

# Deep-Learning-Assisted Single-Pixel Imaging for Gesture Recognition in Consideration of Privacy

Naoya MUKOJIMA<sup>†\*</sup>, Masaki YASUGI<sup>†\*</sup>, Yasuhiro MIZUTANI<sup>††</sup>, Takeshi YASUI<sup>†††</sup>,  
and Hirotsugu YAMAMOTO<sup>†a)</sup>, *Nonmembers*

**SUMMARY** We have utilized single-pixel imaging and deep-learning to solve the privacy-preserving problem in gesture recognition for interactive display. Silhouette images of hand gestures were acquired by use of a display panel as an illumination. Reconstructions of gesture images have been performed by numerical experiments on single-pixel imaging by changing the number of illumination mask patterns. For the training and the image restoration with deep learning, we prepared reconstructed data with 250 and 500 illuminations as datasets. For each of the 250 and 500 illuminations, we prepared 9000 datasets in which original images and reconstructed data were paired. Of these data, 8500 data were used for training a neural network (6800 data for training and 1700 data for validation), and 500 data were used to evaluate the accuracy of image restoration. Our neural network, based on U-net, was able to restore images close to the original images even from reconstructed data with greatly reduced number of illuminations, which is 1/40 of the single-pixel imaging without deep learning. Compared restoration accuracy between cases using shadowgraph (black on white background) and negative-positive reversed images (white on black background) as silhouette image, the accuracy of the restored image was lower for negative-positive-reversed images when the number of illuminations was small. Moreover, we found that the restoration accuracy decreased in the order of rock, scissor, and paper. Shadowgraph is suitable for gesture silhouette, and it is necessary to prepare training data and construct neural networks, to avoid the restoration accuracy between gestures when further reducing the number of illuminations.

**key words:** *single-pixel imaging, deep-learning, gesture recognition, U-net, privacy-preserving*

## 1. Introduction

As the recent advances of information communication technologies, information displays have become pervasive in our daily lives. These information displays can be used as interactive and rich information interfaces by switching the images on the display according to the user's gestures [1]. Examples of gesture detection include the use of high-speed cameras [2], hand-gesture sensor (Leap motion) [3], and a 3D camera (Kinect) [4]. It is essential for these methods to detect the user's gestures.

However, there are many places where gesture detection using a regular camera cannot be implemented due to

privacy issues and the risk of information leakage. This problem is not limited to personal spaces such as washrooms and homes, but also occurs when implementing the system in public spaces and workplaces, where the system visually captures the user's gestures and simultaneously personal (user's face and appearance) and highly confidential information. To solve this problem, several methods have been proposed, such as reducing the resolution of captured images [5], or applying masks to areas other than those necessary for gesture detection [6].

In order to solve this privacy-preserving problem, we propose a method to acquire gestures in consideration of privacy and confidentiality, by utilizing single-pixel imaging to detect only the silhouette of gestures. Single-pixel imaging is a method which can reconstruct images by using a point photodetector and variable illumination patterns [7]–[9]. The object images are reconstructed with correlation calculation between illumination intensity with variable mask pattern and light intensity measured by the photodetector [10]. Single-pixel imaging does not require a camera module. Thus, a privacy-preserving imaging method by use of single-pixel imaging can be used in spaces where the cameras cannot be deployed as described above. The issue of this technology is that it requires a very large number of illuminations to capture the image, and a very large computational cost to achieve high resolution. Regarding gesture recognition, it is enough to keep the resolution low. To reduce the number of illuminations, illumination methods [11]–[13], calculation methods [14], [15] and their combination [16], [17] have been proposed. Then, the use of deep learning for restoration of the original image from data reconstructed with a small number of illuminations has been largely explored in the last five years. Typical approaches are the construction of neural network to reduce the number of illuminations and improve the restoration accuracy (fully-connected neural network [18], U-net [19], convolutional neural network [20], recurrent neural network [21], [22]), and acceleration of image acquisition which aims at real-time performance by reducing the number of illuminations [23]–[25]. Most of these previous studies have shown versatile restoration performance using digits or natural images on databases such as MNIST [26] or ImageNet [27].

The purpose of this study is to propose a single-pixel imaging approach to solve the privacy problem in gesture recognition for the interactive acquisition of information, and to investigate the possibility of drastically reducing the

Manuscript received March 9, 2021.

Manuscript revised June 7, 2021.

Manuscript publicized August 17, 2021.

<sup>†</sup>The authors are with Utsunomiya Univ., Utsunomiya-shi, 321–8585 Japan.

<sup>††</sup>The author is with Osaka Univ., Suita-shi, 565–0871 Japan.

<sup>†††</sup>The author is with Tokushima Univ., Tokushima-shi, 770–8506 Japan.

\*Equally contributed to this paper.

a) E-mail: hirotsugu@yamamotolab.science

DOI: 10.1587/transele.2021DII0002

number of illuminations. Furthermore, we investigate the accuracy of the restored image when deep learning is applied to acquire the image of gesture silhouette. For this purpose, we have acquired silhouette images of the gestures using a simple method with an information display as the light source. Next, we have constructed a network that can accurately restore silhouette images from reconstructed data of single-pixel imaging with reduced number of illuminations. These experiments have been performed on numerical calculation by use of our acquired silhouette images. Finally, we have compared the restoration accuracy among gesture types. To construct the network, Sony Neural Network Console [28] was used, which can be operated graphically and is easy to install even for beginners. At IDW'20, we presented how to make gesture images, the construction of the neural network, and the evaluation of the overall image restoration accuracy using a single index [29]. In this paper, we have detailed our principle and experiments, and analyzed the differences in restoration accuracy among gestures using multiple indicators.

## 2. Principle

### 2.1 Single-Pixel Imaging

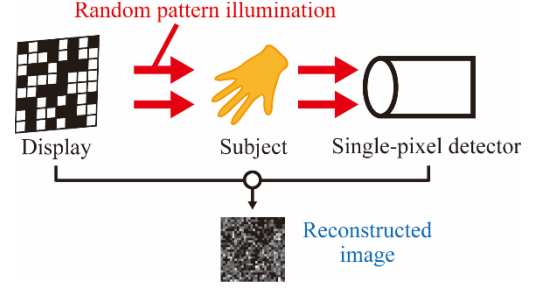
The diagram of single-pixel imaging is shown in Fig. 1. The target object of this study is a subject's hand. The illumination is modulated an arbitrary number of times by a randomly generated mask pattern. Summation of the transmitted light intensity is detected by a point photodetector. By calculating the correlation between the illumination intensity and the detected intensity, it is possible to obtain a reconstructed image as a result of floating-point arithmetic [11]. The detected intensity at by  $k$ -th illumination, denoted by  $B_k$  and obtained by a single-pixel detector, can be expressed as:

$$B_k = \iint_S I_k(x, y)T(x, y) dx dy \quad (1)$$

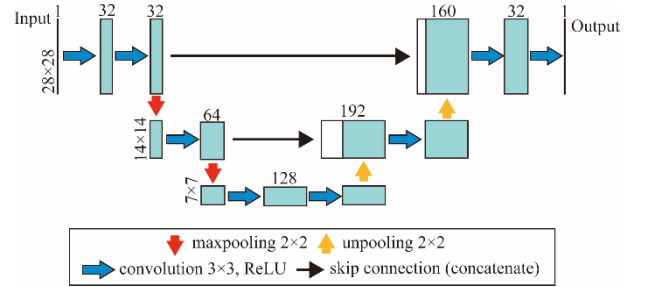
where  $I_k(x, y)$  represents the light intensity of  $k$ -th random mask pattern at the position  $(x, y)$ ;  $T(x, y)$  is the transmittance of the target object on  $(x, y)$ ;  $I_k(x, y)$  is given in advance for  $k = 1, 2, \dots, n$ ;  $n$  is the total number of illuminations; the integration of  $B_k$  is carried out over the area of the object  $S$ . In order to reconstruct the object image, we use the correlation function  $G(x, y)$ .  $G(x, y)$  approaches  $T(x, y)$  with increasing the number of illumination  $n$  [10].  $G(x, y)$  can be expressed as:

$$G(x, y) = \langle \Delta B_k \Delta I_k(x, y) \rangle. \quad (2)$$

$\langle \rangle$  represents the ensemble, which is the average of  $n$  consecutive measurements.  $\Delta I_k(x, y)$  is the deviation between the light intensity of the  $k$ -th random mask pattern and the ensemble.  $\Delta B_k$  is the deviation between the light intensity of the  $k$ -th measurement and the ensemble. If  $N$  consecutive measurements are denoted by  $F_i$ , the ensemble can be expressed as:



**Fig. 1** Diagram of single-pixel imaging to detect a silhouette image of a subject's hand gesture.



**Fig. 2** Example of U-Net architecture. Each box indicates a multi-channel feature map. Numbers on the top and side of the box indicate the number of channels and the size of the feature map, respectively. White boxes are copied feature map on the concatenation. The processes in the arrow are described in the text.

$$\langle F_i \rangle = \frac{1}{N} \sum_{i=1}^N F_i. \quad (3)$$

### 2.2 U-Net

A neural network model called U-Net [19], [30] is used in this study. Figure 2 shows the basic structure of this network, where a  $28 \times 28$  pixel image is used as input for example. 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. The U-net is suitable for capturing the position and contour of objects, and is also used to support single-pixel imaging [19].

## 3. Experiments

To prepare silhouette images of gestures, we took a video of changing gestures among three types ('paper', 'rock' and 'scissors'), as shown in Fig. 3. A video was recorded for 8 minutes by a digital camera (Nikon 1 J3, ISO 800, F2.5, 60 frames/seconds). A twisted-nematic LCD monitor (Dell

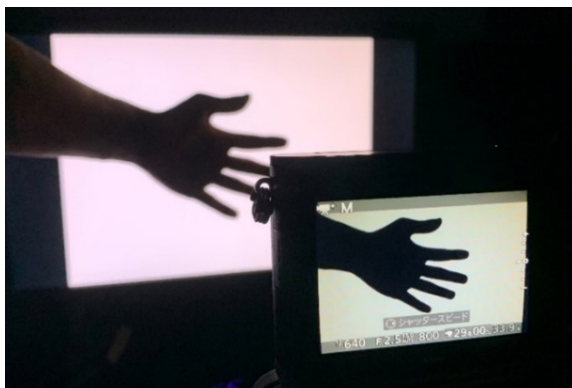


Fig. 3 Scene of taking movie for training data.



Fig. 4 Example of gesture images. Lower images are negative-positive reversed images of Upper shadowgraphs.












Original image	Image converted from reconstructed data with each number of mask pattern				
	100	250	500	784	1000
					
					

Fig. 5 Example of reconstructed data of shadowgraph images.

Alienware AW2310t) was used for a light source. Its luminance was 225.0 cd/m<sup>2</sup>, measured by a luminance meter (Radiant ProMetric IP-PMY29). From the video, 9000 images were exported, and then trimmed and resized so that each image has 28 × 28 pixels. To evaluate the effect of the contrast between the object and its background, we prepared two types of silhouette images, “shadowgraph” (black on white background) and “negative-positive reversed images” (white on black background), in which the personal information is removed. Figure 4 shows examples of the shadowgraphs and reversed images. These images were called as ‘original image’ of gestures.

Next, we performed the numerical experiments on single-pixel imaging and obtained the reconstructed data of those original images. Figure 5 shows the example of reconstructed image when the number of random mask illumination increases. We can see that the reconstructed image will be closer to the original image if the number of illuminations is sufficiently large. For the training and the image restoration with deep learning, we prepared reconstructed data with 250 and 500 illuminations as datasets. Figure 6 shows examples of those datasets. Note that actual reconstructed data

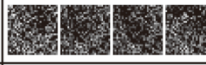
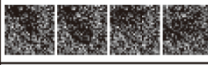
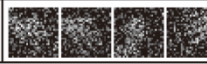
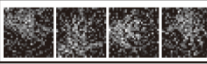
Pattern of original image	Number of mask pattern	
	250	500
Shadowgraph		
Negative-positive reversed		

Fig. 6 Example of reconstructed data of shadowgraphs and negative-positive reversed images.

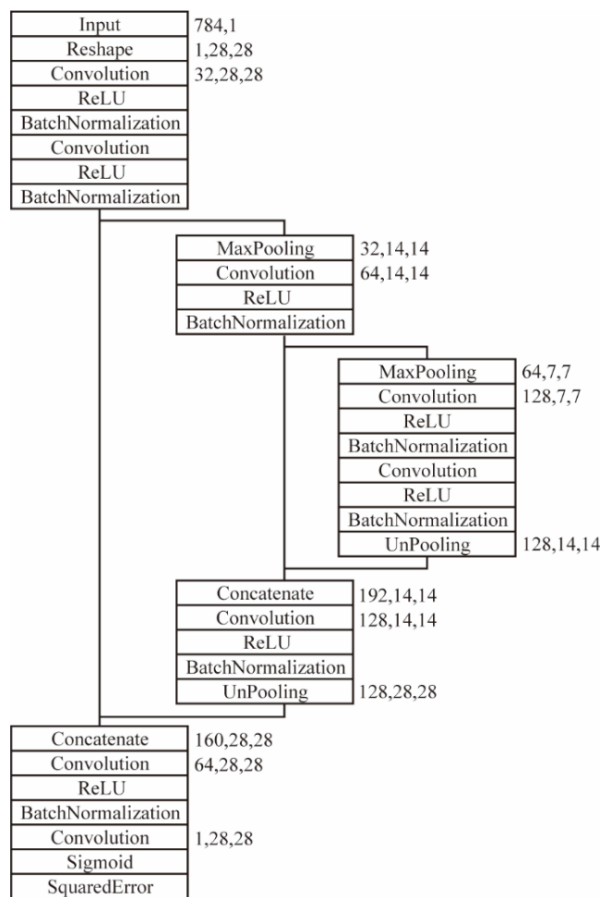


Fig. 7 Structure of our neural network based on U-net. The number on the side of each box means the number of channels and the size of the feature map. This was drawn according to Sony NNC notation.

were float array, while in Fig. 5 and Fig. 6 are shown as images to visualize the data. The preparation of the original image and the numerical simulation of single-pixel imaging were performed by use of Python.

For each of the 250 and 500 illuminations, we prepared 9000 datasets in which original images and reconstructed data were paired. Of these data, 8500 data were used for training a neural network (6800 data for training and 1700 data for validation), and 500 data were used to evaluate the accuracy of image restoration. The reconstructed data were fed to neural network as float array [31]. All of the construction of neural network, training and image restoration were performed on SONY Neural Network Console (Sony NNC) [28]. Figure 7 shows the structure of our neural net-

**Table 1** Detailed parameters in training of deep learning.

Name of parameter	Value
Updater	Adam
Update interval	1
Weight decay	0
Learning rate	0.0001
Beta1, beta2	0.9, 0.999
Batch size	150
Epoch	100

work based on U-net. In Sony NNC, we can build a network by arranging layers represented by blocks like this. Batch normalization was added to improve the speed of training convergence. We adopted squared error as the loss function and used sigmoid as the activation function before the loss function. In training, Adam [32] was used as updater, and hyperparameters are shown in Table 1. Those experiments of deep learning were performed on a custom desktop computer (Windows10, Core i7-6400k and GeForce GTX TITAN X).

To evaluate the accuracy of image restoration, we used peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [33]. PSNR calculates the difference of luminance of each pixel between original image and restored image. The equations for PSNR can be expressed as:

$$\text{PSNR} = 10 \cdot \log_{10} \left( \frac{\text{MAX}^2}{\text{MSE}} \right) \quad (4)$$

where MAX refers to the maximum value that each pixel can take, 255 in this case: MSE indicates mean squared error between two images. SSIM quantifies the difference in brightness, contrast, and structure between the two images. The equations for SSIM can be expressed as:

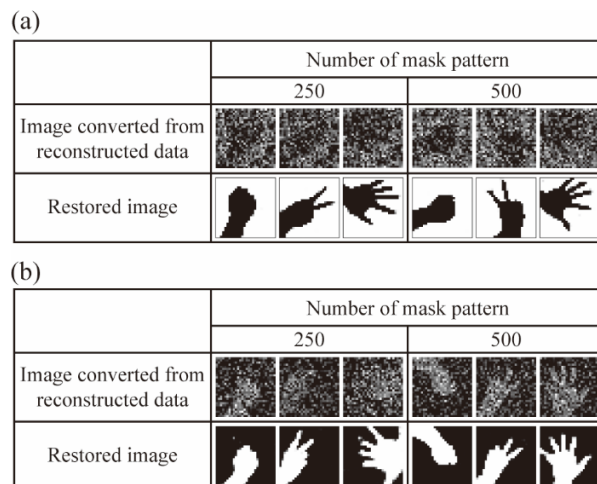
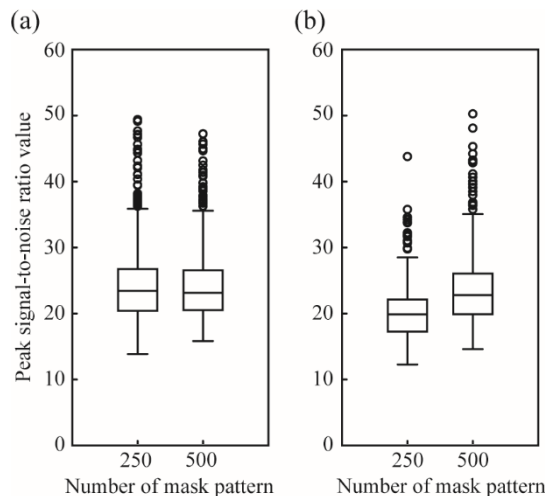
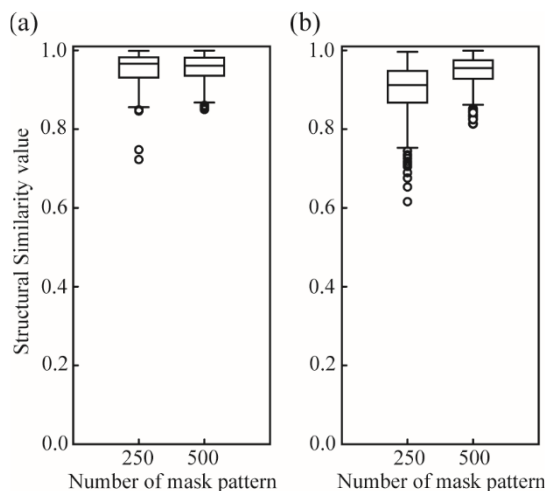
$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (5)$$

where  $\mu$  and  $\sigma$  indicate mean and standard deviation in the small window, respectively.  $\sigma_{xy}$  is the covariance of  $x$  and  $y$ . SSIM takes a value from 0 to 1. Both PSNR and SSIM become larger as the restored image gets closer to the original image. These values were calculated on Python.

#### 4. Results

Figure 8 shows examples of restored images. With these numbers of illuminations in the reconstruction, especially with 250 times, our eyes cannot tell which gesture it is from the reconstructed image. However, in both cases using shadowgraph and negative-positive reversed image as original image, we successfully built a neural network that can restore the images close to original images.

Figure 9 and Fig. 10 show the boxplot of PSNR and SSIM of restored images, respectively. The median PSNR of restored images was more than 20, and their median SSIM was more than 0.9 for each combination of the original image type and the number of random mask illumination. The PSNR and SSIM of the image obtained from the reconstructed data with 10000 illuminations (Fig. 5) were

**Fig. 8** Examples of restored images by our neural network, (a) Shadowgraph image (b) Negative-positive reversed image.**Fig. 9** PSNR boxplot of restored images when using (a) shadowgraph images and (b) negative-positive reversed images as original images.**Fig. 10** SSIM boxplot of restored images when using (a) shadowgraph images and (b) negative-positive reversed images as original images.



7.86 and 0.56, respectively.

Thus, our neural network was able to restore images much closer to the original images even from the reconstructed data with 1/40 of the illumination numbers. In the actual experiment, we plan to use high speed LEDs at 4320 Hz [34] to illuminate the mask patterns. In order to make the mask pattern imperceptible to the human eye, it is necessary to integrate spatio-temporally coded images and alter them at a sufficiently high frame rate [35]. Thus, the number of masks displayed will be doubled and it will take 4629 milliseconds for 10000 mask patterns (20000 display images). On the other hand, the time required for 250 mask patterns (500 display images) is 116 milliseconds. Processing time to obtain gesture silhouette image is largely reduced, however, to integrate systems for image reconstruction, restoration, and gesture recognition and to achieve their real-time processing, the image should be reconstructed with fewer illuminations.

When shadowgraph images were used, the median value of the evaluation index of restored images did not change even with the reduced number of mask patterns, as shown in Fig. 9(a) and Fig. 10(a). On the other hand, when negative-positive reversed images were used, the median value of the evaluation index of restored images became lower as the number of mask pattern decreased. The same trend was observed when using PSNR and SSIM were used, as shown in Fig. 9(b) and Fig. 10(b). These results suggest that shadowgraph is more suitable as original image for restoration by our neural network, compared to negative-positive reversed image, when the number of random mask illumination is smaller. One of the reasons of this difference is asymmetric processing on positive and negative values in our network. Maxpooling keeps only positive correlation peak values.

Tables 2 and 3 show the median values of PSNR and SSIM of the restored images according to the type of gesture, for shadowgraph and negative-positive reversed images. In both evaluation indices, there was a trend that the restoration accuracy decreased in the order of Rock, Scissor, and Paper. This trend was more significant when the number of illuminations was 250, and when the original

**Table 2** Median values of PSNR and SSIM of restored images when using shadowgraph image according to gesture types

Method for evaluation	Number of mask pattern	Gesture type		
		Paper	Rock	Scissor
PSNR	250	21.3	26.7	24.1
	500	21.2	27.4	24.4
SSIM	250	0.94	0.97	0.96
	500	0.94	0.98	0.96

**Table 3** Median values of PSNR and SSIM of restored images when using Negative-positive reversed images according to gesture types

Method for evaluation	Number of mask pattern	Gesture type		
		Paper	Rock	Scissor
PSNR	250	17.5	23.4	20.1
	500	20.0	25.9	23.6
SSIM	250	0.87	0.95	0.91
	500	0.93	0.96	0.96

image was the negative-positive reversed image. This result suggests that the restoration accuracy decreases as the structural complexity of the gesture increases. The difference in restoration accuracy between gestures may have a negative impact when implemented in gesture recognition devices. If the number of illuminations is further reduced, it will be necessary to devise methods to keep the restoration accuracy close between gestures when preparing training data and building neural networks.

## 5. Conclusion

Single-pixel imaging was introduced to solve the privacy problem in gesture recognition for interactive acquisition of information. Deep learning drastically reduced the number of illuminations to reconstruct silhouette images of gestures.

By using a simple method with a display as a light source, we were able to prepare a large number of silhouette images of gestures on our own as the source of training data. Our neural network, based on U-net, was able to recover images close to the original images even from reconstructed data with greatly reduced illumination counts of 250 or 500. We also prepared two patterns of silhouette images of the gesture: shadowgraph (black on white background) and negative-positive reversed images (white on black background) and examined the difference in restoration accuracy. As a result of our numerical experiments, we found that the accuracy of the restored image was lower for negative-positive-reversed images when the number of illuminations was small. This result indicates that using shadowgraph as a silhouette image of the gesture is suitable for further reducing the number of illuminations in our neural network. Moreover, we compared the restoration accuracy among gestures and found that the restoration accuracy decreased in the order of rock, scissor, and paper. Since the difference in the restoration accuracy between gestures may have a negative impact on the implementation of gesture recognition devices, it is necessary to prepare training data and construct neural networks with care to avoid the restoration accuracy between gestures.

In future research, if the neural network alone cannot provide sufficient restoration accuracy when the number of mask patterns is further reduced, one solution would be to combine it with other illumination methods proposed for single-pixel imaging [36], [37]. It is also important to verify the restoration of gesture images using single-pixel imaging on actual devices, although only numerical simulations were performed in this study. As a result, when the imaging area needs to be enlarged for accurate gesture recognition, single-pixel imaging may become possible by applying our proposed aerial display technology [38]–[40].

## Acknowledgments

A part of this work was supported by JST/ACCEL (grant no. JPMJAC1601) and JSPS KAKENHI (19H00871, 20H05702).

## References

- [1] B. Javidi, F. Pla, J.M. Sotoca, X. Shen, P. Latorre-Carmona, M. Martínez-Corral, R. Fernández-Beltrán, and G. Krishnan, “Fundamentals of automated human gesture recognition using 3D integral imaging: a tutorial,” *Adv. Opt. Photonics*, vol.12, no.4, pp.1237–1299, 2020.
- [2] H. Yamamoto, M. Yasui, M.S. Alvisalim, M. Takahashi, Y. Tomiyama, S. Suyama, and M. Ishikawa, “Floating display screen formed by AIRR (Aerial imaging by retro-reflection) for interaction in 3D space,” 2014 International Conference on 3D Imaging (IC3D), pp.1–5, IEEE, 2014.
- [3] N. Rossol, I. Cheng, and A. Basu, “A Multisensor Technique for Gesture Recognition Through Intelligent Skeletal Pose Analysis,” *IEEE Trans. Human-Mach. Syst.*, vol.46, no.3, pp.350–359, 2016.
- [4] M. Nishihori, T. Izumi, Y. Nagano, M. Sato, T. Tsukada, A.E. Kropp, and T. Wakabayashi, “Development and clinical evaluation of a contactless operating interface for three-dimensional image-guided navigation for endovascular neurosurgery,” *Int. J. Comput. Assist. Radiol. Surg.*, vol.16, pp.663–671, 2021.
- [5] J. Dai, J. Wu, B. Saghafi, J. Konrad, and P. Ishwar, “Towards privacy-preserving activity recognition using extremely low temporal and spatial resolution cameras,” 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp.68–76, IEEE, 2015.
- [6] Z. Wu, Z. Wang, Z. Wang, and H. Jin, “Towards privacy-preserving visual recognition via adversarial training: A pilot study,” *Proc. European Conference on Computer Vision (ECCV)*, pp.606–624, Springer, 2018.
- [7] J.H. Shapiro, “Computational ghost imaging,” *Phys. Rev. A - At. Mol. Opt. Phys.*, vol.78, 061802, 2008.
- [8] B. Sun, M.P. Edgar, R. Bowman, L.E. Vittert, S. Welsh, A. Bowman, and M.J. Padgett, “3D computational imaging with single-pixel detectors,” *Science (80-.)*, vol.340, no.6134, pp.844–847, 2013.
- [9] G.M. Gibson, S.D. Johnson, and M.J. Padgett, “Single-pixel imaging 12 years on: a review,” *Opt. Express*, vol.28, no.19, pp.28190–28208, 2020.
- [10] Y. Bromberg, O. Katz, and Y. Silberberg, “Ghost imaging with a single detector,” *Phys. Rev. A*, vol.79, 53840, 2009.
- [11] K. Shibuya, K. Nakae, Y. Mizutani, and T. Iwata, “Comparison of reconstructed ghost imaging and Hadamard transform imaging,” *Opt. Rev.*, vol.22, pp.897–902, 2015.
- [12] Z. Zhang, X. Ma, and J. Zhong, “Single-pixel imaging by means of Fourier spectrum acquisition,” *Nat. Commun.*, vol.6, 6225, 2015.
- [13] Z. Zhang, X. Wang, G. Zheng, and J. Zhong, “Fast Fourier single-pixel imaging via binary illumination,” *Sci. Rep.*, vol.7, 12029, 2017.
- [14] M.F. Duarte, M.A. Davenport, D. Takhar, J.N. Laska, T. Sun, K.F. Kelly, and R.G. Baraniuk, “Single-pixel imaging via compressive sampling,” *IEEE Signal Process. Mag.*, vol.25, no.2, pp.83–91, 2008.
- [15] O. Katz, Y. Bromberg, and Y. Silberberg, “Compressive ghost imaging,” *Appl. Phys. Lett.*, vol.95, 131110, 2009.
- [16] P.G. Vaz, D. Amaral, L.F.R. Ferreira, M. Morgado, and J. Cardoso, “Image quality of compressive single-pixel imaging using different Hadamard orderings,” *Opt. Express*, vol.28, no.8, pp.11666–11681, 2020.
- [17] K. Shibuya, T. Minamikawa, Y. Mizutani, H. Yamamoto, K. Minoshima, T. Yasui, and T. Iwata, “Scan-less hyperspectral dual-comb single-pixel-imaging in both amplitude and phase,” *Opt. Express*, vol.25, no.18, pp.21947–21957, 2017.
- [18] M. Lyu, W. Wang, H. Wang, H. Wang, G. Li, N. Chen, and G. Situ, “Deep-learning-based ghost imaging,” *Sci. Rep.*, vol.7, 17865, 2017.
- [19] T. Shimobaba, Y. Endo, T. Nishitsuji, T. Takahashi, Y. Nagahama, S. Hasegawa, M. Sano, R. Hirayama, T. Kakue, A. Shiraki, and T. Ito, “Computational ghost imaging using deep learning,” *Opt. Commun.*, vol.413, pp.147–151, 2018.
- [20] Y. He, G. Wang, G. Dong, S. Zhu, H. Chen, A. Zhang, and Z. Xu, “Ghost imaging based on deep learning,” *Sci. Rep.*, vol.8, 6469, 2018.
- [21] A.L. Mur, F. Peyrin, and N. Ducros, “Recurrent Neural Networks for Compressive Video Reconstruction,” 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp.1651–1654, 2020.
- [22] I. Hoshi, T. Shimobaba, T. Kakue, and T. Ito, “Single-pixel imaging using a recurrent neural network combined with convolutional layers,” *Opt. Express*, vol.28, no.23, pp.34069–34078, 2020.
- [23] C.F. Higham, R. Murray-Smith, M.J. Padgett, and M.P. Edgar, “Deep learning for real-time single-pixel video,” *Sci. Rep.*, vol.8, 2369, 2018.
- [24] W. Jiang, X. Li, X. Peng, and B. Sun, “Imaging high-speed moving targets with a single-pixel detector,” *Opt. Express*, vol.28, no.6, pp.7889–7897, 2020.
- [25] S. Rizvi, J. Cao, K. Zhang, and Q. Hao, “DeepGhost: real-time computational ghost imaging via deep learning,” *Sci. Rep.*, vol.10, 11400, 2020.
- [26] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol.86, no.11, pp.2278–2324, 1998.
- [27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp.248–255, IEEE, 2009.
- [28] Sony, “Neural Network Console,” <https://dl.sony.com/>.
- [29] N. Mukojima, M. Yasugi, Y. Mizutani, T. Yasui, and H. Yamamoto, “Deep-Learning-Assisted Single-Pixel Imaging for Gesture Recognition Considering Privacy,” *Proc. International Display Workshops*, vol.27, pp.985–988, 2020.
- [30] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” *arXiv:1505.04597*, 2015.
- [31] M. Yasugi, Y. Mizutani, T. Yasui, and H. Yamamoto, “Deep Learning for Single-Pixel Imaging Without Normalization and Image Output,” *JSAP-OSA Joint Symposia 2020*, pp.9p-Z10-6, 2020.
- [32] D.P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *arXiv:1412.6980*, 2017.
- [33] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol.13, no.4, pp.600–612, 2004.
- [34] T. Tokimoto, S. Suyama, and H. Yamamoto, “4320-Hz LED Display With Pulse-Width Modulation by Use of a Nonlinear Clock,” *J. Disp. Technol.*, vol.12, no.12, pp.1581–1587, 2016.
- [35] S. Onose, M. Takahashi, Y. Mizutani, T. Yasui, and H. Yamamoto, “Single pixel imaging with a high-frame-rate LED digital signal,” *Proc. International Display Workshops*, vol.23, pp.1495–1498, 2016.
- [36] X. Yuan and Y. Pu, “Parallel lensless compressive imaging via deep convolutional neural networks,” *Opt. Express*, vol.26, no.2, pp.1962–1977, 2018.
- [37] F. Li, M. Zhao, Z. Tian, F. Willomitzer, and O. Cossairt, “Compressive ghost imaging through scattering media with deep learning,” *Opt. Express*, vol.28, no.12, pp.17395–17408, 2020.
- [38] M. Nakajima, K. Onuki, I. Amimori, and H. Yamamoto, “Polarization State Analysis for Polarized Aerial Imaging by Retro-Reflection (PAIRR),” *Proc. IDW 22*, 429–432, 2015.
- [39] S. Morita and H. Yamamoto, “Single Pixel Imaging with pAIRR,” *OPJ-OSA Joint Symposia on Nanophotonics and Digital Photonics*, 31aOD5, 2017.
- [40] S. Morita, S. Onose, M. Sasaki, and H. Yamamoto, “Single Pixel Imaging on Aerial Display with AIRR,” *Proc. IDW 17*, pp.958–961, 2017.



**Naoya Mukojima** is currently a graduated student of Department of Optical Engineering at Utsunomiya University. He received Bachelor of Engineering degree from the Dept. of Information Science at Utsunomiya University. His research interests include deep-learning assisted single-pixel imaging. He was a recipient of Outstanding Poster Paper Award at IDW'20.



**Masaki Yasugi** is currently a project associate professor of Center for Optical Research and Education (CORE), at Utsunomiya University. He received PhD from Kyoto University, Japan in 2012. He was a researcher in Kyoto University from April 2012 to March 2015, and a NIBB Research Fellow in The National Institute for Basic Biology from April 2015 to March 2018. His research interests are animal ecology, ethology, cognitive science and information photonics. Recently, he focuses on the application

of computer-graphics, neural network for image processing, and aerial floating display to our lives and the experiment for animals. He was a recipient of Zoological Science Award in 2018, Best Poster Paper Award at IMID2018, and Outstanding Poster Paper Award at IDW'20.



**Yasuhiro Mizutani** obtained his Ph.D. degree in Mechanical engineering from Tokyo university of Agriculture & Technology, Japan, in 2008. He has also BE and ME degrees in nuclear engineering from Osaka university in 1997 and 1999, respectively. From 1999 to 2003, he joined a researcher at Panasonic corporation. Then, from 2003 to 2009, he worked as a research associate in the Department of Mechanical systems engineering at Tokyo university of Agriculture & Technology, Japan. He joined the

University of Tokushima, Japan, in 2009 as an associate professor. He is currently an associate professor at Osaka university from 2015. Professor Mizutani's research interests include interferometry, polarimetry, 3D surface measurement, optical trapping and 3D lithography. In these areas, he has published over 100 papers in refereed international journals and conferences. He is a member of SPIE and OSA.



**Takeshi Yasui** received the first Ph.D. degree in engineering from the University of Tokushima in 1997, and the second Ph.D. degree in medical science from the Nara Medical University in 2013. From 1997 to 1999, he worked as a Post-Doctoral Research Fellow in the National Research Laboratory of Metrology, Japan. He was with the Graduate School of Engineering Science, Osaka University from 1999 to 2010, and was briefly with the University of Bordeaux I in 2007 and 2012, and with the Uni-

versity of Littoral Côte d'Opale in 2010 as an Invited Professor. He is currently a Director and a Professor in the Institute of Post-LED Photonics, Tokushima University, Vice Director of Research Support in the same university since 2016. His research interests include THz instrumentation and metrology, second-harmonic-generation microscopy, and optical frequency comb.



**Hirotsugu Yamamoto** was born in Wakayama, Japan, in 1971. He received his B.E., M.E., and Ph.D. degrees from the University of Tokyo. From 1996 to 2008, he was an assistant professor at Tokushima University. From 2009 to 2014, he was an associate professor at Tokushima University. From 2014 to 2019, he was an associate professor at Utsunomiya University. Since Sep. 2019, he has been a full professor at Utsunomiya University. His recent work activities have included aerial

display, aquatic display, a high-speed LED display, information photonics, and VR biology. He was a recipient of Young Scientist Award for the Presentation of an Excellent Paper, The Japan Society of Applied Physics, Outstanding Poster Paper Award at IDW'03, IDW'04, IDW'07, IDW'08, IDW'09, IDW'10, IDW'11, IDW'12, IDW'13, IMID2014, IDW'14, IMID2015, IDW'15, IMID2016, IDW'16, IDW'17, IMID2018, IDW'18, and IDW'20, Best Paper Award at DHIP2011, IDW'11, and IDW'15, Best 3D Demonstration Award at Stereoscopic Displays and Applications 2012, the Gen-Nai Grand Prize from the Ozaki Foundation of Japan, and SPIE the Fumio Okano Best 3D Paper.