R-CNN model pretrained by the COCO dataset [34] was used for the estimation.

The mIoUs of these conditions are shown in Table 3. The mIoUs decreased under all conditions. In the evaluation of (a), it was assumed that the mIoU decreased because the shape of the mannequin was different from that of the training dataset and the reflected waves differed because of the surface material. In the evaluation of (b), (c), and (d), it was assumed that the mIoUs decreased because the sound images were different from that of the dataset due to the difference in the reflection angle, the frequency, and the distance of the wall. To address the limitations obtained from these results, we will increase the variety of data and improve the sound image generation process.

Table 2. Comparison of mIoU with camera-based approach. Mask R-CNN pretrained by COCO dataset, which had less noise, was used for the evaluation.

| \( \kappa \) | RGB images | Noisy images | CL-VAE |
|---|---|---|---|
| PSNR | - | - | - |
| mIoU | 0.764 | 0.731 | 0.689 | 0.619 | 0.545 | 0.472 | 0.384 | 0.310 | 0.240 | 0.202 | 0.174 | 0.532 |

Table 3. mIoU in various conditions.

| Condition | mIoU |
|---|---|
| (a) Person | 0.357 |
| (b) Sensor position | 0.215 |
| (c) Ultrasound frequency | 0.206 |
| (d) Distance of the wall | 0.489 |

(a) A person not included in the training dataset. We used a mannequin to simulate that condition.
(b) A change of a sensor position. Although we captured the training data in front of people, we captured the data from an angle in this condition.
(c) A change of ultrasound frequency from 62 to 40 kHz. To match this condition, the analysis frequency band was changed in the preprocessing.
(d) A change of the distance between the wall and the sensor from 3 to 1.5 m. To match this condition, the analysis section was limited to the range of 1.5 m when generating the sound images.
6. Conclusions

We proposed privacy-aware human instance segmentation from airborne ultrasonic using CL-VAE. Our approach can produce human instance segmentation images by learning the latent space shared by both segmentation and sound images for training. This approach can be applied to detect human actions in situations where consideration for privacy is required, such as home surveillance, because the sound/segmentation images cannot be reconstructed to RGB images. To improve the accuracy in unknown environments, we will increase the variety of data and improve the sound image generation process in the future.

References

[1] Abnormal behavior recognition for intelligent video surveillance systems: A review. Expert Systems with Applications, 91:480–491, 2018.
[2] Reem Abdel-Salam, Rana Mostafa, and Mayada Hadhood. Human activity recognition using wearable sensors: review, challenges, evaluation benchmark. arXiv preprint arXiv:2101.01665, 2021.
[3] Hande Alemdar, Halil Ertan, Ozlem Durmaz Incel, and Cem Ersoy. ARAS human activity datasets in multiple homes with multiple residents. In 2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops, pages 232–235. IEEE, 2013.
[4] Inigo Alonso and Ana C. Murillo. EV-SegNet: Semantic segmentation for event-based cameras. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1624–1633, June 2019.
[5] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12):2481–2495, 2017.
[6] Alessio Carullo and Marco Parvis. An ultrasonic sensor for distance measurement in automotive applications. IEEE Sensors Journal, 2(1):143–147, 2001.
[7] D.E. Chimenti. Review of air-coupled ultrasonic materials characterization. Ultrasonics, 54(7):1804–1816, 2014.
[8] Maria Cornacchia, Koray Ozcan, Yu Zheng, and Senem Veli-pasalar. A survey on activity detection and classification using wearable sensors. IEEE Sensors Journal, 17(2):386–403, 2016.
[9] Amit Das, Ivan Tashev, and Shoaib Mohammed. Ultrasound based gesture recognition. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 406–410, 2017.
[10] Chandni J Dhansamia and Tushar V Ratanpara. A survey on human action recognition from videos. In 2016 online international conference on green engineering and technologies (IC-GET), pages 1–5. IEEE, 2016.
[11] Tran Hiep Dinh, Minh Trien Pham, Manh Duong Phung, Due Manh Nguyen, Van Manh Hoang, and Quang Vinh Tran. Image segmentation based on histogram of depth and an application in driver distraction detection. In 2014 13th International Conference on Control Automation Robotics Vision (ICARCV), pages 969–974, 2014.
[12] Emilien Dupont. Learning disentangled joint continuous and discrete representations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems (NeurIPS), volume 31. Curran Associates, Inc., 2018.
[13] Yiming Fang, Lujun Lin, Hailin Feng, Zhixiong Lu, and Grant W. Emms. Review of the use of air-coupled ultrasonic technologies for nondestructive testing of wood and wood products. Computers and Electronics in Agriculture, 137:79–87, 2017.
[14] Biying Fu, Jakob Karolus, Tobias Grosse-Puppendahl, Jonathan Hermann, and Arjan Kuiper. Opportunities for activity recognition using ultrasound doppler sensing on unmodified mobile phones. In Proceedings of the 2nd international Workshop on Sensor-based Activity Recognition and Interaction, pages 1–10, 2015.
[15] D. Gowseikhaa, S. Abirami, and R. Baskaran. Automated human behavior analysis from surveillance videos: A survey. Artificial Intelligence Review, 42(4):747–765, 2014.
[16] Monica Gurosso, Nicola Capece, and Ugo Erra. Human segmentation in surveillance video with deep learning. Multimedia Tools and Applications, 80(1):1175–1199, 2021.
[17] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. Tool release: Gathering 802.11n traces with channel state information. SIGCOMM Computer Communication Review, 41(1):53, Jan. 2011.
[18] Le Han, Xuelong Li, and Yongsheng Dong. Convolutional edge constraint-based U-Net for salient object detection. IEEE Access, 7:48980–48990, 2019.
[19] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 2961–2969, Oct 2017.
[20] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In International Conference on Learning Representations (ICLR), 2017.
[21] Gunpil Hwang, Seohyeon Kim, and Hyeon-Min Bae. Batnet: Bat-inspired high-resolution 3D image reconstruction using ultrasonic echoes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alch´e-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems (NeurIPS), volume 32. Curran Associates, Inc., 2019.
[22] Go Irie, Mirela Ostrek, Haochen Wang, Hirokazu Kameoka, Akisato Kimura, Takahito Kawashima, and Kunio Kashino. Seeing through sounds: Predicting visual semantic segmentation results from multichannel audio signals. In International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 3961–3964, 2019.
[23] Yanli Ji, Yang Yang, Fumin Shen, Heng Tao Shen, and Xuelong Li. A survey of human action analysis in hri applica-
[50] Prafull Sharma, Miika Aittala, Yoav Y. Schechner, Antonio Torralba, Gregory W. Wornell, William T. Freeman, and Fredo Durand. What you can learn by staring at a blank wall. arXiv preprint arXiv:2108.13027, 2021.

[51] Jae Mun Sim, Yonnim Lee, and Ohbyung Kwon. Acoustic sensor based recognition of human activity in everyday life for smart home services. International Journal of Distributed Sensor Networks, 11(9), 2015.

[52] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Semantic object prediction and spatial sound super-resolution with binaural sounds. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, Computer Vision – ECCV 2020, pages 638–655, Cham, 2020. Springer International Publishing.

[53] Kentaro Wada. labelme: Image Polygonal Annotation with Python. https://github.com/wkentaro/labelme, 2016.

[54] Fei Wang, Sanping Zhou, Stanislav Panev, Jinsong Han, and Dong Huang. Person-in-WiFi: Fine-grained person perception using WiFi. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 5451–5460, 2019.

[55] Huaijun Wang, Jing Zhao, Junhuai Li, Ling Tian, Pengjia Tu, Ting Cao, Yang An, Kan Wang, and Shancang Li. Wearable sensor-based human activity recognition using hybrid deep learning techniques. Security and Communication Networks, 2020, 2020.

[56] Lei Wang, Du Q Huynh, and Piotr Koniusz. A comparative review of recent kinect-based action recognition algorithms. IEEE Transactions on Image Processing, 29:15–28, 2019.

[57] W. Wang, Y. Song, J. Zhang, and H. Deng. Automatic parking of vehicles: A review of literatures. International Journal of Automotive Technology, 15(6):967–978, 2014.

[58] Yancheng Wang, Yang Xiao, Fu Xiong, Wenxiang Jiang, Zhiguo Cao, Joey Tianyi Zhou, and Junsong Yuan. 3DV: 3D dynamic voxel for action recognition in depth video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 511–520, 2020.

[59] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[60] Yi Zhou, Guillermo Gallego, Xiuyuan Lu, Siqi Liu, and Shaojie Shen. Event-based motion segmentation with spatio-temporal graph cuts. arXiv preprint arXiv:2012.08730, abs/2012.08730, 2020.

[61] Muhammad Zubair, Kibong Song, and Changwooo Yoon. Human activity recognition using wearable accelerometer sensors. In 2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), pages 1–5. IEEE, 2016.