Optical Review (2024) 31:116–125
https://doi.org/10.1007/s10043-023-00848-2

SPECIAL SECTION: REGULAR PAPER

Laser Display and Lighting Conference (LDC' 23), Yokohama, Japan

Gesture recognition using deep-learning in single-pixel-imaging with high-frame-rate display with latent random dot patterns

Hiroki Takatsuka[1] · Masaki Yasugi[2] · Shiro Suyama[1] · Hirotsugu Yamamoto[1]

Received: 31 May 2023 / Accepted: 1 November 2023 / Published online: 11 December 2023

Abstract

Gesture recognition using cameras capable of capturing detailed images for gesture recognition is not feasible in many places due to concerns regarding privacy and information leakage. To address this problem, we have proposed a method of capturing shadow pictures using single-pixel-imaging to realize privacy-conscious gesture recognition. As an implementation method of single-pixel-imaging in public spaces, we have studied using a high-frame-rate LED display as a light source. By using a high-frame-rate LED display, random patterns can be latent while the observer perceives an apparent image. However, the image reconstructed by single-pixel-imaging using a high-frame-rate LED display is influenced by the apparent image, making gesture recognition difficult. In this study, we show that the influence of the apparent image can be removed by restoring the restored image using deep learning with a convolutional network called U-Net, and high classification accuracy with a small number of illuminations by using LeNet to classify restored images.

Keywords Gesture recognition · Single-pixel-imaging · Deep learning · U-Net · LeNet

1 Introduction

With the recent advancements in information and communication technology, information displays have become pervasive in our daily lives. By combining these information displays with gesture recognition technology, it becomes possible to create interactive information interfaces that can switch images on the display based on the user's gestures. Examples of gesture recognition applications include patient monitoring, anomaly detection using surveillance cameras, master–slave operations for robots, and sign language recognition [1]. To perform gesture recognition, various devices are used, such as stereo high-speed cameras [2], stereo infrared cameras (Leap Motion) [3], and Time of Flight (ToF) 3D cameras (Kinect) [4]. However, using cameras capable of capturing detailed images for gesture recognition is not feasible in many places due to concerns regarding privacy and information leakage. Examples of such places include personal spaces like toilets and bathrooms, as well as public spaces. Particularly in bathrooms, it is not possible to use electrostatic sensors, and voice recognition is difficult due to water sounds. To address this problem, research has been conducted on methods such as reducing the resolution of captured images [5] and performing masking operations outside the required areas [6]. We have proposed a method of capturing shadow pictures using single-pixel-imaging to realize privacy-conscious gesture recognition [7].

Single-pixel-imaging is a technique that utilizes spatially modulated illumination and a single light detector to capture images [8]. It allows imaging under low-light conditions and with light sources other than visible light, making it applicable in a wide range of scenes. To perform single-pixel-imaging, a modulable light source is required, and various displays already present in public spaces can serve as suitable light sources. We have previously proposed single-pixel-imaging using a high-speed modulable LED display for banner advertisements and news display [9]. In this case, the content of the banner display can be directly utilized as the spatial light intensity distribution of the light source [10]. Alternatively, by embedding random patterns while maintaining the apparent image recognizable to observers [11], it becomes possible to achieve a balance between digital

✉ Hirotsugu Yamamoto
hirotsugu@yamamotolab.science

1 Utsunomiya University, Utsunomiya, Tochigi, Japan

2 Fukui Prefectural University, Obama, Fukui, Japan

signage display and imaging without constraints on the content. However, this approach presents a challenge where the reconstructed images through single-pixel-imaging are influenced by the apparent image, making gesture recognition difficult [12].

To solve this problem, we propose to use deep learning to restore the original image from which the apparent image has been removed from the reconstructed image of single-pixel-imaging. Although deep learning has been proposed to reduce the number of illumination times for single-pixel-imaging [7], this study aims to achieve both reduction of illumination times and removal of apparent images. Preliminary results of this study were presented at LDC2023 [13]. The purpose of this paper is to investigate the classification accuracy of reconstructed single-pixel-imaging images with latent random patterns in the illumination by removing the influence of apparent images through deep learning. To achieve this, a neural network, U-Net, is used to train pairs of reconstructed and original images, and the image is restored by the network. LeNet was then used to determine the classification accuracy of the restored image.

2 Principle

2.1 Single-pixel-imaging with random-dot-embedded apparent images

The principle of the single-pixel-imaging with random-dot-embedded apparent images is shown in Fig. 1. The encoded images are displayed on an LED display at a sufficiently high frame rate, so the observer perceives an apparent image that integrates the encoded images. The light transmitted through the subject is measured by a single detector and reconstructed using the principle of single-pixel-imaging with 2D encoding images and 1D temporal signals. The reconstruction of single-pixel-imaging is expressed by

$$\begin{aligned} G(x,y,n) &= \langle \Delta I(x,y,n)\Delta A(n)\rangle \\ &= \left\langle \left[I(x,y,n) - \langle I(x,y,n)\rangle\right]\left[A(n) - \langle \Delta A(n)\rangle\right]\right\rangle \\ &= \langle I(x,y,n)A(n)\rangle - \langle I(x,y,n)\rangle\langle A(n)\rangle \end{aligned} \tag{1}$$

where $\Delta I(x,y,n)$ is the deviation between the light intensity $I(x,y,n)$ and the mean $\langle I(x,y,n)\rangle$ of the n-th 2D encoding images in the coordinates (x,y). $\Delta A(n)$ is the deviation of average value of 1D temporal signals. $A(n)$ can also be given by

$$A(n) = \iint T(x,y)I(x,y,n)\,\mathrm{d}x\mathrm{d}y \tag{2}$$

where $T(x,y)$ denotes the transmission function [14]. Thus, the reconstructed image from n-th measurements

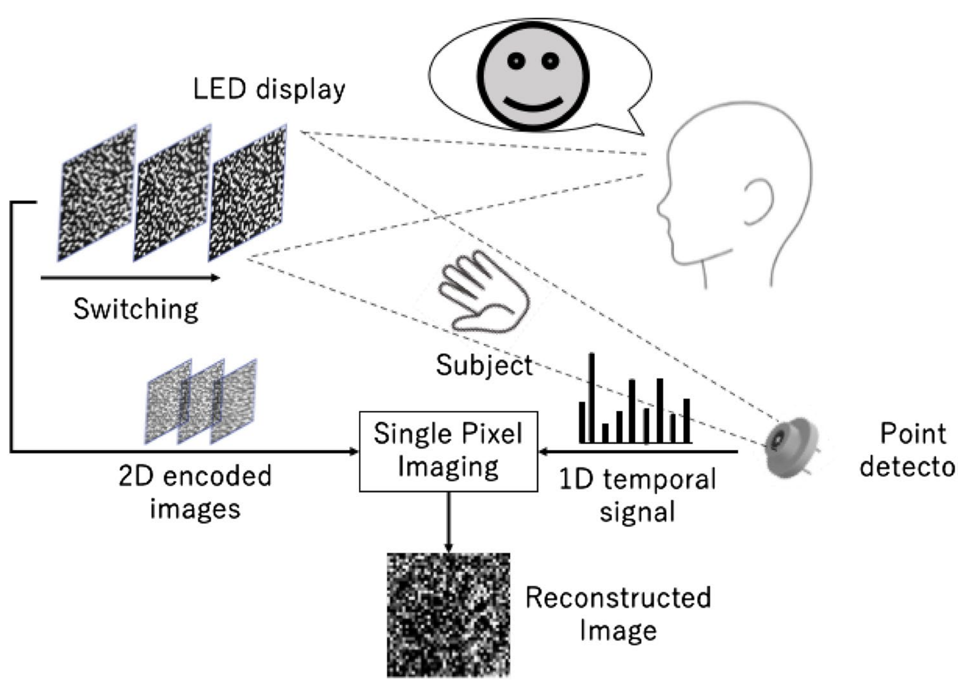


Fig. 1 Principle of single-pixel-imaging with apparent image latent with random pattern

can be obtained from 2D encoded images displayed on the LED display and the 1D temporal signal measured by a single detector. The reconstructed images are influenced by noise and apparent images, making gesture recognition difficult.

2.2 Encoding of apparent images

The LED display is updated at a sufficiently high frame rate so that the observer perceives an integrated image of latent random patterns. This principle has been confirmed with LED displays at 960 fps [11]. Encode m frames to latent random patterns in the apparent image. The latent random pattern satisfies:

$$V(x,y) \equiv \sum_{n=1}^{m} E(x,y,n) \tag{3}$$

where $V(x,y)$ be the pixel value of the apparent image at coordinate (x,y) and $E(x,y,n)$ be the pixel value of the n-th coded image [15].

Fig. 2 Apparent image

Table 1 Composition of pixel values by two encoded images

Original pixel value	Number of pixel value 0	Number of pixel value 190	Number of pixel value 255	Total encoded images
190	1	1	0	2
255	1	0	1	2

In this study, the apparent image was also encoded to satisfy Eq. (3). The apparent image used in the experiment was a binary image with pixel values (190,255) as shown in Fig. 2. When $m = 2$ is used as an example of encoding, Fig. 3 shows two coded images of Fig. 2. Table 1 shows the composition of pixel values by encoding two images. By displaying these two images at high speed on an LED display, the observer perceives the apparent image shown in Fig. 2.

2.3 U-Net

Structure of U-Net is shown in Fig. 4. U-Net is a convolutional neural network (CNN) that is good at capturing and restoring features of input images [16]. In the convolutional process, a filter-based convolution is performed on the input to output a feature map. Maxpooling reduces the resolution of the input by extracting the maximum value in the filter and aggregating it into one. Then, unpooling brings the resolution back to the original. These processes enable capturing the features of an object. However, since the positional information of the object is lost in these processes, the feature maps before the convolution is concatenated to complement the positional information, which is called skip-connection.

U-Net was developed for medical image segmentation and was also used in this study because it is suitable for single-pixel-imaging that contains a lot of noise.

Fig. 3 Two encoded images

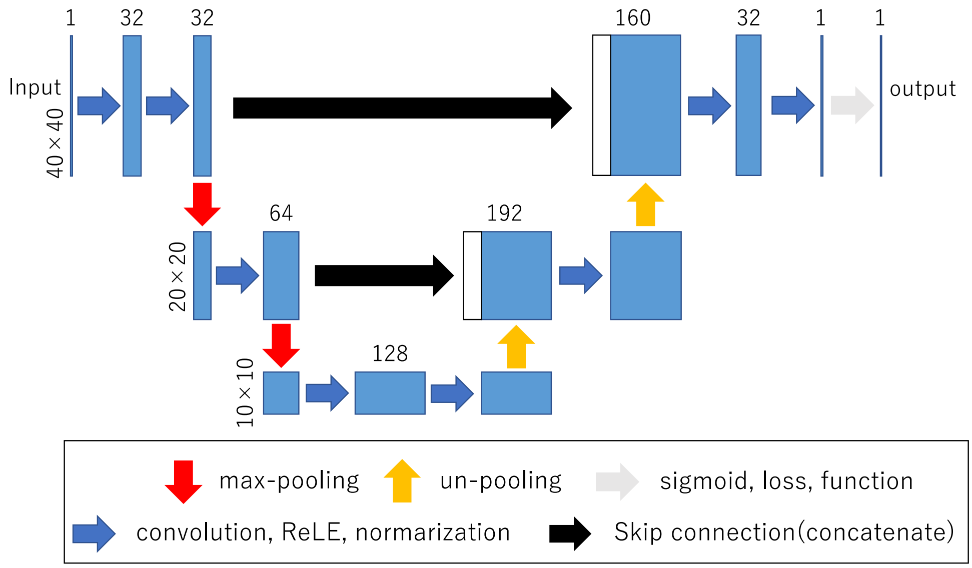


Fig. 4 Structure of U-Net

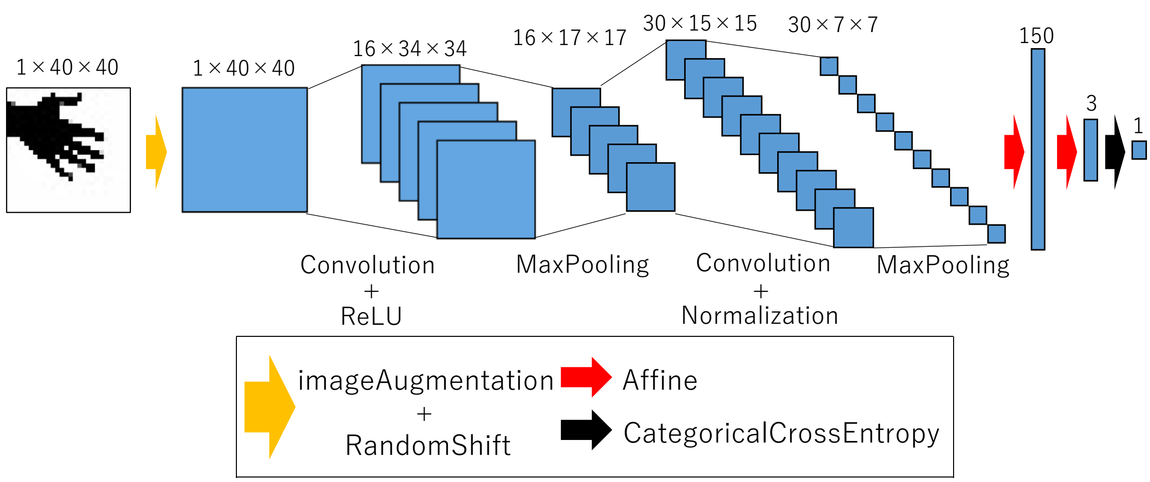


Fig. 5 Structure of LeNet

Table 2 Composition of pixel values by 20 encoded images

Original pixel value	Number of pixel value 0	Number of pixel value 19	Number of pixel value 65	Total encoded images
190	10	10	0	20
255	9	10	1	20

2.4 LeNet

Structure of LeNet is shown in Fig. 5. LeNet is a network model suitable for image classification that consist of CNN [17]. This network performs classification by repeating the convolutional layer and the max-pooling layer, and then

Fig. 6 A part of 20 encoded images

Table 3 Composition of gesture images

Rock	Scissors	Paper	Total
6000	6000	6000	18,000

repeating the affine layer. In this paper, we added layers for image augmentation to compensate for the lack of training data.

3 Experiments

3.1 Reconstruction of single-pixel-imaging

Hand gesture images were reconstructed using single-pixel-imaging with random patterns and single-pixel-imaging with apparent images. The SSIM value is a measure of structural similarity, and the closer the value is to 1, the higher the similarity. The apparent images were encoded into 20 images, and the pixel value composition of the 20 encoded images is shown in Table 2 and a part of 20 encoded images are shown in Fig. 6. The order in which these are displayed is random for each pixel. Hand gesture images are 18,000 images of 40×40 pixels and are simulated on a computer. Composition of the hand gesture images is shown in Table 3 and hand gesture images are shown in Fig. 7.

3.2 Elimination of apparent image with U-Net

To remove the influence of the apparent image, the reconstructed image was restored by learning with U-Net. By using pairs of original gesture images and reconstructed images from single-pixel-imaging, U-Nets were trained to remove the influence of apparent image. To obtain the transition of the SSIM value in response to changes in the number of illuminations in the reconstructed image, Network settings were the same, and training was performed for each number of illuminations. Training was performed using the Neural Network Console (NNC) provided by Sony. Dataset structure of U-Net is shown in Table 4, Network settings of U-Net is shown in Table 5, and U-Net implemented on NNC is Fig. 8.

3.3 Classification of hand gesture

We performed learning to classify the restored images using LeNet. Restored images were given labels corresponding to gestures. Training was performed using the labeled restored images. Network settings were the same and training was performed for each number of illuminations. Training was performed using NNC. Dataset structure of LeNet is shown in Table 6, Network settings of LeNet is shown in Table 7, and LeNet implemented on NNC is Fig. 9. In "ImageAugmentation" layer, input images are rotated, and in "RandomShift" layer, patterns are increased by shifting left and right.

Table 4 Dataset structure of U-Net

	Training data set	Validation data set	Test data set
Rock	4000	1200	800
Scissors	4000	1200	800
Paper	4000	1200	800
Total	12,000	3600	2400

Table 5 Network setting of U-Net

Updater	Adam
Update interval	1
Weight decay	0
α (Learning rate)	0.001
Beta1, beta2	0.9, 0.999
Batch size	64
Epoch	100

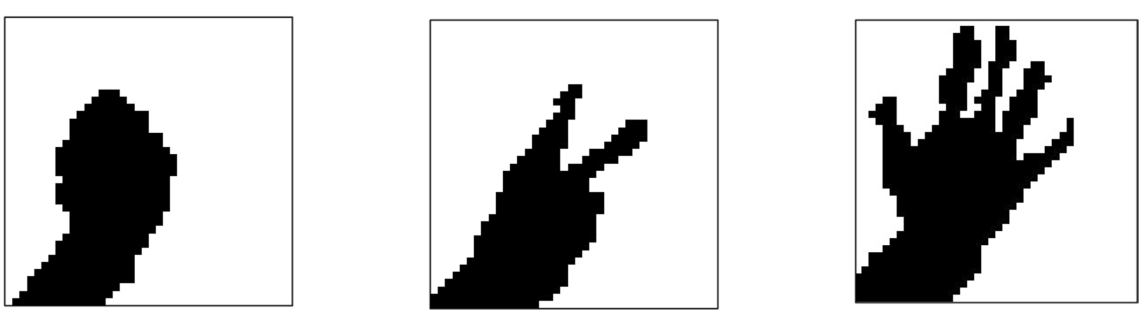


Fig. 7 Images of hand gesture

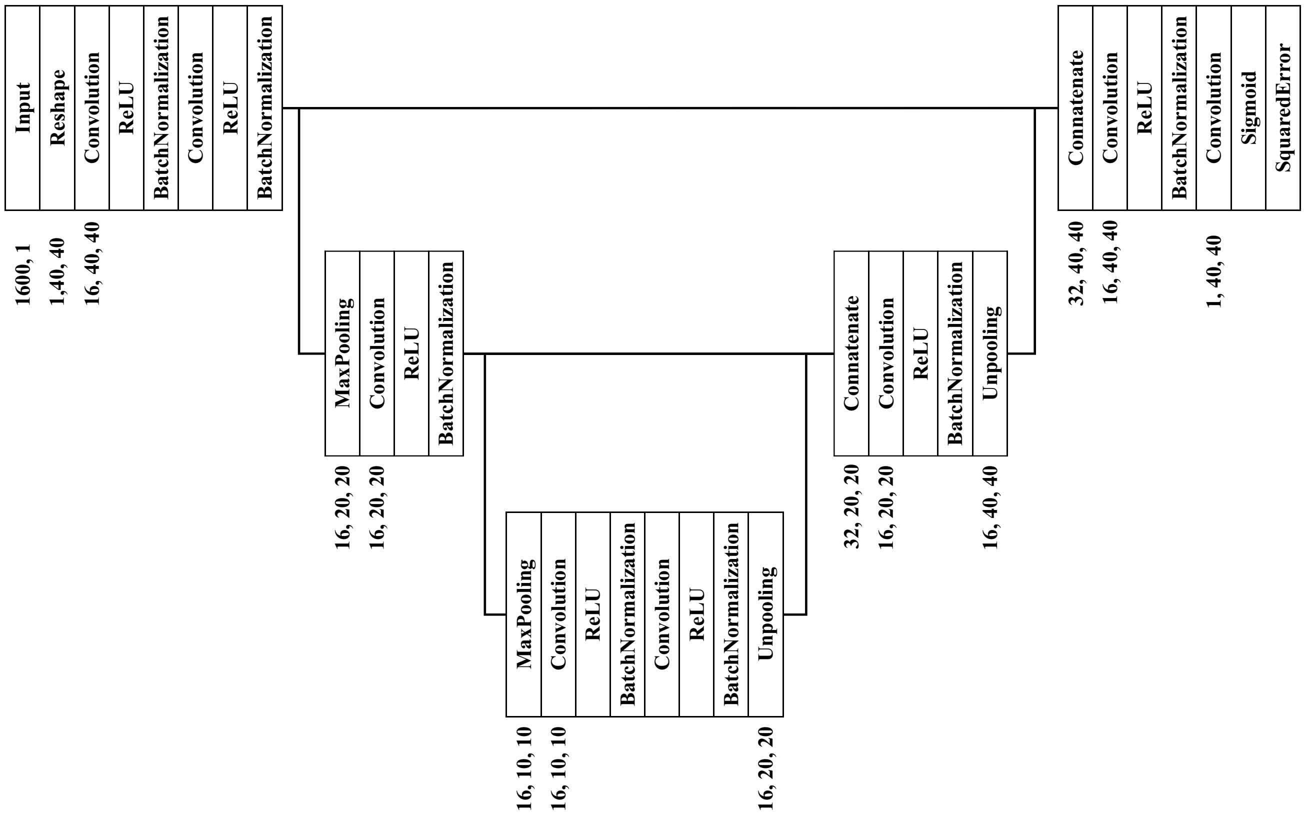


Fig. 8 U-Net implemented on NNC

Table 6 Dataset structure of LeNet

	Training data set	Validation data set	Test data set
Rock	500	200	100
Scissors	500	200	100
Paper	500	200	100
Total	1500	600	300

Table 7 Network setting of LeNet

Updater	AMSGrad
Update interval	1
Weight decay	0
α (Learning rate)	0.001
Beta1, beta2	0.9, 0.999
Batch size	64
Epoch	100

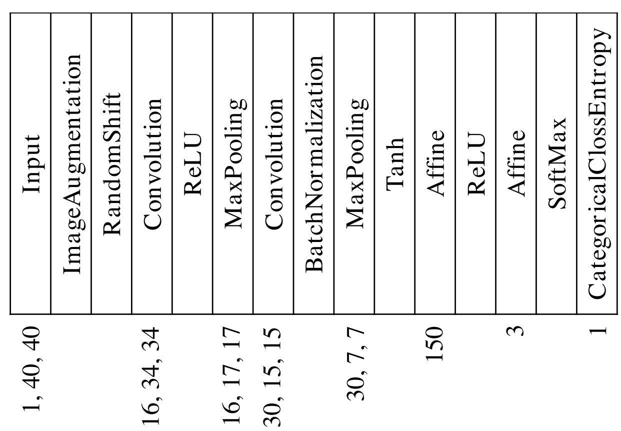


Fig. 9 LeNet implemented on NNC

4 Result

4.1 Reconstruction of single-pixel-imaging

Figure 10 shows reconstruction results of single-pixel-imaging with random patterns and single-pixel-imaging with apparent images, and Fig. 11 shows the SSIM values of single-pixel-imaging with random patterns and single-pixel-imaging with apparent images.

Figure 10 shows that the reconstruction results of the random pattern and the apparent image are clearer when there is more illuminations, and noisier when there is less

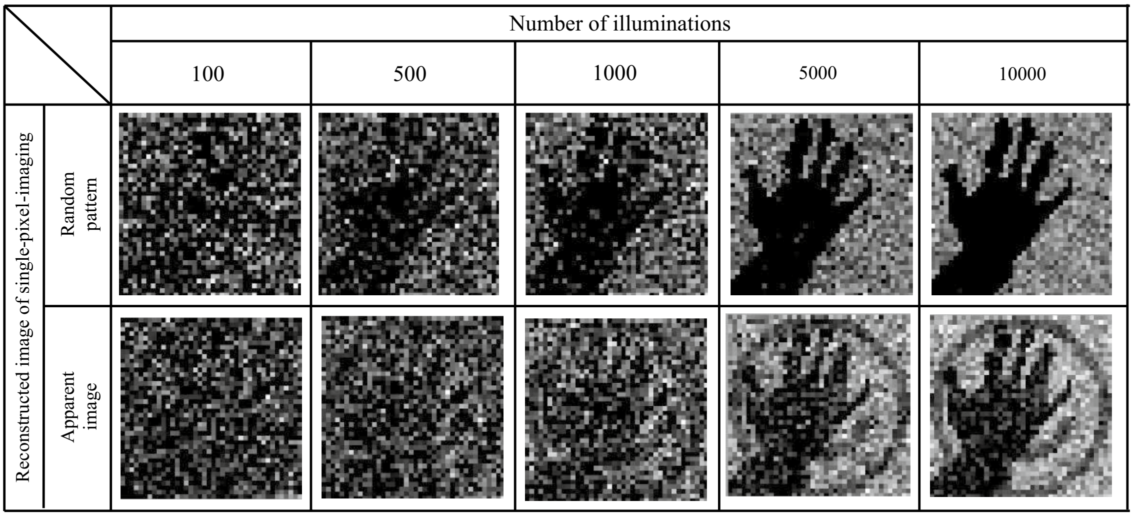


Fig. 10 Reconstruction results of single-pixel imaging with random patterns and single-pixel-imaging with apparent images

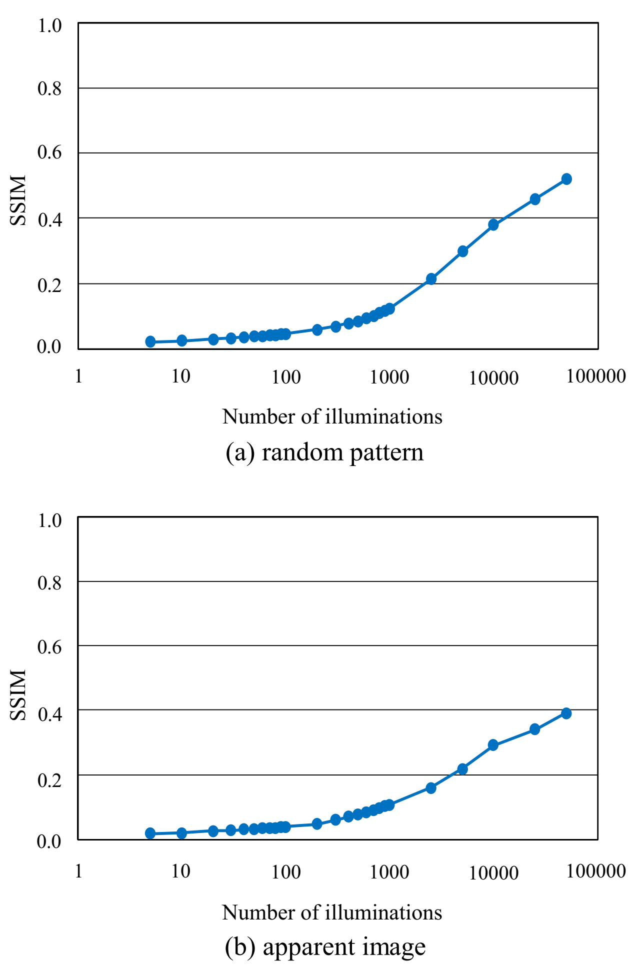


Fig. 11 SSIM value for reconstructed image of **a** random pattern and **b** apparent image

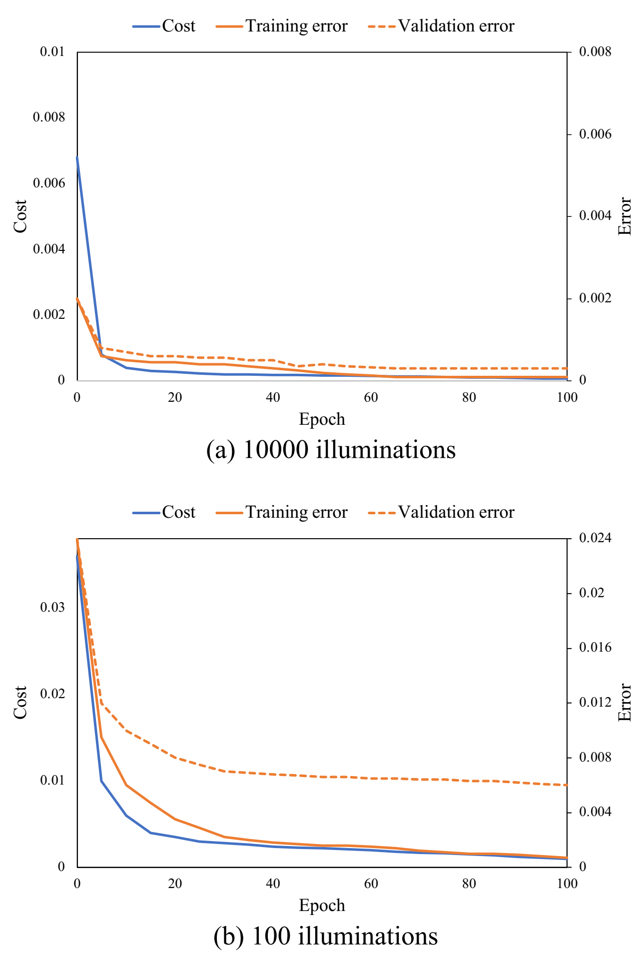


Fig. 12 Learning curves of U-Net for 10,000 illuminations and 100 illuminations in single-pixel imaging using apparent images

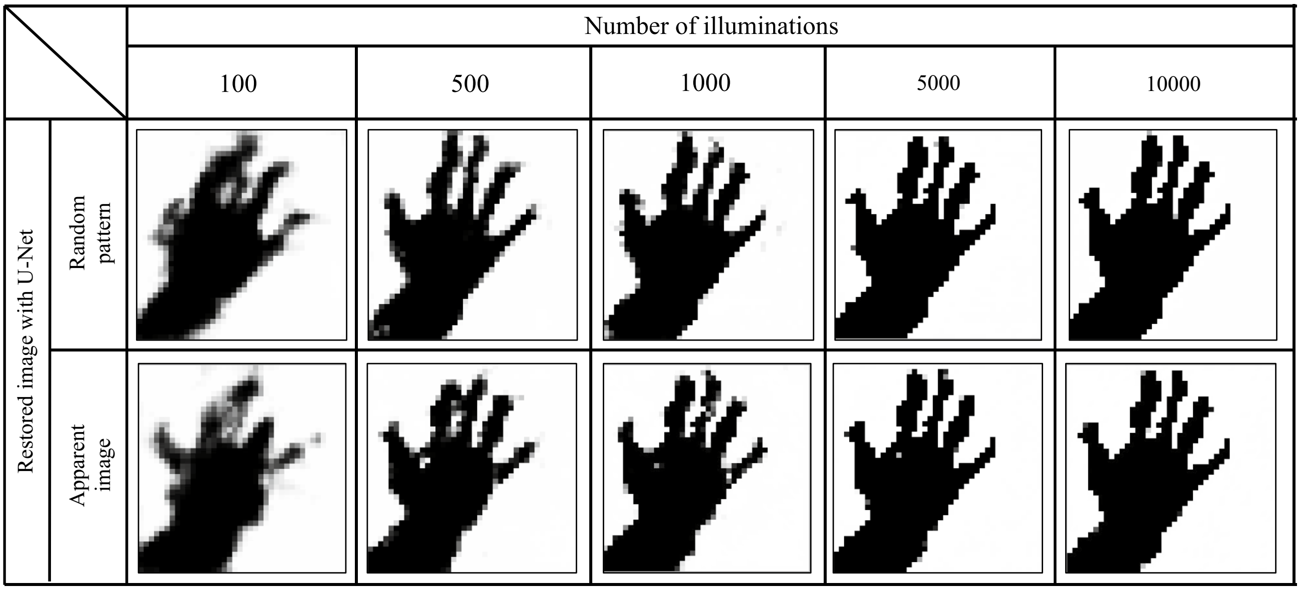


Fig. 13 Restoration result of **a** random pattern and **b** apparent image

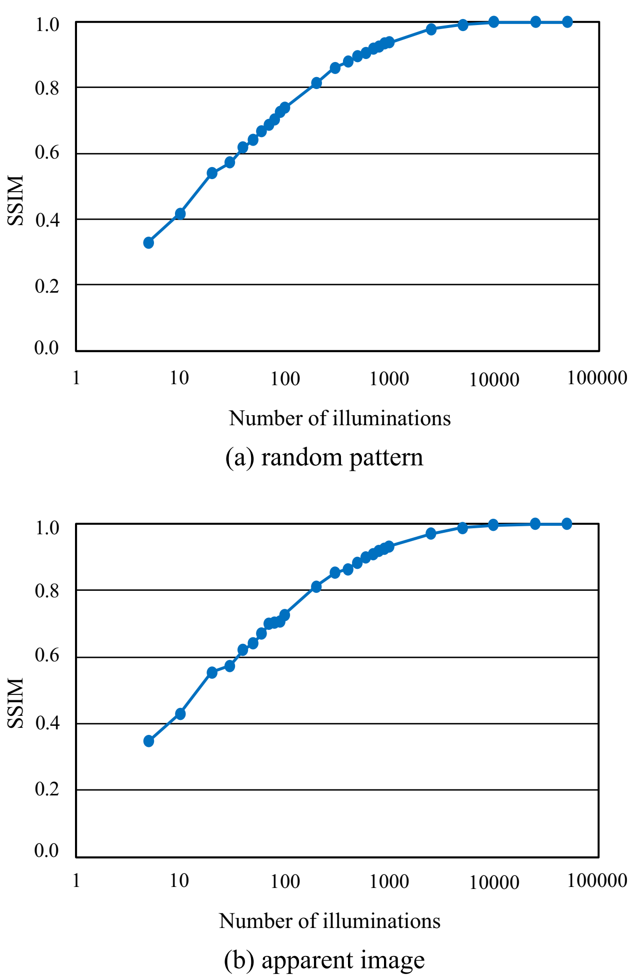


Fig. 14 SSIM value for restored image of **a** random pattern and **b** apparent image

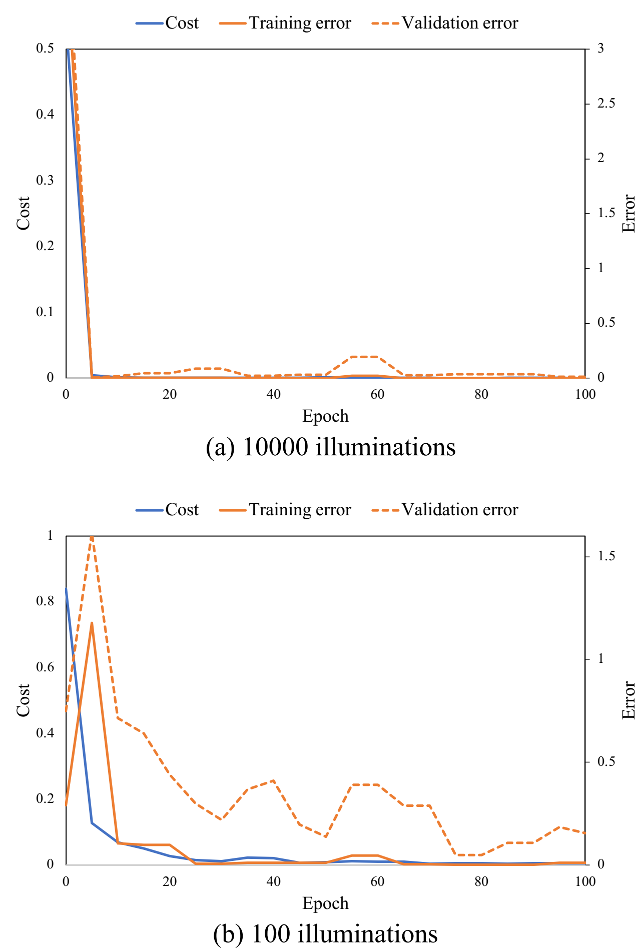
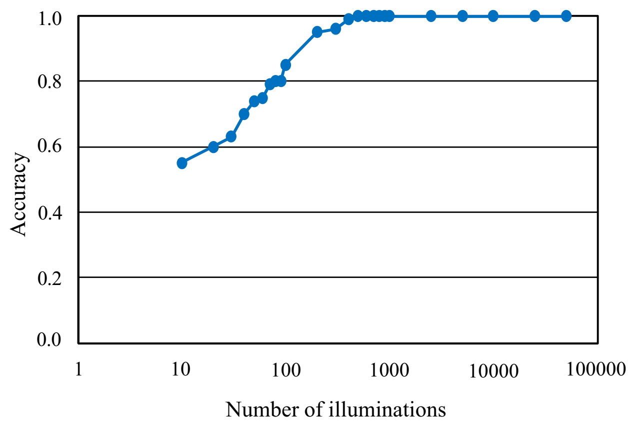
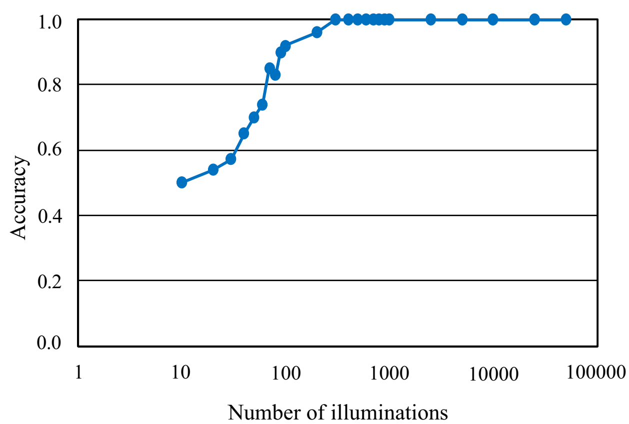


Fig. 15 Learning curves of LeNet for **a** 10,000 illuminations and **b** 100 illuminations in single-pixel imaging using apparent images



(a) random pattern



(b) apparent image

Fig. 16 The relationship between the number of illuminations and classification accuracy of **a** random pattern and **b** apparent image

illuminations. The single-pixel-imaging using the apparent image shows the influence of the apparent image.

Figure 11 shows that when the number of illuminations is 1000 or less, the SSIM values of the random pattern and the apparent image are comparable. When the number of illuminations exceeds 1000, the random pattern has a higher SSIM value.

4.2 Elimination of apparent image with U-Net

U-Net was trained to restore the reconstructed image. Learning curves of U-Net for 10,000 illuminations and 100 illuminations in single-pixel-imaging using apparent images are shown in Fig. 12, and restoration result of random pattern and apparent image are shown in Fig. 13. SSIM values of the restored image using single-pixel-imaging with random patterns and the restored image of single-pixel-imaging with apparent images are shown in Fig. 14.

Figure 12 shows that the error value converges to a small value when the number of illuminations is set to 10,000. As the number of illuminations decreases, the error value gradually increases, and the error value for 100 illuminations is about ten times larger than that for 10,000 illuminations.

Figure 13 shows that the effect of the apparent image was removed by the U-Net restored image. In addition, it was confirmed that the gestures in the reconstructed image could be restored when the number of illuminations was 500 or more, but the reconstructed image could not be completely restored when the number of illuminations was 100.

Figure 14 shows that there is no difference in SSIM values between the restored image of single-pixel-imaging with random patterns and the restored image of single-pixel-imaging with apparent images.

4.3 Classification of hand gesture

LeNet was trained to classify the restored image. Learning curves of LeNet for 10,000 illuminations and 100 illuminations in single-pixel-imaging using apparent images are shown in Fig. 15, and the relationship between the number of illuminations and classification accuracy of random pattern and apparent image are shown in Fig. 16.

Figure 15 shows that the error value converges to a small value when the number of illuminations is set to 10,000. As the number of illuminations decreases, the error value gradually increases, and the error value for 100 illuminations not only decrease when the number of epochs increases, but also increase in some places.

Figure 16 show that classification accuracy depends on the number of illuminations. When the number of illuminations was 300 or more, all restored images could be classified, and when the number of illuminations was less than 200, the classification accuracy began to decrease. The classification accuracy was similar for both random patterns and apparent images.

5 Discussion

Figures 12 and 15 show that there is a large difference in error values when comparing the error values resulting from 10,000 illuminations and 100 illuminations, and there are apparent signs of over-learning in the case of 100 illuminations. To solve this problem, it is considered necessary to improve the network and adjust parameters.

From Fig. 11, the difference in SSIM values between the reconstructed image of single-pixel-imaging with random patterns and the reconstructed image of single-pixel-imaging with apparent images can be seen. However, from Fig. 14, the SSIM values of the reconstructed image of single-pixel-imaging with random patterns and the reconstructed image of single-pixel-imaging with apparent images are similar.

Also, from Fig. 16, the classification accuracy of the restored image of single-pixel-imaging using random patterns by LeNet and that of the restored image of single-pixel-imaging using apparent images by LeNet are similar. Therefore, using U-Net and LeNet in single-pixel-imaging with apparent images, it is possible to classify more than 80% of the restored images with more than 200 illuminations. We expect that the measurement with 200 illuminations and a 3000 Hz LED display can realize gesture classification with a sampling rate of 15 fps.

6 Conclusion

Reconstructed images by single-pixel-imaging using apparent images are influenced by the apparent images, and it is difficult to classify gestures. Using U-Net for restoration and LeNet for classification, it is possible to classify all of them with more than 200 illuminations.

Author contributions HT contributed for this paper as first author. He conducted the experiments, analyzed the data, and wrote the original draft. MY and SS and HY designed the experiments and edited the manuscript.

Funding A part of this work was supported by JSPS KAKENHI (20H05702).

Data availability The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare no conflicts of interest associated with this manuscript.

References

1. Mitra, S., Acharya, T.: Gesture recognition: a survey. IEEE Trans Syst Man Cybern Part C **37**(3), 311–324 (2007)
2. Yasui YM, Alvissalim MS, Takahashi M, Tomiyama Y, Suyama S, Ishikawa M. Floating display screen formed by AIRR (Aerial imaging by retro-reflection) for interaction in 3D space. In: 2014 International Conference on 3D Imaging (IC3D) (IEEE, 2014), pp. 1–5.
3. Rossol, N., Cheng, I., Basu, A.: A Multisensor technique for gesture recognition through intelligent skeletal pose analysis. IEEE Trans Hum Mach Syst **46**, 350–359 (2016)
4. Nishihori, M., Izumi, T., Nagano, Y., Sato, M., Tsukada, T., Kropp, A.E., Wakabayashi, T.: Development and clinical evaluation of a contactless operating interface for three-dimensional image-guided navigation for endovascular neurosurgery. Int J Comput Assist Radiol Surg **16**, 663–671 (2021)
5. Dai J, Wu J, Saghafi B, Konrad J, Ishwar P. Towards privacy-preserving activity recognition using extremely low temporal and spatial resolution cameras. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (IEEE, 2015), pp. 68–76.
6. Wu Z, Wang Z, Wang Z, Jin H. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In: Proceedings of the European Conference on Computer Vision (ECCV) (Springer, 2018), pp. 606–624.
7. Mukojima, N., Yasugi, M., Mizutani, Y., Yasui, T., Yamamoto, H.: Deep-learning-assisted single-pixel imaging for gesture recognition in consideration of privacy. IEICE Trans Electron E105-C. **2**, 79–85 (2022)
8. Gibson, G.M., Johnson, S.D., Padgett, M.J.: Single-pixel imaging 12 years on: a review. Opt Express **28**, 28190–28208 (2020)
9. Onose, S., Takahashi, M., Mizutani, Y., Yasui, T., Yamamoto, H.: Single pixel imaging with a high-frame-rate LED digital signage. Proc Int Display Worksh **23**, 1495–1498 (2016)
10. Mukojima, N., Talatsuka, H., Yasugi, M., Suyama, S., Yamamoto, H.: Reconstruction of gesture images by using banner as illumination of single-pixel imaging. Proc. IDW **29**, 1039–1042 (2022)
11. Takahashi M, Yamamoto H. Encryption by spatiotemporal scrambling on a high-frame-rate display. In: The 63rd JSAP Spring Meeting, 21a-S224–5. 2016. [in Japanese].
12. Mukojima N, Yasugi M, Suyama S, Yamamoto H. The possibility of using banner images as the mask pattern of single-pixel imaging. In: 2022 Information Photonics (IP) (OSJ, 2022) IPp-09.
13. Takatsuka H, Yasugi M, Suyama S, Yamamoto H. Reconstruction performance of U-Net in single-pixel-imaging with random-dot-embedded apparent images. In: The 12th laser display and lighting conference 2023, p. LDC7–05. 2023.
14. Shibuya, K., Minamikawa, T., Mizutani, Y., Yamamoto, H., Minoshima, K., Yasui, T., Iwata, T.: Scan-less hyperspectral dual-comb single-pixel-imaging in both amplitude and phase. Opt Express **25**, 21947–21957 (2017)
15. Takatsuka H, Yasugi M, Mukojima N, Suyama S, Yamamoto H. Elimination of apparent image on single-pixel-imaging by use of high-frame-rate display with latent random dot patterns. In: Proc. IDW 29, 1035–1038. 2022.
16. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. 2015. arXiv:1505.04597.
17. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc IEEE **86**(11), 2278–2324 (1998)