

Image-free single-pixel semantic segmentation for complex scene based on multi-scale U-Net

Tengfei Liu(刘腾飞), Yanfeng Bai(白艳锋), Jianxia Chen(陈健霞), Jintao Zhai(翟锦涛), Siqing Xiang(向思卿), Xianwei Huang(黄贤伟), and Xiquan Fu(傅喜泉)

Citation: Chin. Phys. B, 2026, 35 (**1**): 014202. DOI: 10.1088/1674-1056/ade4b0
http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Color Fourier single-pixel imaging with random color filter array

Jialiang Chen(陈佳亮), Wei Zhu(朱维), Le Wang(王乐), and Shengmei Zhao(赵生妹)
Chin. Phys. B, 2025, 34 (**9**): 094210. DOI: 10.1088/1674-1056/adf1e9

Reconstruction algorithm for cross-waveband optical computing imaging

Jin-Tao Xie(谢锦涛), Shu-Hang Bie(别书航), Ming-Fei Li(李明飞), Yuan-Jin Yu(余远金), Yi-Fei Li(李毅飞), Jin-Guang Wang(王进光), Bao-Gang Quan(全保刚), and Ling-An Wu(吴令安)
Chin. Phys. B, 2024, 33 (**10**): 104203. DOI: 10.1088/1674-1056/ad6ccc

Efficient single-pixel imaging encrypted transmission based on 3D Arnold transformation

Zhen-Yu Liang(梁振宇), Chao-Jin Wang(王朝瑾), Yang-Yang Wang(王阳阳), Hao-Qi Gao(高皓琪), Dong-Tao Zhu(朱东涛), Hao-Li Xu(许颢砾), and Xing Yang(杨星)
Chin. Phys. B, 2024, 33 (**3**): 034204. DOI: 10.1088/1674-1056/acf204

Complex-amplitude Fourier single-pixel imaging via coherent structured illumination

Hong-Yun Hou(侯红云), Ya-Nan Zhao(赵亚楠), Jia-Cheng Han(韩佳成), De-Zhong Cao(曹 德忠),Su-Heng Zhang(张素恒), Hong-Chao Liu(刘宏超), and Bao-Lai Liang(梁宝来)
Chin. Phys. B, 2023, 32 (**6**): 064201. DOI: 10.1088/1674-1056/ac8e9a

Fast Fourier single-pixel imaging based on Sierra-Lite dithering algorithm

Zhen-Yu Liang(梁振宇), Zheng-Dong Cheng(程正东), Yan-Yan Liu(刘严严), Kuai-Kuai Yu(于快快), Yang-Di Hu(胡洋頔)
Chin. Phys. B, 2019, 28 (**6**): 064202. DOI: 10.1088/1674-1056/28/6/064202

Image-free single-pixel semantic segmentation for complex scene based on multi-scale U-Net

Tengfei Liu(刘腾飞), Yanfeng Bai(白艳锋)[†], Jianxia Chen(陈健霞), Jintao Zhai(翟锦涛), Siqing Xiang(向思卿), Xianwei Huang(黄贤伟), and Xiquan Fu(傅喜泉)[‡]

College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, *China*

(Received 29 March 2025; revised manuscript received 12 May 2025; accepted manuscript online 16 June 2025)

Single-pixel imaging (SPI) receives widespread attention due to its superior anti-interference capabilities, and image segmentation technology can effectively facilitate its recognition and information extraction. However, the complexity of the target scene and plenty of imaging time in SPI make it challenging to achieve high-quality and concise segmentation. In this paper, we investigate the image-free intricate scene semantic segmentation in SPI. Using "learned" illumination patterns allows for the full extraction of the object's spatial information, thereby enabling pixel-level segmentation results through the decoding of the received measurements. Simulation and experimentation show that, in the absence of image reconstruction, the mean intersection over union (MIoU) of segmented image can reach higher than 85%, and the Dice coefficient (DICE) close to 90% even at the sampling ratio of 5%. Our approach may be favorable to applications in medical image segmentation and autonomous driving field.

Keywords: single-pixel imaging, image-free, deep learning, complex scene

PACS: 42.30.–d, 42.30.Va, 42.30.Tz

DOI: 10.1088/1674-1056/ade4b0 **CSTR:** 32038.14.CPB.ade4b0

1. Introduction

Modern visual applications need to deal with a large number of complex scenes, especially in dynamic environments where occlusion and the similarity between objects present inevitable challenges.[1] For example, in the field of autonomous driving and the overlapping scenes in medical images, accurate identification is indispensable to support subsequent decision-making. Image segmentation plays an important role in the field of computer vision and image processing,[2] which is used to divide an image into multiple meaningful regions or objects. Different from the image classification which divides the image into a label or a class, the detail of whole scene is receiving more attention through image segmentation. It provides a better understanding of the target shape and specific details, and prepares for subsequent object recognition. There are many common methods for segmentation, including thresholding, edge detection, region growing, clustering methods,[3–5] etc. In recent years, image segmentation methods based on deep learning, such as U-Net, fully convolutional networks (FCN), and mask R-CNN,[6–9] have attracted much attention due to their excellent performance in complex scenes. These methods use convolutional neural network (CNN) to automatically learn image features, achieve higher precision and more detailed segmentation results.

However, high-fidelity images and the choice of appropriate algorithms are prerequisites for achieving excellent segmentation results. But the acquisition of high-fidelity images relies on complex imaging hardware, which makes it challenging to implement in resource-constrained environments or for real-time applications. Recently, the image-free sensing method based on SPI is gaining popularity. By combining the unique imaging characteristics of SPI with the strength of deep learning for feature learning, special hardware defects can be overcome. According to the optimized structured illumination and decoding network, the effective feature information of the target scene can be captured by network-optimized patterns, and the trained decoder can implement specific tasks based on the received light intensity vector in the absence of an image. Due to its foundation on SPI,[10–12] it has natural advantages in imaging radar,[13,14] target recognition,[15–20] and special bands.[21–23] Therefore, it is also possible to obtain sufficient information about the target scene under these harsh conditions.

In the most recent studies, there is also some research on the segmentation of the target scene based on image-free method. The group of Bian employed a two-stage training strategy to optimize network patterns, achieving excellent segmentation performance on the various datasets according to one-dimensional measurements.[24] Taking it a step further, by utilizing an array-net, Liu *et al*. implement high quality image reconstruction and segmentation based only on one-dimensional (1D) bucket detection signals at low sampling times.[25] These methods dismiss the traditional patterns and

[†]Corresponding author. E-mail: yfbai@hnu.edu.cn

[‡]Corresponding author. E-mail: fuxq@hnu.edu.cn

© 2026 Chinese Physical Society and IOP Publishing Ltd. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

turn to network-optimized patterns for illumination, making accurate judgments only based on the 1D information. However, in general, these methods incur extra time costs due to the setup of intermediate steps for the preliminary recovery of image spatial features and have limited investigation on multi-object and occlusion scene. Besides, the choice of loss function in Ref. [24] makes the research tend to be image restoration instead of image segmentation, which can be seen from blur of the result images.

In this paper, we propose an image-free segmentation scheme that is used to deal with complex scenarios. Illumination patterns are optimized during network training, only a detector without spatial resolution is relied to capture a scene's total light intensity, thereby the trained decoder can implement image segmentation in case of no image. Specifically, our contributions are summarized as follows.

The segmentation of complex scenes, including overlapping, occluded, and visually similar objects, is investigated by considering the relationships between different categories at the pixel level, resulting in superior segmentation performance.

A multi-scale U-Net architecture is proposed to enhance semantic segmentation through the recovery of feature information from compressed 1D vectors.

The need for image reconstruction during both the training and segmentation phases is eliminated, significantly reducing data storage requirements and alleviating computational burden.

2. Principle and methods

2.1. Principle

The experiment setup of SPI is illustrated in Fig. 1. The most basic process involves two ways: structured illumination and structured detection. In the process of structured illumination, a pulsed laser is firstly irradiated on a DMD (digital micromirror device: DLP7000) for modulation. By controlling the tilt state of each microlens on the DMD, the incident beam is selectively redirected to perform spatial filtering, thereby generating different projection patterns. After passing through the collimation system (the focal length of lens L1 and L2 are 50 mm and 100 mm, respectively), the illumination patterns on the DMD are irradiated on the object, and then a single-pixel detector (Thorlabs PDA100A2, the gain ranges from 0 to 70 dB) is used to measure the total backscattered light.

By irradiating a range of prefabricated patterns $R(x,y)_m$ (m is the pattern sequence number) onto a DMD using a laser, the detector collects different intensity signals I_m multiple times, and the estimation of the two-dimensional (2D) image can be obtained by[26]

$$O(x,y) = \frac{1}{M}\sum_{m=1}^{M} I_m R(x,y)_m, \tag{1}$$

where M is the total number of patterns. Obviously, the reconstructed image is the weighted sum of patterns. Specifically, the light intensity value I_m can be expressed as[27]

$$I_m = R(x,y)_m s + \varepsilon, \tag{2}$$

where ε denotes the possible noise during measurement, and s represents the target object.

Assuming that the noise generated during the measurement is ignored, equation (2) states that the light intensity value I_m is the result of the interaction between the prefabricated pattern $R(x,y)_m$ and the object s, namely the inner product calculation result of the illumination patterns and the original object.

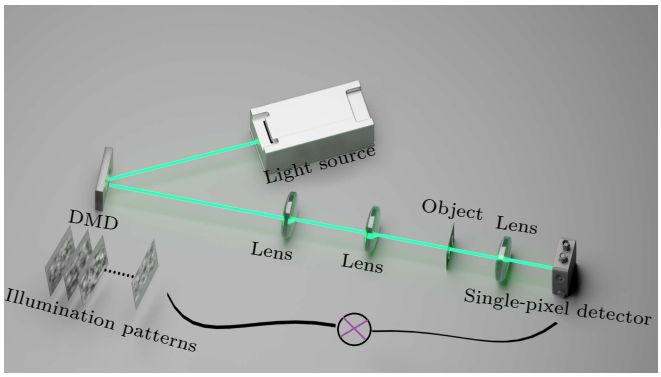


Fig. 1. Schematic diagram of the SPI.

In our design, network convolutional layers are treated as modulation patterns. For convolution operations, the convolution kernel moves step by step across the input 2D image to capture specific spatial information, performing an element-wise product at each position and then summing the results to obtain a single output value. The convolution process can be abbreviated as

$$O(i,j) = \sum_{m=0}^{k-1}\sum_{n=0}^{k-1} I(i+m, j+n)\cdot K(m,n), \tag{3}$$

where $K(m,n)$ is the convolution kernel, and $I(i+m, j+n)$ denotes the input image. This process will generate the feature map $O(i,j)$, which is the extraction of some spatial features of the original image. Therefore, once the convolution kernel is equivalent to the image size, only a single value can be output, and equation (3) is reduced to the following form:

$$O(i,j) = \sum_{i=0}^{k-1}\sum_{j=0}^{k-1} I(i,j)\cdot K(i,j). \tag{4}$$

It means that when we have a sufficient number of kernels, various spatial information of the image can be fully extracted, which corresponds exactly to the variation of sampling ratio (SR) in SPI.

However, the values in the convolution kernel need to be manually designed to extract specific spatial information,

which makes it hard to obtain the optimal value. Besides, A certain type of fixed pattern such as random pattern or Hadamard pattern, is not considered due to its fixed characteristics towards a variety of objects. By making use of them, it may not be able to obtain the most effective measurements, which may lead to pending segmentation results. Therefore, we consider using neural network training to optimize the dynamic convolution kernels. During the network training stage, both illumination patterns and decoding network are optimized together. A series of "learned" convolution layers equivalent to the image size as structured illumination patterns, so the patterns can be applied to a wide variety of objects.

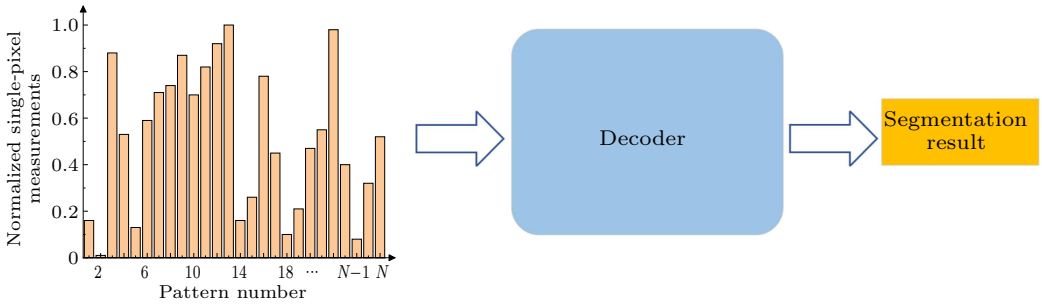


Fig. 2. Schematic diagram of receiving 1D vector for image segmentation.

Then the subsequent image segmentation is carried on. The segmentation result is obtained by inputting N single-pixel measurements into the decoder of the trained network, as shown in Fig. 2. As long as the spatial information contained in the 1D vector is sufficient, it can be decoded by the decoder to achieve excellent segmentation results.

2.2. Multi-scale U-Net

For image segmentation tasks, the most commonly used network nowadays is U-Net.[6] It is known for its skip connections and efficient segmentation performance. But it is based on an entire image to be input. Therefore, the extracted single-pixel measurements need to be preprocessed.

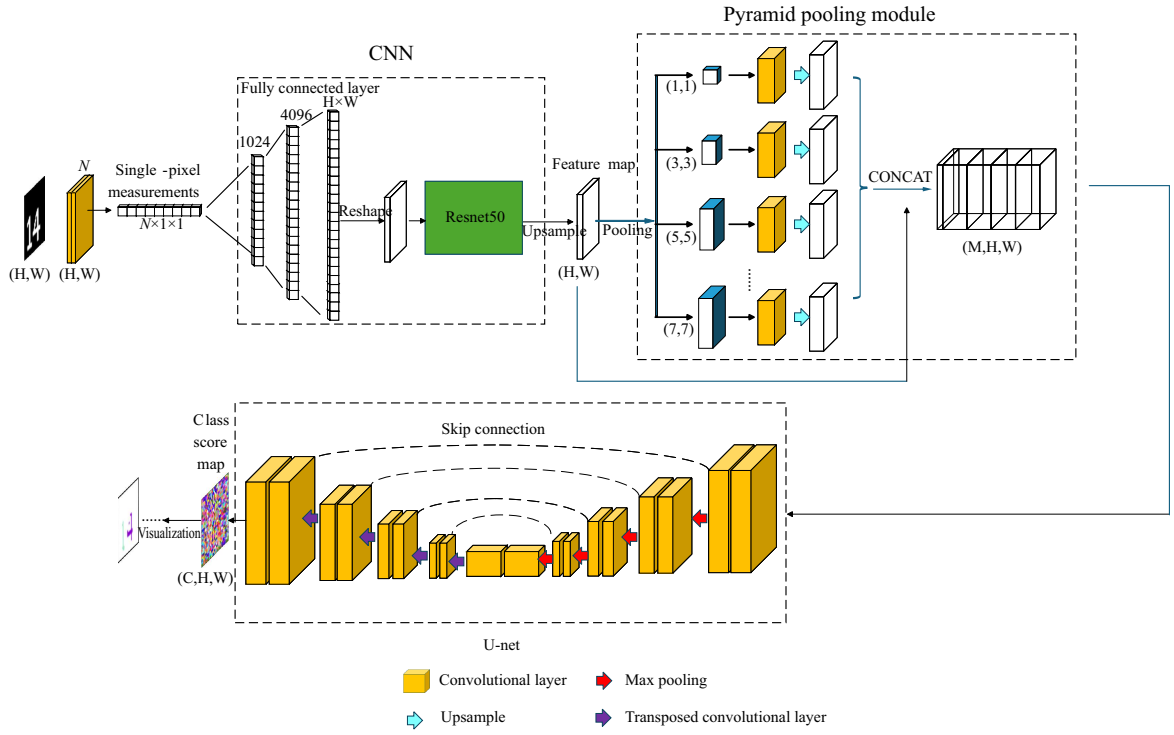


Fig. 3. Network structure of multi-scale U-Net.

Our network architecture is composed of an encoder and a decoder, which is depicted in Fig. 3. Specifically, the convolutional layer utilized for illumination functions as the encoder, tasked with extracting feature information from the input image. On the other hand, the decoder is a combination of CNN, pyramid pooling module (PPM),[28] and U-Net, managing to decode the compressed image information to achieve image segmentation. However, the original rich image in-

formation is significantly condensed into 1D vectors, leading to the loss of a portion of the original detailed image data. Consequently, this 1D feature map necessitates sufficient feature information restoration to facilitate the subsequent segmentation task. To this end, the decoder is divided into two modules: the feature information restoration module and the segmentation module. The former, constituted by CNN and PPM, is designed to recover the vital features within the image; the latter relies on U-Net to execute the ultimate segmentation task. Initially, single-pixel measurements recover spatial feature information via CNN. The CNN includes fully connected layers and the initial stages of ResNet50,[8] with partial loading of the pre-trained weights of ResNet50 on ImageNet. With over 1.2 million training images in the ImageNet dataset, the ResNet50 model's pre-trained weights demonstrate robust general feature extraction capabilities. Therefore, when training starts, only fine-tuning is performed on the ResNet50 segment, thereby significantly enhancing training efficiency. Subsequently, to capture contextual information at various scales within the output feature map, PPM is employed to pool the feature map at different scales and then fuses them, thereby augmenting the model's perceptual capacity. Ultimately, the feature map that integrates multi-scale features is fed into a U-Net architecture to achieve pixel-level image segmentation.

In the training process, the network not only learns to produce a better segmented image but also prepares for the patterns required for the SPI process. Once the training is completed, the convolution kernels that have been optimized in the network are used as the structured illumination patterns. Some features of the input network image are perceived. In the testing stage, only the single-pixel detection values detected by the single-pixel detector are input into the decoder of the network to acquire the image segmentation result, as shown in Fig. 2. Overall, the network adopts an end-to-end training strategy, eliminating the need for intermediate layer validation, thereby ensuring that complex data processing and lengthy training times are avoided. Furthermore, the entire segmentation process is imageless, which guarantees the high efficiency of data acquisition.

2.3. Dataset

For semantic segmentation in complex cases, we first consider the problems of overlapping and occlusion of objects. Therefore, it is particularly important to select a dataset with a large number of occluded object annotations. Since this kind of suitable dataset is relatively rare, we consider using a large number of simple image combinations to obtain the required dataset. By superimposing two images with the same resolution, some objects in one image are overlapped or blocked by objects in another image, which meets the requirements of the complex scene studied. Here, we contemplate the utilization of the EMINST dataset,[29] which boasts a more voluminous data corpus, as an augmentation to the MNIST dataset commonly used for classification tasks. In addition, by using a character grayscale image with a specific fixed font (printed dataset), it is considered to randomly combine it with the image in EMINST. To ensure data balance, each type of character in the EMINST dataset and printed dataset is selected evenly. After sufficient random combination, data augmentation is considered to be used, and the position, size and rotation angle of superimposed characters in the image are randomly changed to ensure the credibility of the final dataset. The final dataset contains 5×10^4 images. For the dataset, the effectiveness of image segmentation is reflected in the separation of different characters. Each pixel in the image may belong to a different character concurrently.

2.4. Loss function

For the image segmentation task, it is treated as a pixel-level classification task, which achieves the semantic segmentation by classifying each pixel specifically. So the loss function in classification is still suitable for image segmentation. In the selection of a loss function for classification problems, different loss functions are specifically applicable to various classification scenarios. Based on the analysis, since we investigate image segmentation involving occlusions and overlaps, it can be regarded as a multi-label image segmentation task. In this context, a given pixel may belong to multiple categories. Consequently, we assume that each category is treated as independent of the others, and the network performs a binary classification for each pixel with respect to each category. In this work, binary cross entry loss (BCE Loss) is considered. Additionally, Dice loss (DICE Loss) is particularly good at detail capture and small target detection, especially when dealing with imbalanced data. Besides, in the presence of noise, the DICE loss enhances spatial consistency in the target region, thereby improving robustness against local noise interference. Meanwhile, the BCE Loss refines edge details, compensating for the potential neglect of fine structures by DICE loss. The joint optimization of these two losses enhances the model's ability to maintain segmentation performance under noisy conditions. In general, loss function is computed as follows:

$$L = \alpha \cdot \text{BCE Loss} + \beta \cdot \text{DICE Loss}, \tag{5}$$

where $\alpha + \beta = 1$ with α and β representing the respective weights assigned to each component.

2.5. Training hyperparameters and evaluation metrics

The training is performed on the GeForce RTX 3070Ti GPU using the PyTorch framework. The Adam algorithm is

adopted for optimization. Moreover, in order to use the optimized convolutional layers as the illumination patterns, the bias term of all convolutional layers in the encoder is set to 0. Through multiple experimental tests, the learning rate for each epoch is fixed at 0.0001, and the batch size is determined as 250. The dataset is divided into training, validation, and test sets with a ratio of 7:2:1.

Besides, in order to quantitatively analyze the effect of image segmentation, mean intersection over union (MIoU) and Dice coefficient (DICE) are used in the following discussion, which are defined by the following equation:

$$\mathrm{MIoU} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{IoU}_i, \tag{6}$$

$$\mathrm{DICE} = \frac{2\cdot|A\cap B|}{|A|+|B|} = \frac{2\cdot\mathrm{TP}}{2\cdot\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}, \tag{7}$$

with $\mathrm{IoU} = \frac{|A\cap B|}{|A\cup B|} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$. A and B denote the predicted and real segmentation region, respectively. N represents the total number of categories. For a certain pixel category k, where TP, TN, FP, FN are the number of pixels that are correctly predicted as category k, correctly predicted as other categories, wrongly predicted as category k, and wrongly predicted as other categories, respectively.

3. Results

3.1. Simulation

Figure 4 documents the segmentation outcomes of overlapping character images at varying sampling ratios 25%, 5%, and 1%, respectively. Five representative images are selected as subjects. The resolution of each image is 32×32. Each image comprises two characters randomly chosen from the corresponding dataset. The images in the ground truth column are the visualizations of the original dataset labels. We apply different colors to mark distinct categories of objects. Even in instances where certain pixels belong to multiple categories, rendering them indiscernible as to which category they belong, one can distinctly observe that the pixels in the original image are accurately classified. The primary target object within the image can be clearly identified. When the sampling ratio drops to a level that cannot support the segmentation of accurate pixels, the segmentation effect of the edge of the target object is gradually weakened, yet the principal contours of the characters remain qualitatively preserved.

Furthermore, to validate the feasibility of the method in more complex scenarios, we consider incorporating regional occlusion on top of overlapping. In the original image, occlusion blocks of varying sizes are introduced, partially or entirely obscuring the characters to different extents, rendering some characters completely indiscernible to the naked eye. The results recorded in Fig. 5 demonstrate that the effective segmentation can still be achieved even in the presence of partial random occlusion, as illustrated by the complete coverage of a character in the last row of images. When the sampling ratio is reduced to 1%, the segmentation performance fluctuates depending on the simplicity of the image.

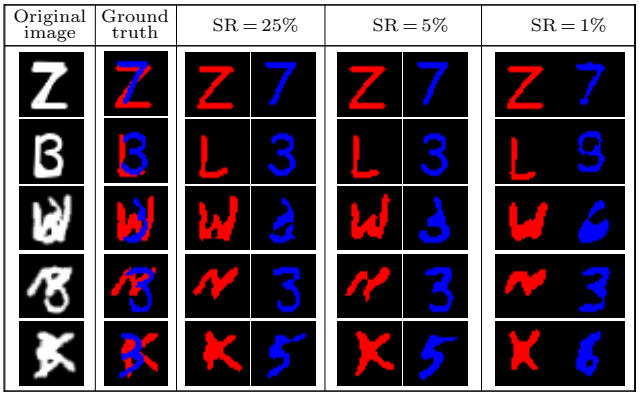


Fig. 4. Simulation results of overlapping scenes at different sampling ratios. The ground truth displays the visualized label images, and each column in the results section represents the visualized segmentation results for the corresponding category.

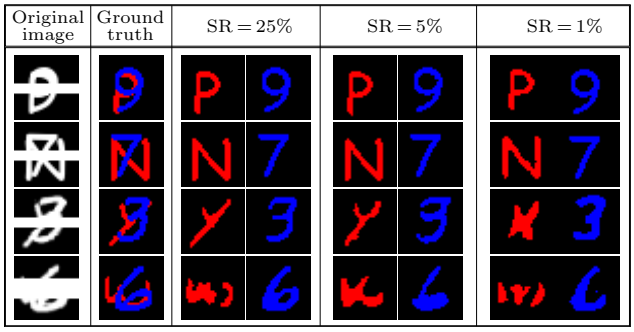


Fig. 5. Simulation results of additional occlusion scenarios.

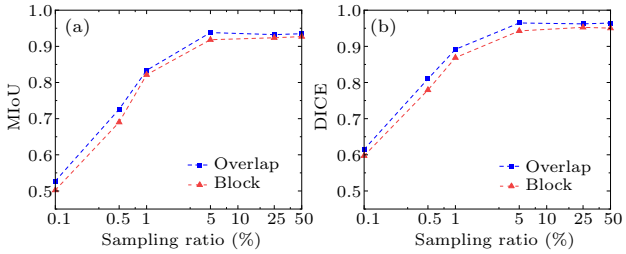


Fig. 6. Segmentation performance curves for both scenarios (overlap and occlusion) at different sampling ratios.

The above result images present the feasibility and excellent performance of our method, but a few examples still have contingency. Therefore, to quantitatively investigate the segmentation performance across different sampling ratios, an additional 200 images not included in the previous dataset serve as a benchmark for comprehensively and scientifically evaluating the segmentation performance. Building upon the three previously mentioned sampling ratios, we also explore the performance at higher and lower sampling ratios. Figures 6(a) and 6(b) present the segmentation performance metrics at six sampling ratios. It is evident that the performance in the presence

of occlusion is inferior to that with mere overlap at each sampling ratio. When the sampling ratio is greater than or equal to 5%, the quality of segmentation can be maintained at an exceptional level, with MIoU greater than 90% and DICE close to 95%.

There is no guarantee that better segmentation results can be obtained with an increase in the sampling ratio. From the perspective of the neural network, an excessive number of convolutional layers make the network redundant, because image features do not require an extensive number of convolutional layers for extraction. Conversely, when the sampling ratio is too low, such as at a sampling ratio of 1%, the number of measurements is reduced to single digits due to resolution constraints, leading to the loss of edge details in the target object. Due to the excessively large size and relatively small number of convolutional kernels in the encoder, the convolutional layer's capacity to extract features becomes inadequate, leading to a deterioration in overall performance and a gradual blurring of the edges in image segmentation. However, overall, from the perspective of SPI, the method can acquire sufficient spatial information at a relatively low sampling ratio to estimate the specific location details of the image, which illustrates its superiority relative to traditional SPI.

Table 1. Quantitative performance comparison with array-net at two sampling ratios.

Sampling ratio	Metrics	Array-net	**Ours**
0.5%	MIoU	70.44%	**71.30%**
	DICE	79.28%	**81.83%**
1%	MIoU	83.18%	**83.79%**
	DICE	88.35%	**89.16%**

Additionally, to highlight the advantages of our method, we consider a performance comparison with the segmentation component of array-net under the identical overlap and occlusion datasets. According to Ref. [25], array-net initially simulates the imaging process at the physical layer, roughly reconstructing the image through correlation calculations, and then achieves segmentation via its segmentation network. We select two representative sampling ratios for comparison, and the comparative results are displayed in Table 1. The numerical results indicate that our network's segmentation efficacy surpasses that of array-net at both sampling ratios. In practice, aside from the segmentation performance comparison, our method is more time-efficient, as it eliminates the need for imaging simulation, significantly reducing computational burden and additional time consumption.

In simulation, to demonstrate the effectiveness of the feature information restoration module in the multi-scale U-Net, we compare the performance of different feature information restoration modules at a representative sampling ratio. Table 2 presents the segmentation performance under various backbones, with the segmentation module being the same U-Net throughout. The numerical results indicate that our network module is capable of better recovering the compressed feature information, thereby enhancing the final network's segmentation performance.

Table 2. Quantitative performance comparison with different backbones at a 0.5% sampling ratio.

Backbone	MIoU	DICE
–	70.79%	79.85%
Resnet50	71.25%	80.21%
PPM	70.30%	79.57%
FSRCNN[30]	70.26%	79.01%
Ours (Resnet50+PPM)	**71.30%**	**81.83%**

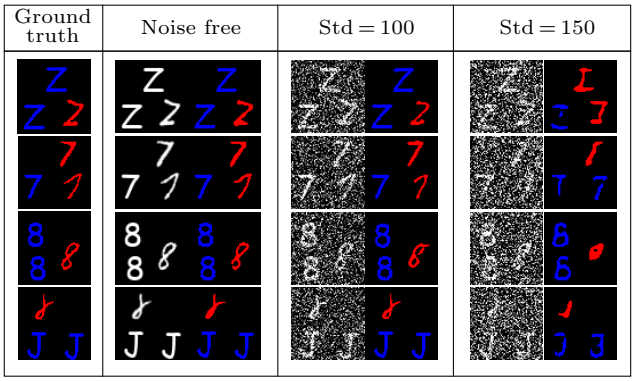


Fig. 7. Simulation results of similar object scenarios. The ground truth displays the visualized label images, and each column in the results section shows the target scene on the left and the segmentation result on the right.

Furthermore, the segmentation of images containing similar targets has also been investigated. We consider combining identical characters from the printed dataset and the EMINST dataset within the same image, with their positions and sizes being variable. The resolution of each image is 64×64. The aim is to accurately distinguish between these two distinct types of the same character. The results depicted in Fig. 7 demonstrate that efficient segmentation is effortlessly achieved. Moreover, to differentiate from traditional threshold segmentation methods and to demonstrate the superiority and noise resistance of our approach, Gaussian noise with a mean of 0 and a standard deviation of Std is introduced to the target scene, and the corresponding results are displayed in Fig. 7. Compared to passively allowing the segmentation performance of the network to be compromised by noise, our proposed method proactively incorporates noise into the training process. This approach enables the network to actively learn the characteristics of noise distribution and intrinsically optimize the joint mechanism between feature extraction and noise suppression. Consequently, the network achieves robust target identification and segmentation across varying noise levels. It was shown that even when different objects in the target scene become indistinguishable to the naked eye or by pixel value distribution, our method can differentiate between two types of distinct characters within a certain range of noise.

Figure 8 quantitatively records the variation in segmentation performance as noise levels change. As the Std of the noise progressively increases, the objects within the target scene become increasingly difficult to discern, leading to segmentation performance diminishes. Nonetheless, our method exhibits robustness within a certain range of noise, ensuring a relatively high-quality segmentation effect.

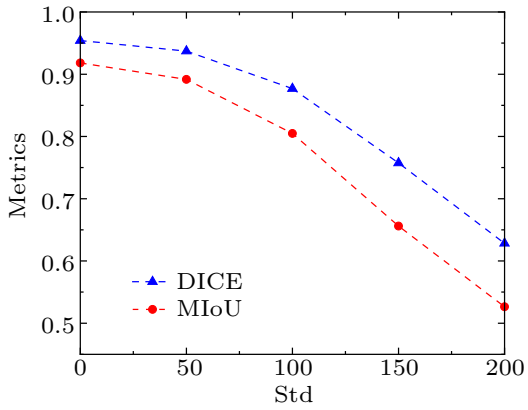


Fig. 8. The curve of segmentation performance as a function of noise intensity.

3.2. Experiment

To prove the practical feasibility and verify segmentation performance, a series of actual SPI experiments are conducted. The experimental device is configured as shown in Fig. 1. The imaging objects are chosen randomly from the image in the testing set. The objects are printed on the disk with a size of 50 mm×50 mm to adapt to the illumination area and pattern size. The modulated light is projected onto a transmission disk printed with the target scene, and the transmitted light is collected by a single-pixel detector after being focused by the lens. The gain of the detector is set to 60 dB.

It is noteworthy that we do not restrict the network-optimized convolutional layers during the network training process, resulting in the possibility of negative and floating-point values in its kernel. However, the experimental DMD can only generate non-negative binary patterns. Therefore,a binarization method is considered used,[31] which is divided into three steps.

To enhance the gray kernel resolution for improved detail retention during subsequent binarization, we first apply bicubic interpolation to upsample the original convolution kernel.[32] This algorithm generates smooth transitions through local polynomial function fitting, effectively avoiding artifacts like sawtooth edges or information loss that typically occur with direct scaling methods. Following this upsampling process, we implement the Floyd–Steinberg dithering algorithm on the processed patterns,[33] which strategically distributes quantization errors to neighboring pixels, thereby achieving global grayscale simulation through binary pattern representation.

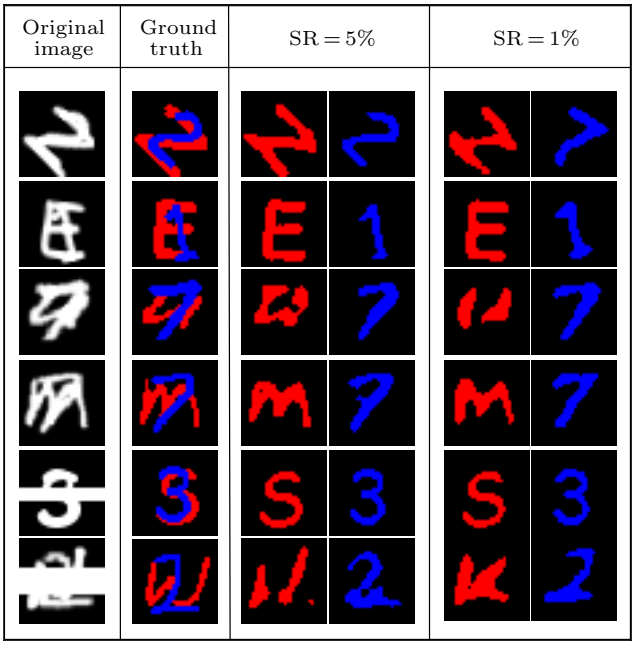


Fig. 9. Experimental results of overlapping and occluded scene at different sampling ratios.

And then we consider using positive and negative pattern differentiation: each kernel pattern is divided into positive and negative components based on its values, with each component containing only the corresponding values from the original pattern. The positive component contains only the positive values, with negative positions set to 0, and vice versa for the negative component. This ensures that the values of the resulting patterns are non-negative. By using a double number of binary patterns on the DMD, a single-pixel detector will receive double light-intensity values. Furthermore, the difference of the corresponding single-pixel measurement serves as the effective measurement.

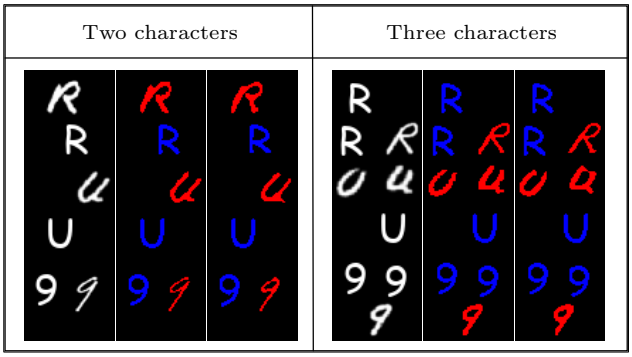


Fig. 10. Experimental results of similar objects. Each section shows the target scene on the left, the visualized label image in the middle, and the segmentation result on the right.

The experimental results of overlap and occlusion are illustrated in Fig. 9. Even in the presence of rotation and scaling within the images, segmentation is accurately achieved. As the sampling ratio gradually decreases to 1%, the precision of segmentation at the finer details of the original characters diminishes, yet the initial contours remain accurately segmented. In practice, when the number of samples is reduced to single digits, the segmentation can only recover the main body of the

target, with a significant decline in the segmentation of edge details. Regarding the segmentation results of similar objects at a 10% sampling ratio in Fig. 10, the results indicate that the network is capable of efficient segmentation even on untrained dual-character images, suggesting that the model has successfully captured the key features of individual objects, and that feature learning is effective.

4. Conclusion

In summary, an image-free segmentation method for complex scenes based on single-pixel imaging (SPI) has been proposed. By utilizing a multi-scale U-Net architecture, the network-optimized patterns are applied for structured illumination, allowing for the automatic capture of object-specific details without relying on postprocessing techniques for imaged objects. The trained network's decoder restores and decodes the light intensity vector carrying crucial features, enabling image-free segmentation. Both simulation results and real-world experiments demonstrate the feasibility of our method across various sampling ratios.

Several advantages of our method can be emphasized. First, in the presence of overlap and occlusion, where pixels may belong to multiple categories, our approach excels in effectively segmenting the pixels corresponding to multiple objects, surpassing other methods. The optimized network facilitates the aggregation of spatial features, enhancing segmentation accuracy. Second, by achieving direct end-to-end segmentation, we can skip the step of obtaining image and directly achieving segmentation only through the spatial information contained in the single-pixel measurements. This not only overcomes the limitations posed by poor imaging quality in low-sampling scenarios but also eliminates the complexities of traditional postprocessing, ensuring efficient segmentation of the target object. Furthermore, unlike the traditional camera using a 2D array sensor which is costly and difficult to produce in specific bands, SPI offers significant advantages, especially for imaging in specific bands such as infrared and terahertz frequencies, where traditional methods face challenges. Thus, the image-free segmentation method has natural advantages in the field of national defense and medical imaging.

Acknowledgement

Project supported by the Fundamental Research Funds for the Central Universities of China (Grant No. 531118010757).

References

[1] Saleh K, Szénási S and Vámossy Z 2021 *IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)* 000477
[2] Zaitoun N M and Aqel M J 2015 *Procedia Comput. Sci.* **65** 797
[3] 1978 *IEEE Trans. Syst. Man Cybern.* **8** 630
[4] Canny J 1986 *IEEE Trans. Pattern Anal. Mach. Intell. PAMI-* **8** 679
[5] Adams R and Bischof L 1994 *IEEE Trans. Pattern Anal. Mach. Intell.* **16** 641
[6] Ronneberger O, Fischer P and Brox T 2015 *Medical image computing and computer-assisted intervention, MICCAI* 2015*: 18th international conference*, Munich, Germany, October 5–9, 2015, proceedings, part III 18, pp. 234–241
[7] Badrinarayanan V, Kendall A and Cipolla R 2017 *IEEE Trans. Pattern Anal. Mach. Intell.* **39** 2481
[8] He K, Zhang X, Ren S and Sun J 2016 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778
[9] He K, Gkioxari G, Dollár P and Girshick R 2017 *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969
[10] Zhou Y, Guo S X, Zhong F and Zhang T 2019 *Chin. Phys. B* **28** 84204
[11] Hou H Y, Zhao Y N, Han J C, Cao D Z, Zhang S H, Liu H C and Liang B L 2023 *Chin. Phys. B* **32** 064201
[12] Bie S H, Wang C H, Lv R B, Bao Q Q, Fu Q, Meng S Y and Chen X H 2023 *Chin. Phys. B* **32** 128702
[13] Hardy N D and Shapiro J H 2013 *Phys. Rev. A* **87** 023820
[14] Hardy N D and Shapiro J H 2011 *Phys. Rev. A* **84** 063824
[15] Liu J J, Bai Y F, Huang X W, Tan W, Nan S Q, Zou X P F and Fu X Q 2021 *Appl. Sci.-Basel* **11** 4115
[16] Zou X, Huang X, Liu C, Tan W, Bai Y and Fu X 2023 *Opt. Laser Technol.* **167** 109807
[17] Zhou L Y, Huang X W, Fu Q, Zou X P F, Bai Y F and Fu X Q 2021 *Chin. Opt. Lett.* **19** 121101
[18] Pan L, Deng C J, Yu C P, Yue S, Gong W L and Han S S 2021 *Chin. Opt. Lett.* **19** 041103
[19] Lin L X, Cao J, Zhou D, Cui H and Hao Q 2022 *Opt. Express* **30** 11243
[20] Fan Y R, Bai Y F, Fu Q, Zhang R, Zhou L Y, Zhu X H, Zou X P F and Fu X Q 2024 *Opt. Commun.* **566** 130684
[21] Cheng J and Han S S 2004 *Phys. Rev. Lett.* **92** 093903
[22] Liu H C and Zhang S 2017 *Appl. Phys. Lett.* **111** 031110
[23] Wang C L, Mei X D, Pan L, Wang P W, Li W, Gao X, Bo Z W, Chen M L, Gong W L and Han S S 2018 *Remote Sens.* **10** 732
[24] Liu H, Bian L and Zhang J 2023 *Opt. Laser Technol.* **157** 108600
[25] Liu X, Han T, Zhou C, Huang J, Ju M, Xu B and Song L 2023 *Opt. Express* **31** 9945
[26] Gibson G M, Johnson S D and Padgett M J 2020 *Opt. Express* **28** 28190
[27] He Z Q and Dai S S 2024 *AIP Adv.* **14** 045316
[28] Zhao H, Shi J, Qi X, Wang X and Jia J 2017 *Proceedings of the IEEE conference on computer vision and pattern recognition* 2881
[29] Cohen G, Afshar S, Tapson J and Van Schaik A 2017 *International joint conference on neural networks (IJCNN)* 2921
[30] Dong C, Loy C C and Tang X 2016 *Computer Vision, ECCV* 2016*: 14th European Conference*, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II **14** 391
[31] Zhang Z B, Li X, Zheng S J, Yao M H, Zheng G A and Zhong J G 2020 *Opt. Express* **28** 13269
[32] Hou H and Andrews H 1978 *IEEE Trans. Acoust. Speech Signal Process.* **26** 508
[33] Floyd R W 1976 *Proc. Soc. Inf. Disp.* **17** 75