Journal of Systems Engineering and Electronics
Vol. 37, No. 3, June 2026, pp.816 – 825



# Multi-scale optical convolutional neural network for target classification

YU Zijian, LI Lijing, WANG Siyuan, and ZHENG Yue*

School of Instrumentation and Optoelectronic Engineering, Beihang University, Beijing 100191, China

**Abstract:** The physical architecture of optical convolution restricts its capacity to capture multi-scale features from targets, thus impeding the precision of network recognition. In this work, we propose a multi-scale optical convolutional neural network (MS-OCNN), which uses convolution kernels with different resolutions in the convolution layer to extract different scale features, along with attention mechanism and residual structure to analyze features. By separating the training and inference platforms of the network, we facilitate the electronic training of model parameters on a computer and the optical deployment on a system equipped with a spatial light modulator, enabling efficient target classification. The proposed MS-OCNN exhibits 1% to 3% improvement in classification performance on the modified national institute of standards and technology (MNIST) and Fashion-MNIST datasets compared to single-scale optical inference models. Online experimental systems in real-world scenarios have validated the target recognition capabilities of this method, which yielded classification accuracies of 97% and 87% on the MNIST and Fashion-MNIST datasets, respectively. This work enhances the feature acquisition capabilities of optical convolutional networks, elevates network recognition accuracy, and significantly propels the application of optical computing in domains such as guidance systems, autonomous driving, and robotics.

**Keywords:** optical computing, multi-scale features, target classification, guidance system.

**DOI:** 10.23919/JSEE.2026.000059

## 1. Introduction

As the fields of computer vision and artificial intelligence have burgeoned, there has been a marked evolution in smart technologies. Convolutional neural networks (CNNs), emblematic of this progress, facilitate advanced perception tasks tailored to diverse environmental contexts. These include object identification, detection, and semantic segmentation [1], with extensive applications in autonomous vehicles, robotics, and military defense [2–4]. The paradigm of distilling prior knowledge from data to predict the outcomes for unknown datasets has propelled CNNs to efficiently tackle perception challenges across a broad spectrum of scenarios. Moreover, this approach has spurred the continual refinement of CNNs for enhanced recognition in complex settings [5].

In pursuit of augmented neural network efficacy, a plethora of studies has concentrated on scaling up network architectures and leveraging high-performance computational hardware to manage the computational demands of extensive datasets [6]. The visual geometry group (VGG) network, introduced in 2014, corroborated a direct relationship between network depth and model efficacy, which points out that deeper networks markedly contribute to improving performance metrics [7]. This insight has been perpetuated and expanded in subsequent research, culminating in the development of prominent and high-performing large-scale models, such as YOLO and Transformer [8–10]. Despite the proficiency of these models in executing recognition tasks, the expansion of network size has entailed a proportional increase in computational requirements for inference. This development poses significant challenges for edge devices, which demands greater computational resources and performance, thereby circumscribing the deployment of intelligent technologies in contexts that prioritize real-time processing or are computationally constrained [11–13].

Optical computing is a method of performing numerical calculations using the propagation of light fields. Compared to traditional electronic computing, optical computing offers advantages such as high speed, high bandwidth, and low power consumption [14]. Implementing neural network components optically can accelerate the inference process. Moreover, transitioning the compu-

Manuscript received April 15, 2024.
*Corresponding author.
This work was supported by the National Natural Science Foundation of China (62201025), the Fundamental Research Funds for the Central Universities (YWF-23-L-1225), and the Chinese Aeronautical Establishment (2022Z037051001).

tational platform from dielectric to optical media reduces the workload of conventional computing devices, thereby decreasing the overall computational power requirements.

Lin et al. [15] established a comprehensive optical deep learning framework predicated on multi-layer diffractive surfaces, which capitalized on the diffraction phenomena of light fields during propagation to facilitate convolutional computations and inferential processes. In a distinct approach, Li et al. [16] utilized a two-dimensional array of passive pixels to create diffractive layers that encoded spatial characteristics through multi-spectral diffraction, thereby achieving optical classification. Jiao et al. [17] innovatively constructed an optical machine learning framework within a single-pixel imaging configuration, employing spatial light modulators (SLM) to execute convolutional operations between targets and kernels, thus enabling inference under incoherent lighting conditions. These studies tend to deploy models within complete perception systems, allowing the signal sensing process to directly participate in model inference. Compared to neural networks computed on traditional electronic platforms, this approach offers higher processing speed and computational efficiency, positioning it at the forefront of emerging computational technologies. However, this novel computational method faces certain limitations in simulating the convolution process for feature extraction, which constrains further enhancement of network performance.

Traditional digital CNNs perceive features at different scales using small-sized convolutional kernels, pooling, and increased network depth [18–20]. However, in optical CNNs (OCNNs) utilizing optical computing, implementing convolution through optical modulation and similar techniques makes it challenging to simulate the sliding process of convolutional kernels. Additionally, it is difficult to capture multi-scale (MS) features of the target by increasing the number of layers. Of the above methods, MS convolution has emerged as the most effective method for improving efficiency, particularly in the implementation of incoherent optical convolution. This approach employs convolutional kernels of varying sizes to perform convolutional operations on input data, thereby capturing feature information at different scales and effectively enhancing the accuracy of decision-making.

Accordingly, this work introduces a MS optical convolution method tailored for the optical computing process within neural networks, addressing the limitations of single-scale (SS) feature extraction. We have designed and trained an optical convolutional inference model, which significantly improves recognition accuracy. By establishing a complete network deployment, we have constructed an online classification system that validates the feasibility and effectiveness of this model in the object classification of real-world entities.

## 2. Methods

### 2.1 Multiscale optical convolution

Deep neural networks have demonstrated exceptional performance in the extraction of image information, with CNNs commonly employed for the extraction and compression of feature dimensions [21]. Convolution, the process by which a kernel slides across the input plane to perform computations, is significantly influenced by the size of the kernel and the stride setting, both of which affect the efficiency of target feature extraction. The computational process of a two-dimensional convolutional layer is as follows:

$$\mathrm{out}(C_{\mathrm{out}}) = b(C_{\mathrm{out}}) + \sum_{k=0}^{C_{\mathrm{in}}-1} w(C_{\mathrm{out}}, k) * \mathrm{input}(k) \tag{1}$$

where $C_{\mathrm{in}}$ and $C_{\mathrm{out}}$ correspond to the quantities of input and output channels, respectively. out(·), which can be derived by the convolution (∗) of weights $w(\cdot)$ and input image input(·) and then adding the bias term $b(\cdot)$, signifies the output features. The process essentially constitutes a cross-correlation operation. When the bias term is nullified and the convolution is executed optically, (1) is transformed to be

$$\mathrm{out}(C_{\mathrm{out}}) = w(C_{\mathrm{out}}) * \mathrm{input}(k). \tag{2}$$

The input to the simplified convolutional layer can be considered as a monochrome image with a single channel, and the computation is equivalent to the inner product of the illumination and detection process, which means that the convolution process can be implemented by the optical manner to extract target feature. By comparing (1) and (2), optical convolution can be perceived as an unbiased, single-channel convolutional layer. In this process, the kernel is distributed over the input optical image to execute a template matching operation, of which the result is collected by a detector and can be represented as

$$F = O_{(x,y)} * K_{(x,y)} = \sum_{x,y} O_{(x,y)} \cdot K_{(x,y)} \tag{3}$$

where $F$ denotes the collected feature intensity values, $O_{(x,y)}$ represents the object's grayscale distribution at the pixel location $(x,y)$, and $K$ refers to the kernel distribution that is projected.

Leveraging the principles of optical convolution, a hybrid optoelectronic neural network framework can be implemented as depicted in Fig. 1. The network architecture comprises an input layer, convolutional layers, and a

decoding classification network. During the training process, the input layer receives the target images, while the output layer provides the predicted class labels. The convolutional and decoding networks are jointly optimized to extract target features and classify the targets accurately. The kernels of the convolutional layer, as the components of network optimization, can be mapped into the optical inference process. The inference process of the network involves a programmable SLM (in Fig. 1), that modulates the light field according to the kernel distribution, and a detector completing the feature collection of the actual target. Finally, the decision network is deployed on a platform with limited computing power to implement feature inference.

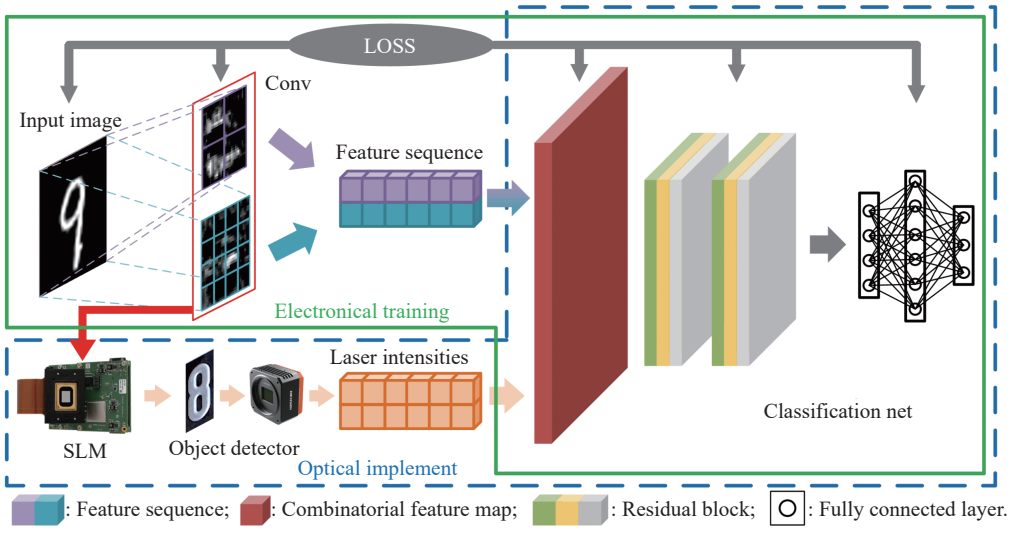



**Fig. 1 Optoelectronic hybrid neural network framework**

The process of implementing optical convolution through the encoding of light fields is equivalent to single-kernel full convolution, which does not facilitate the sliding of the kernel to capture the coupled feature information between the pixels at different spatial locations. The optical convolution process necessitates an incoherent optical system for the physical mapping. The compressive nature of convolution precludes the ability to achieve deep convolution through stacking processes, leading to weaker feature extraction capability of optical convolution compared to fully optical diffraction neural networks. Consequently, optical convolution has inherent limitations in feature extraction, impeding the extraction of more information from complex scenes.

The importance of analyzing targets at multiple scales stems from the nature of the targets themselves. Scenes in the world contain objects of various sizes, and objects themselves have features at different scales. Applying any analytical process at a single scale may result in the loss of information at other scales. A better solution is to perform analysis simultaneously at multiple scales [22,23]. Modern image classification networks aggregate MS features using various methods, such as integrating information through consecutive convolutional and pooling layers or performing convolutions with kernels of different sizes. The specificity of optical convolution makes it challenging to achieve the former, as it often requires the additional creation of non-programmable diffractive media to construct the optical network. For the latter, the physical mapping method of the projected coded light field allows optical convolution to freely adjust kernel sizes and achieve light field illumination to extract features at different scales without facing computational issues associated with MS convolution. This inspired us to construct the MS optical convolution module shown in Fig. 2.

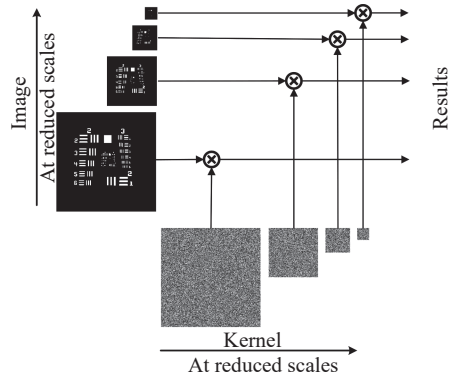



**Fig. 2 Schematic diagram of the proposed multiscale optical convolutional module**

By progressively downsampling to generate images of different resolutions, an image pyramid is formed [24]. Optical convolution is then performed with kernels of the same size as each respective image, outputting features at different scales. Smaller convolutional kernels, corresponding to lower resolution, are capable of capturing detailed features, whereas larger convolutional kernels can capture the global features of the target. The integration of information from different scales enhances the recognition capability of the backend network. The principle of extracting MS features of the target using this method can be expressed as

$$\begin{cases} F_{m\times m} = O^{m\times m} * K^{m\times m} = \sum\limits_{i=0,j=0}^{i=m,j=m} O_{(i,j)} \cdot K_{(i,j)} \\ \vdots \\ F_{n\times n} = O^{n\times n} * K^{n\times n} = \sum\limits_{i=0,j=0}^{i=n,j=n} O_{(i,j)} \cdot K_{(i,j)} \end{cases} \tag{4}$$

where $F_{m\times m}$ ($F_{n\times n}$) denotes the image features collected by optical convolution at a scale of $m\times m$ ($n\times n$), $O_{(i,j)}$ represents the object's grayscale distribution at the pixel location $(i, j)$, and $K^{m\times m}$ ($K^{n\times n}$) refers to the kernel distribution that is projected at an $m\times m$ ($n\times n$) resolution.

Furthermore, unlike traditional convolutional layers, optical convolution kernels of different sizes possess the same semantic level. This means that during inference, all convolution kernels can directly illuminate the target objects based on the trained distribution without the need for additional serial arrangement of more optical systems, facilitating the simplification and integration of hardware systems.

## 2.2 Image-free neural network for target classification

Traditional target recognition neural networks rely on sensor-captured target images for inference and classification. In contrast, OCNNs based on spatial light modulation achieve target feature perception through light field propagation and photoelectric detection, which are then used for decision-making on target categories. Unlike traditional image-based neural networks, OCNNs do not generate image matrices. Instead, they directly utilize the sequence of light intensity values captured by detectors as target features for numerical computation. This approach significantly reduces the complexity and computational load of the network, increases computation speed, and lowers the deployment difficulty of neural networks on resource-constrained devices.

Therefore, introducing MS optical convolution in non-imaging target recognition networks can significantly enhance the model's recognition capabilities and contribute to the advancement of non-imaging target recognition applications.

Fig. 3 illustrates the schematic of the designed MS optical convolution non-imaging target recognition training framework. To simulate the MS characteristics of the target, we use the process of hierarchical downsampling of the original image to multiple resolutions as the generative model to generate MS dataset from the original images in the dataset. The training batch consisting of the generated images is fed into the MS-OCNN object classification model. We build parallel multi-resolution convolution layers to optimize the distribution of optical convolution kernels. What is more, we combine channel attention mechanisms and residual blocks to improve the performance of the model.

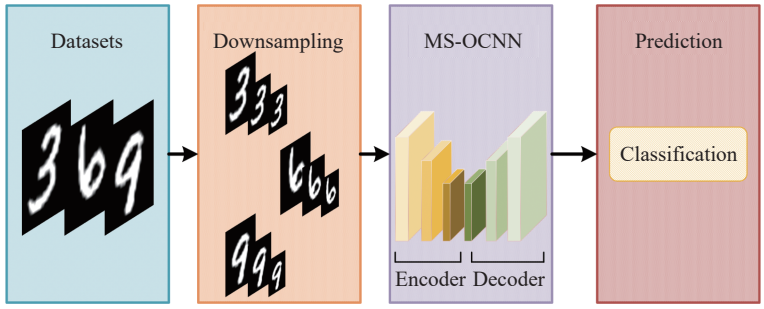



**Fig. 3 Training framework for MS optical convolutional non-imaging neural network**

The parallel gradient-dimension convolutional structure ensures that the convolutional kernel in each channel operates at an equivalent depth, facilitating the real-time optimization of optical convolution kernels. In consistency with the execution modality of optical convolution, we implement the extraction of target features through full convolution. For the inputs with varying resolutions, the convolutional kernels matching the respective sizes are endowed with their channel resources, achieving the extraction of MS features. In this setup, the number of channels within the convolutional kernels corresponds to the number of modulations performed on the light field.

Furthermore, specialized fusion blocks are engineered to cater to feature information at different hierarchical levels, maintaining the distinctiveness among the features of various scales.

In the feature inference module, residual blocks and attention mechanisms are utilized for feature identification. As shown in Fig. 4(a), features from various semantic levels are reshaped to form feature maps. Once multiple-channel feature maps are established, the channel attention module (CAM) is integrated to derive attention vectors for the respective channels. These vectors are combined with the original channel distributions through weighting, resulting in weighted feature maps [25]. The introduction of the attention mechanism also facilitates the prioritization of modulation patterns based on the participatory weights of the convolutional kernels in inference, thereby identifying the more beneficial kernels for inference and discarding those with minimal influence, which can further reduce the sampling rate by retaining these kernels [26]. Furthermore, for computations involving deep features, residual blocks are implemented for the processing and transmission of deep features [27]. Fig. 4(b) illustrates the residual architecture, which consists of convolutional layers (conv in Fig. 4) and rectified linear unit (ReLU). This architecture is introduced to mitigate the excessive consumption of computational resources and the problem of vanishing gradients, as well as for effective feature learning and signal propagation, while significantly reducing the network's parameter count. Focusing on the lightness and efficiency throughout the target inference process, we have crafted multi-resolution convolutional layers and embedded channel attention mechanisms. The complete network architecture is delineated in Fig. 4(c) with fully connected layers (fc in Fig. 4) inserted. Within the MS-OCNN that accommodates the inputs of various resolutions, we have realized the extraction, integration, and inference of MS features from the target, culminating in the use of fully connected layers to yield the prediction outcomes.

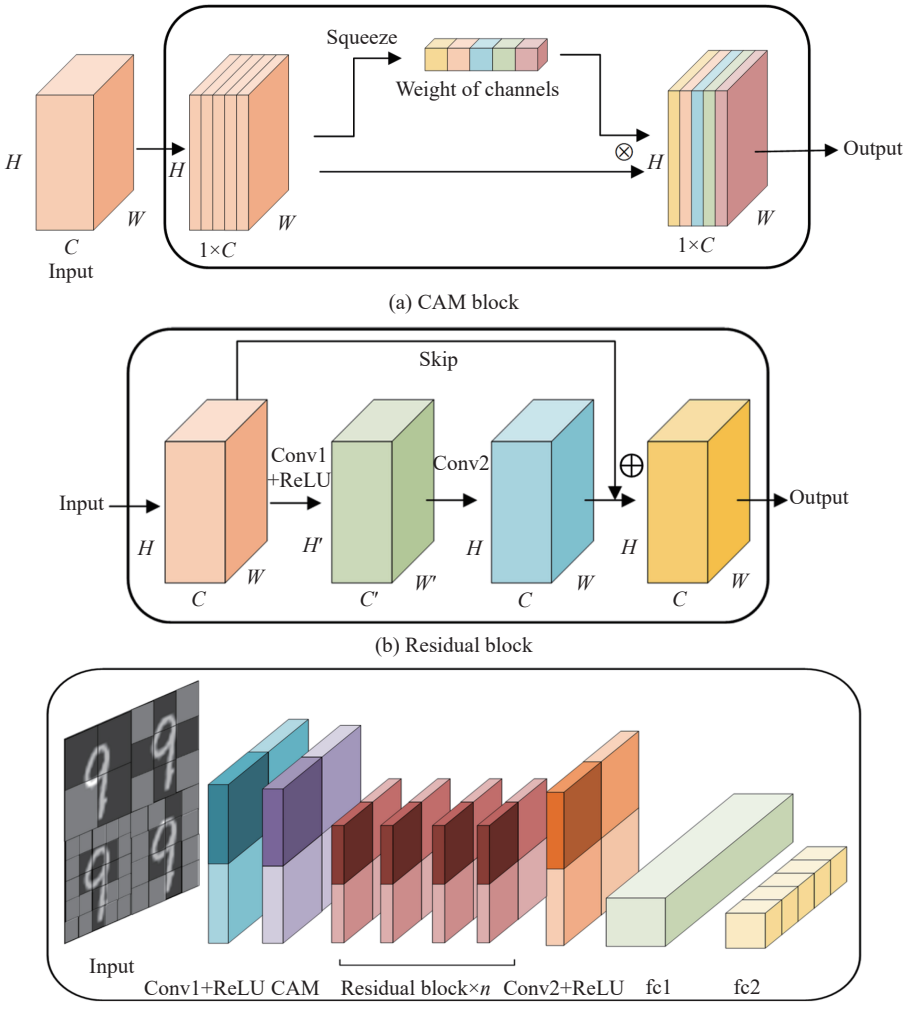



**Fig. 4 Model structure diagram**

## 3. Simulation and experiment

### 3.1 Simulation

We train and test the model on the PyTorch platform. To assess the model's generalization capabilities, we select the modified national institute of standards and technology (MNIST) handwritten digit dataset and the Fashion-MNIST apparel dataset, which feature grayscale details, as our training data inputs. The MNIST dataset encompasses digits '0' to '9' and consists of simple binary images, whereas the Fashion-MNIST dataset comprises images of 10 popular types of clothing, presenting sparser grayscale images [28]. We generate multi-resolution datasets from both datasets through hierarchical downsampling. Each dataset is composed of a training set, a validation set and a test set, while maintaining the original class configurations. The higher-resolution images contain more detailed information among images of varying resolutions, while the lower-resolution images exhibit more pronounced contour features. For full convolution, lower-resolution input images provide intuitive contour information, which is less susceptible to the interference from the distribution of details within the image, such as texture and noise. High-resolution images, on the other hand, are employed to supplement fundamental detail information. Compared to the MNIST dataset, the Fashion-MNIST dataset is more complex, with a certain degree of grayscale distribution, making it more suitable for testing the robustness and generalization capabilities of the proposed model, especially when subjected to added noise.

To independently verify the effectiveness of the designed MS optical convolutional layer, we employ a fully connected neural network (FCNN) as the baseline and construct fully connected networks with both SS and MS optical convolutional structures. We align the number of convolutional kernel channels for the SS recognition model with that of the MS recognition model, using the number of convolutional kernel channels as the variable for recognition simulation. The results are displayed in Fig. 5. As the number of convolutional kernel channels increases, the recognition accuracy for targets is improved within a certain range and then becomes nearly stable. Compared to the SS FCNN model, the introduction of the MS optical convolutional structure endows the network with a higher ceiling of recognition accuracy. With direct decision-making on input features, the accuracy is increased by nearly 1%. This indicates that the MS optical convolutional structure can effectively augment the quantity of features fed into the network which are derived from various scales of the images. The enhancing of the volume of valid information drives correct model decisions, which is consistent with our theoretical analysis.

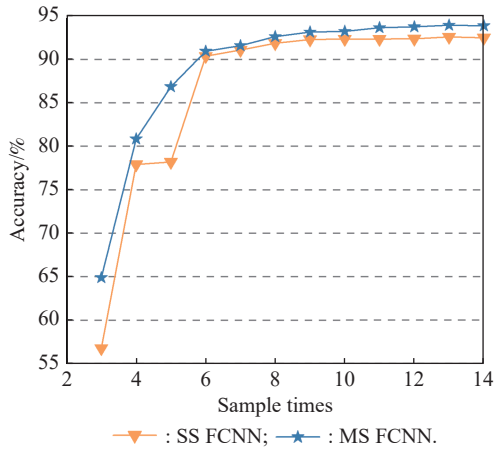



**Fig. 5 Comparison of MS FCNN and SS FCNN**

To substantiate the efficacy of the introduced residual structures and attention mechanisms in enhancing model performance, we compare four models: (i) a multi-scale fully convolutional network MS-FCN classification model constructed with MS optical convolutional layers and fully connected layers; (ii) a MS-FCN classification model incorporating an attention mechanism; (iii) a MS-FCN classification model integrating a residual structure; (iv) the proposed MS-OCNN, a multi-resolution CNN classification model employing both attention mechanisms and residual structures. The experimental platform is configured with an Ubuntu 22.04 operating system, 16 GB random-access-memory (RAM), and an Nvidia MX450 graphics card. All four models are trained under identical training parameters. Training is completed over 100 epochs with a batch size of 256. We employ a stochastic gradient descent (SGD) optimizer with the initial learning rate of 0.002, the momentum set to 0.9, and the weight decay at $1.2\times10^{-4}$. A learning rate scheduler is also utilized, which dynamically reduces the learning rate based on the loss changes of the model on the validation set, combined with an early-stopping mechanism to ensure better convergence and improved training efficiency. Table 1 presents the accuracy metrics of the aforementioned four models on the two datasets. The results indicate that the integration of the attention modules and the residual structures significantly enhances the recognition capabilities of the model. Our proposed MS-OCNN achieves 3.8% and 5.9% higher recognition accuracies on MNIST and Fashion-MNIST datasets compared to the basic multi-resolution CNN classification model, as indicated by the bold values in Table 1.

**Table 1 Ablation experiments** %

| Model | MNIST | Fashion-MNIST |
|---|---|---|
| MS-FCN | 94.86 | 84.74 |
| MS-FCN+CAM | 94.92 | 86.66 |
| MS-FCN+Residual block | 98.07 | 89.66 |
| MS-OCNN (proposed) | **98.67** | **90.67** |

To validate the performance of the proposed MS-OCNN, we compare it with several reported SS non-imaging recognition models. These selected SS models utilize neural network-optimized optical convolutional distributions and single-pixel detection recognition methods. They are trained and fine-tuned on the same dataset to achieve efficient non-imaging classification of specific targets. The classification simulation results with different sampling times (ST) are shown in Tables 2 and 3. In the tables, sample rates (SR) represents the ratio of the actual number of samples to the number of samples required to reconstruct an image using a single-pixel detector, the latter typically being the total number of pixels in the image. With the same number of convolution kernel channels, which corresponds to the same sample times in the optical convolution process, the model with the specifically selected combination of resolutions demonstrated higher recognition accuracy compared to SS models, thus exhibiting superior feature extraction and recognition efficiency. This is validated across recognition tasks for both datasets. However, in the more complex task of recognizing the Fashion-MNIST dataset, hyper-parameters including optical convolution resolution need to be fine-tuned. This discrepancy is primarily attributed to the distinct characteristics of objects within different datasets, necessitating the model to adjust the resolution as a hyperparameter through training to achieve optimal performance. In our simulations, we incorporate Gaussian noise of varying intensities into the test dataset to simulate pixel-level fluctuations characteristic of light source noise. This approach is utilized to evaluate the effect of noise on the accuracy of the recognition process. Fig. 6(a) illustrates the images with superimposed Gaussian noise of different magnitudes. The recognition outcomes are depicted in Fig. 6(b), where the proposed model can resist a certain level of noise interference. However, as the noise intensity increases (when the standard deviation exceeds 0.1), there is a significant drop in recognition accuracy.

**Table 2 Simulation results of the classification accuracy for different models with low sample times**

| Model | ST | SR/% | MNIST/% | Fashion-MNIST/% |
|---|---|---|---|---|
| FCFNN [29] | 8 | 0.78 | 58.17 | 63.50 |
| DICNN [30] | 8 | 1 | 58.94 | – |
| SPSCNN [31] | 8 | 1 | 91.73 | 82.15 |
| MS-OCNN (proposed) | **8** | **0.78** | **91.93** | **84.00** |

**Table 3 Simulation results of the classification accuracy for different models with high sample times**

| Model | ST | SR/% | MNIST/% | Fashion-MNIST/% |
|---|---|---|---|---|
| FCFNN [29] | 64 | 6.25 | 97.25 | 87.46 |
| DICNN [30] | 39 | 5 | 94.82 | – |
| SPSCNN [31] | 40 | 5 | 96.98 | 84.98 |
| MS-OCNN (proposed) | **36** | **3.5** | **98.27** | **90.02** |

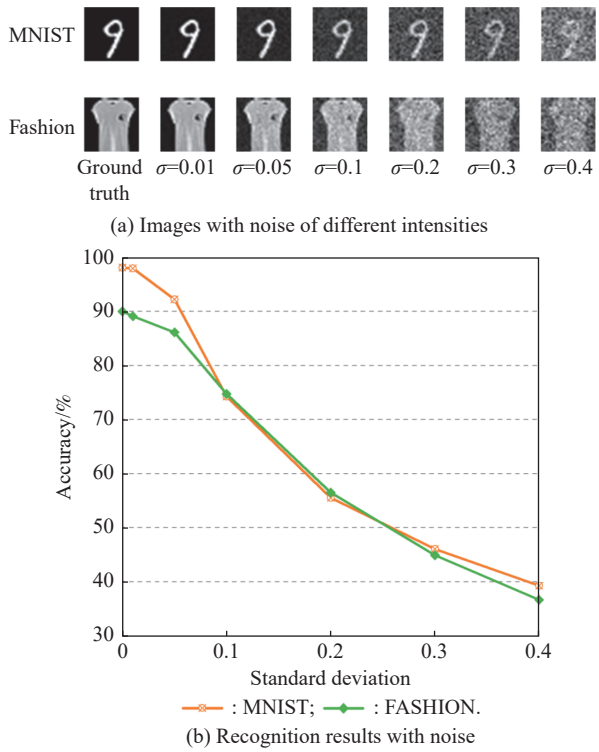



**Fig. 6 Recognition results under different light source noise interference**

## 3.2 Experiment

To validate the practical effectiveness of the proposed method, we construct an experimental setup illustrated in Fig. 7 to realize the optical CNN for target recognition and classification. The emitted continuous-wave laser (Rayziss, 532 nm) beam is expanded by a beam expander to produce a uniform light spot. The projected light is modulated by the pre-imported convolution kernel distribution on a digital micromirror device (DMD, Texas Instruments

Discovery 7001, resolution of 1 024×768, refresh rate of 22 kHz), and the exiting modulated light field is reflected off the target surface. The reflected light, after passing through the collection lens, is gathered by a quadrant detector (First sensor) and converted into point signals. These signals are then amplified by a signal conditioning circuit and acquired by a microprocessor chip with a 12-bit analog-to-digital converter (ADC), converting them into digital signals. The processor performs real-time calculations to produce recognition results. The ADC used in the experiment is configured with four channels to detect the light intensity signals of the four pixels of the quadrant detector. Each channel's conversion time is set to more than 1 μs to ensure sufficient sampling accuracy. Given the SLM's modulation rate of 20 kHz, the detection time required for each recognition is approximately 1.2 ms, providing the system with good real-time performance.

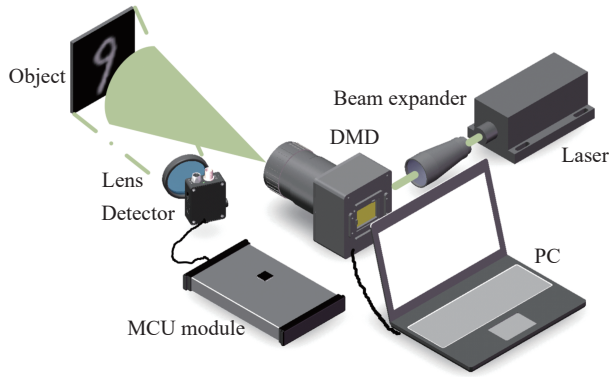
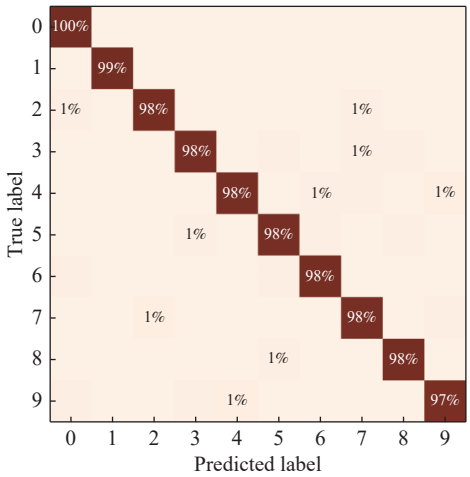


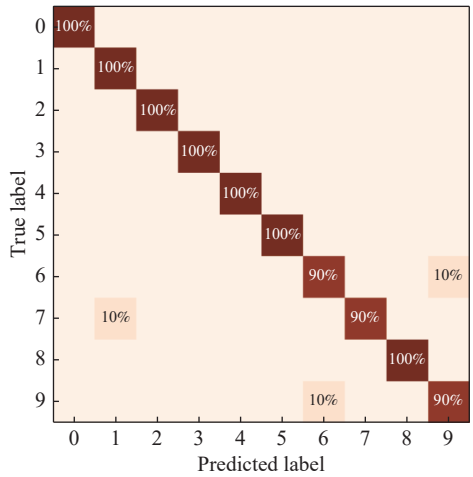

**Fig. 7 Experiment Setup for online non-imaging target recognition with MS features**

In the experiments, to approximate the recognition of real-world objects, we print randomly selected images from the test set on A4 papers. Through printed grayscale calibration, we correct the reflectance of the images. After training the proposed model, we extract the convolutional kernels used for feature extraction. Utilizing a grayscale dithering algorithm, we transform the continuously distributed convolutional kernels into binary discrete modulation patterns for spatial light modulation. Furthermore, we increas the ADC sampling period to 10 ms. After five repeated samplings, we take the mean value as the effective signal for processing, thereby reducing the impact of random fluctuations in the light source on recognition accuracy. Leveraging the effectiveness of the proposed network model, we decouple the decision-making part of the model and deploy it into the microcontroller unit (MCU) after quantization and compression, achieving online target recognition.

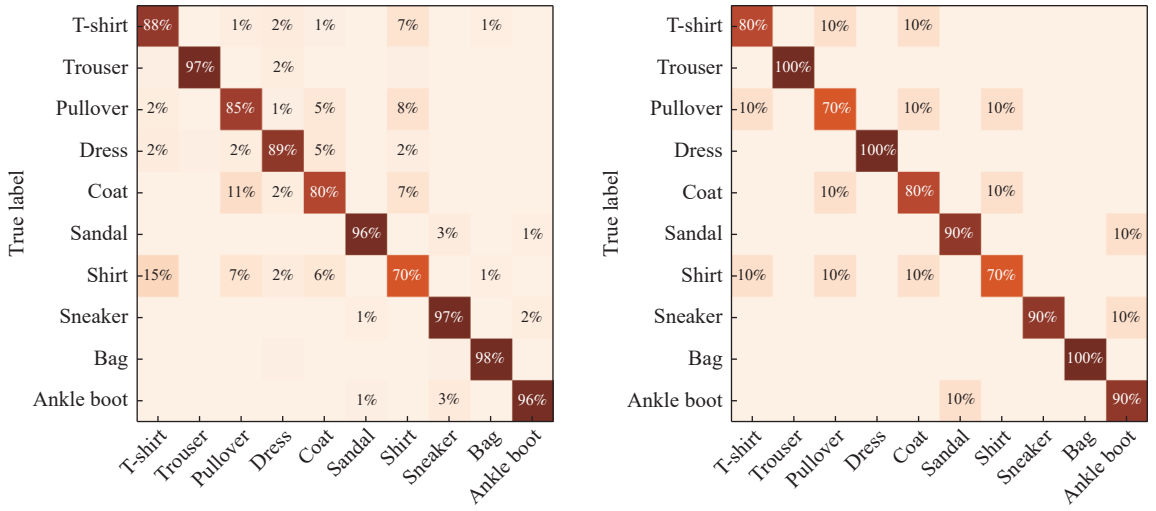
To evaluate the performance of our trained target classification model, we conduct recognition experiments with images from different datasets. Fig. 8 illustrates the recognition results for 100 randomly selected printed images from each dataset in actual experiments. The results indicate that the proposed method achieves the accuracy of 97% in the recognition task of printed MNIST images (Fig. 8(b)) and the accuracy of 87% on the Fashion-MNIST dataset (Fig. 8(d)). Numerically, the MS-OCNN attains classification accuracies of 98.6% over 10 000 handwritten digits in the blind testing set (Fig. 8(a)) and of 90.6% for Fashion-MNIST dataset in the blind testing set (Fig. 8(c)). The experimental classification results demonstrate consistency with the simulation results, despite an approximate 2% loss in accuracy. This reduction is primarily attributed to some interference factors such as the quantization of network parameters in deployment, the deviations in light spot illumination, and the circuit sampling errors. The experiment demonstrates that the proposed model and recognition method can effectively enhance the sampling efficiency of target features and improve the recognition accuracy of targets.




(a) Confusion matrix of numerical classification results over handwritten digits




(b) Confusion matrix of experimental classification results over handwritten digits




(c) Confusion matrix of numerical classification results for Fashion-MNIST dataset

(d) Confusion matrix of experimental classification results for Fashion-MNIST dataset

**Fig. 8 MS image-free experiment results based on MNIST and Fashion-MNIST**

## 4. Conclusions

In summary, we have proposed a MS optical convolution strategy guided by the architecture of OCNNs for non-imaging target classification and have completed the construction of an online recognition system. Through sampling and recognition experiments with real images from various datasets, the results demonstrate that the proposed classification model exhibits commendable performance in practical scenarios. Compared to traditional SS recognition models, the model we proposed achieves enhanced recognition accuracy while maintaining superior efficiency. This work enhances the feature acquisition capabilities of optical convolution networks and elevates network recognition precision, thereby significantly advancing the applications of optical computing in areas such as guidance, autonomous driving, remoting sensing and robotics [32–34].

In this work, a certain degree of discrepancy is observed between the features utilized for model training and directly perceived from the sensor after the kernel projection, which is primarily attributed to the interference or noise introduced in the optical detection process. The inclusion of detector-acquired images in the training dataset can lead to a marked enhancement in the classification performance. Furthermore, the mechanism for setting the MS convolutional resolutions in the model, as well as the execution methodology of the target classification experiments, requires additional refinements. When the scale of the full convolutional kernel, which matches the image size, is set excessively high, it may compromise the robustness and increase the number of parameters. The mechanism for setting resolution levels can be optimized through algorithmic enhancements to approach the optimal configuration during the training process. In the future, we will verify the upper limit of the model's performances under more challenging scenarios. With the upgrading of system hardware, more advanced networks, such as Transformer-based ones, can be applied to provide superior performances.

## References

[1] YU C P, XIONG W, LI X Q, et al. Deep convolutional neural network for meteorology target detection in airborne weather radar images. Journal of Systems Engineering and Electronics, 2023, 34(5): 1147–1157.

[2] WANG S Y, LI L J, YU Z J, et al. Image-free target classification with semiactive laser detection system. IEEE Sensors Journal, 2022, 22(23): 23088–23094.

[3] NIRANJAN D R, VINAYKARTHIK B C, MOHANA. Deep learning based object detection model for autonomous driving research using CARLA simulator. Proc. of the 2nd International Conference on Smart Electronics and Communication, 2021: 1251−1258.

[4] YAO Q H, WANG Y, YANG Y X. Range estimation of few-shot underwater sound source in shallow water based on transfer learning and residual CNN. Journal of Systems Engineering and Electronics, 2023, 34(4): 839–850.

[5] LI Z W, LIU F, YANG W J, et al. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE Trans. on Neural Networks and Learning Systems, 2022, 33(12): 6999–7019.

[6] VOULODIMOS A, DOULAMIS N, DOULAMIS A, et al. Deep learning for computer vision: a brief review. Computational Intelligence and Neuroscience, 2018, 2018(1): 7068349.

[7] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition. https://doi.org/10.48550/arXiv.1409.1556.

[8] REDMON J, DIVVALA S, GIRSHICK R, et al. You only

look once: unified, real-time object detection. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779−788.
[9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need. Advances in Neural Information Processing Systems. https://arxiv.org/abs/1706.03762.
[10] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale. https://doi.org/10.48550/arXiv.2010.11929.
[11] HOWARD A G. MobileNets: efficient convolutional neural networks for mobile vision applications. https://doi.org/10.48550/arXiv.1704.04861.
[12] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2818–2826.
[13] FENG T X, ZHANG S Y, WU T, et al. Entangled photon-pair source using a wedge-shaped nonlinear crystal. Optical Materials, 2023, 145: 114441.
[14] WETZSTEIN G, OZCAN A, GIGAN S, et al. Inference in artificial intelligence with deep optics and photonics. Nature, 2020, 588(7836): 39–47.
[15] LIN X, RIVENSON Y, YARDIMCI N T, et al. All-optical machine learning using diffractive deep neural networks. Science, 2018, 361(6406): 1004–1008.
[16] LI J X, MENGU D, YARDIMCI N T, et al. Spectrally encoded single-pixel machine vision using diffractive networks. Science Advances, 2021, 7(13): eabd7690.
[17] JIAO S M, FENG J, GAO Y, et al. Optical machine learning with incoherent light and a single-pixel detector. Optics Letters, 2019, 44(21): 5186–5189.
[18] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4510–4520.
[19] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector. Proc. of the Computer Vision–ECCV , 2016: 21–37.
[20] ZHAO B J, ZHAO B Y, TANG L B, et al. Multi-scale object detection by top-down and bottom-up feature pyramid network. Journal of Systems Engineering and Electronics, 2019, 30(1): 1–12.
[21] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 2012: 25.
[22] HE K M, ZHANG X Y, REN S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. on Pattern Analysis Machine Intelligence, 2015, 37(9): 1904–1916.
[23] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149.
[24] ADELSON E H, ANDERSON C H, BERGEN J R, et al. Pyramid methods in image processing. RCA Engineer, 1984, 29(6): 33–41.
[25] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132–7141.
[26] ZHAN X R, ZHU C L, SUO J L, et al. Weighted sampling-adaptive single-pixel sensing. Optics Letters, 2022, 47(11): 2838–2841.
[27] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770–778.
[28] DENG L. The MNIST database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine, 2012, 29(6): 141–142.
[29] CAO J N, ZUO Y H, WANG H H, et al. Single-pixel neural network object classification of sub-Nyquist ghost imaging. Applied Optics, 2021, 60(29): 9180–9187.
[30] LOHIT S, KULKARNI K, TURAGA P. Direct inference on compressive measurements using convolutional neural networks. Proc. of the IEEE International Conference on Image Processing, 2016: 1913–1917.
[31] FU H, BIAN L H, ZHANG J. Single-pixel sensing with optimal binarized modulation. Optics Letters, 2020, 45(11): 3111–3114.
[32] ZHU X X, MONTAZERI S, ALI M, et al. Deep learning meets SAR: concepts, models, pitfalls, and perspectives. IEEE Geoscience and Remote Sensing Magazine, 2021, 9(4): 143–172.
[33] LIN Z, JI K F, LENG X G, et al. Squeeze and excitation rank faster R-CNN for ship detection in SAR images. IEEE Geoscience and Remote Sensing Letters, 2019, 16(5): 751–755.
[34] KANG M, JI K F, LENG X G, et al. Contextual region-based convolutional neural network with multilayer fusion for SAR ship detection. Remote Sensing, 2017, 9(8): 860.