Research Article

ADVANCED IMAGING

Ultra-robust imaging restoration of intrinsic deterioration in graded-index imaging systems enabled by classified-cascaded convolutional neural networks

Zaipeng Duan,[a,b,c,†] Yang Yang,[a,b,†] Ruiqi Zhou,[a,b] Jie Ma,[c] Jiong Xiao,[a,b] Zihang Liu,[a,b] Feifei Hao,[a,b] Jinwei Zeng,[a,b,*] and Jian Wang[a,b,*]

[a]Wuhan National Laboratory for Optoelectronics and School of Optical and Electronic Information, Huazhong University of Science and Technology, Wuhan, China

[b]Optics Valley Laboratory, Wuhan, China

[c]National Key Laboratory of Science and Technology on Multispectral Information Processing, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China

Abstract. Endoscopic imaging is crucial for minimally invasive observation of biological tissues. Notably, the integration between the graded-index (GRIN) waveguides and convolutional neural networks (CNNs) has shown promise in enhancing endoscopy quality thanks to their synergistic combination of hardware-based dispersion suppression and software-based imaging restoration. However, conventional CNNs are typically ineffective against diverse intrinsic distortions in real-life imaging systems, limiting their use in rectifying extrinsic distortions. This issue is particularly urgent in wide-spectrum GRIN endoscopes, where the random variation in their equivalent optical lengths leads to catastrophic imaging distortion. To address this problem, we propose a novel network architecture termed the classified-cascaded CNN (CC-CNN), which comprises a virtual-real discrimination network and a physical-aberration correction network, tailored to distinct physical sources under prior knowledge. The CC-CNN, by aligning its processing logic with physical reality, achieves high-fidelity intrinsic distortion correction for GRIN systems, even with limited training data. Our experiment demonstrates that complex distortions from multiple random-length GRIN systems can be effectively restored using a single CC-CNN. This research offers insights into next-generation GRIN-based endoscopic systems and highlights the untapped potential of CC-CNNs designed under the guidance of categorized physical models.

Keywords: imaging transmission; graded-index waveguide imaging; classified-cascaded convolutional neural network.

Received Jun. 15, 2024; revised manuscript received Jul. 29, 2024; accepted Aug. 19, 2024; published online Sep. 19, 2024.

[DOI: 10.3788/AI.2024.10009]

1. Introduction

Observing the intricate interiors of interested organisms is an intellectually stimulating pursuit, especially given its crucial role in enhancing medical diagnostics through *in vivo* imaging[1–3]. To achieve minimally invasive imaging, researchers have proposed endoscopes that integrate optical waveguide imaging and algorithm imaging recovery technologies[4–6]. Since the Second World War, rigid endoscopes, with Hopkins endoscopes as the mainstream representatives, have been predominantly utilized for procedures such as arthroscopy, laparoscopy, and thoracoscopy, involving access to body cavities

*Address all correspondence to Jinwei Zeng, zengjinwei@hust.edu.cn; Jian Wang, jwang@hust.edu.cn

†These authors contributed equally to this work.

through small incisions or natural orifices[7–9]. However, the relative lack of efficient restoration techniques and inherent aberrations from the rod lens jointly result in significant deficiencies in both imaging fidelity and compactness of the Hopkins endoscope[10]. Due to the introduction of multiple reflections and aberration accumulation caused by the structure of several glass rod combinations, these endoscopes often necessitate complex aberration correction systems and delicate mechanical assembly structures, leading to inconveniently larger caliber that might cause excessive risks to the patients. This issue urgently calls for an innovative endoscopic design solution to fabricate the next generation of endoscopes.

The promise emerges from the synergy between the graded-index (GRIN) waveguide, as the hardware engineered for dispersion suppression[11], and the convolutional neural network (CNN), as the software tailored for image restoration[12]. We append the imaging principles of GRIN waveguides in Sec. 1 of Supplementary Material and the analysis of the imaging aberration from GRIN waveguides and the virtual-real problem from structural deterioration in Sec. 2 of Supplementary Material. On the one hand, the low mode dispersion in GRIN waveguides can efficiently mitigate the laboriousness of CNNs. On the other hand, the CNNs can adaptively address complex distortions in GRIN waveguides, facilitating the transmission and restoration of high-resolution, wide-field images. Notably, a previous study has demonstrated the effectiveness of the CNN in restoring GRIN imaging, successfully recovering distorted images from extrinsic deteriorations such as defocusing, tilting, and contaminating among others[13]. However, practical constraints in manufacturing and the need for wide-spectrum light sources in medical devices often require restoring imaging of GRIN waveguides with intrinsic deterioration, as elaborated in Sec. 2.1 of this paper. Here, extrinsic deteriorations stem from optical errors external to the components, while intrinsic deteriorations originate from flaws within the optical components themselves, often unavoidable in broadband operations. Indeed, the length mismatch, as a primary factor, introduces complex coupling aberrations from multiple physical sources[14,15], resulting in nonlinear image degradation, decreased image fidelity, and laborious training efforts, thus serving as a primary constraint delaying the introduction of GRIN endoscopes.

Although CNNs may achieve effective imaging restoration mainly in controlled environments, the complexity of real-life optical systems poses challenges, requiring considerable effort to train CNNs and rendering this approach impractical[16–18]. Thus, most CNN-based image restoration researches are confined to rectifying distortions in laboratory conditions. Controlled through precise interference conditions, the imaging system is mostly limited to fixed imaging systems, like multimode fiber imaging, where aberrations are predominantly extrinsic and tend to remain constant, in which case imaging distortions usually perform as straightforward linear changes. Therefore, conventional CNNs may prove insufficient enough when confronted with complex intrinsic deteriorations such as those found in random-length GRIN imaging systems. Although researchers employ the Universal Approximation Theorem as a theoretical guide to address this limitation, it necessitates extensive datasets and computational power to train cumbersome CNNs, leading to an exponential increase in the demand for training data[19,20]. Since top-tier imaging datasets are extremely difficult to acquire, large-scale image data training becomes an almost unattainable task. Hence, the development of novel compact network architectures tailored for intrinsic deterioration imaging systems becomes a prerequisite for designing high-performance GRIN endoscopes, while also catalyzing the broader application of CNNs in the field of medical imaging.

To address the challenge coupled with the intrinsic deterioration in GRIN imaging systems, we propose a novel architecture termed the classified-cascaded CNN (CC-CNN), which emerges as a critical problem-solver. Indeed, through simulation and experimentation, we demonstrate that GRIN waveguides of deteriorated lengths can cause catastrophic distortion in both imaging aberrations and virtual-real problems. Relatively, a typical U-net network may perform barely satisfactory only in one type of image (such as the real images), while serious distortions are still observed in the restoration of the other type (such as the virtual images). Leveraging insights from the prior knowledge of GRIN physical properties, we postulate an unrelated relationship between the virtual-real problem and the imaging aberration. Consequently, we propose the CC-CNN, which is designed to incorporate two distinct networks—one for discriminating between virtual and real images, and the other for correcting the physical aberration. By intentionally avoiding the intricate coupling of the aforementioned two types of distortions, the network decomposes the complex problems into two relatively simpler tasks for the two classified networks, respectively, and then makes them cascaded to solve the issue. As a result, the overall complexity is significantly reduced. Thus, compared to the traditional U-net architecture, the CC-CNN exhibits a lower dependence on data volume, enabling it to address challenges that traditional methods struggle with in data-limited scenarios.

The following sections will detail the experimental setup used for data acquisition involving GRIN lenses of randomly varied lengths. A CNN based on the U-Net architecture has been trained to correct comprehensive distortions in the images produced by GRIN lenses of all lengths. However, the network demonstrates superior restorative effects for real images (represented as positive images) but shows comparatively modest efficacy for virtual images (depicted as inverted images). As a comparison, the CC-CNN tailored for this scenario comprises a ResNeSt50-based[21] image real-virtual discrimination network and a U-Net-based[22] image physical-aberration correction network. When combined in sequence, this CC-CNN has successfully resolved the imaging restoration issues presented by GRIN lenses of varying lengths within the same training dataset.

2. Materials and Methods

2.1. Imaging of random-length GRIN systems

Due to the imaging characteristics of GRIN waveguides being highly dependent on their length, as discussed in Supplementary Material, and considering that the position and length of GRIN waveguides are often fixed in practical endoscopes, addressing the impact of length mismatch in waveguides through deep learning at the backend becomes a crucial topic. It is noteworthy that, despite achieving high precision in length through mechanical machining of GRIN materials, the issue of length mismatch in GRIN waveguides remains unsolved. This discrepancy stems from two main factors. First, the refractive index design of GRIN waveguides is typically optimized for a specific wavelength, such as in the visible light range with a reference wavelength of 550 nm. This specialization results

in the optical length of the GRIN waveguide differing from its designed value when used to transmit the light of other wavelengths[23]. Consequently, various light colors undergo different equivalent pitch (P) numbers, leading to pronounced chromatic dispersion. Second, the P of the GRIN waveguide, which denotes the periodicity of light's trajectory within it, is directly proportional to the waveguide's diameter[24]. Particularly in applications that require smaller-diameter endoscopes, GRIN waveguides often feature smaller P values. This means that even minor mechanical discrepancies can cause notable changes in the number of Ps, significantly influencing the GRIN waveguide's imaging characteristics (the quantitative discussion can be seen in Sec. 2 of Supplementary Material). Therefore, developing effective methods to mitigate the impact of length mismatch in GRIN waveguides is critically important.

To tackle this challenge, we designed an experiment, which is depicted in Fig. 1. This figure illustrates the imaging system, which employs GRIN lenses of randomly varied lengths, and the restoration achieved by the CC-CNN. Figure 1 has been divided into three parts. In Part 1, the optical path configuration of the experimental random-length GRIN imaging system is illustrated. Part 2 showcases the imaging principles and simulation results of the random-length GRIN lens. Part 3 depicts the process of image restoration through a CNN and the expected results for the images captured by the system.

As shown in Part 1, a 15 mm × 15 mm sized color image with a dimension of 320 pixel × 320 pixel was loaded onto a liquid crystal display (LCD, 1080 × 1920, BOE). Subsequently, the image was transmitted by 7 GRIN lenses of randomly different lengths (18.4–36.3 mm, Putian Huayue Electronics). Considering the principle of a single variable, each GRIN lens is cut from the same batch of customized 36.3 mm GRIN lenses with the same parameters. The core diameter of these lenses was 1 mm, the numerical aperture (NA) was 0.5, and the refractive index parameter $\sqrt{A}$ was consistently 0.334 mm^{-1}. The imaging process involved amplification and focusing through a combination of a 20× objective lens (OL, RMS20X-PF, Thorlabs) and a focusing lens, ultimately forming images on a CMOS camera (MER-630-60U3C-L, Daheng Optics). The positions of the GRIN lens and CMOS camera were adjusted using six-axis and three-axis translation stages (MAX609L and RBL13D, Thorlabs), respectively. The positions of these components were kept constant when switching between GRIN lenses of different lengths, ensuring their alignment with the objective lens. The LCD was adjusted using a linear translation stage (DDS300/M, Thorlabs) to the appropriate position, ensuring consistent image size received by the CMOS camera.

In Part 2, additionally, to visually illustrate the imaging characteristics of the GRIN waveguide, we conducted ray tracing simulations using ZEMAX. These simulations provided a theoretical perspective, which we then compared side by side with the corresponding experimental imaging results. This approach not only helps in understanding the optical behavior of GRIN waveguides but also in validating the effectiveness of the CNN in correcting imaging distortions caused by length mismatches. Among the results obtained from these 7 imaging systems, 5 sets of real images (by lenses 1, 2, 3, 4, and 7), and 2 sets of virtual images (by lenses 5 and 6) were generated. The imaging results reveal that the images transmitted through GRIN lenses of different lengths exhibit significant aberrations with substantial variations. The ray tracing simulations and experimental images corresponding to GRIN lenses are well-matched. In Fig. 1, yellow circles indicate the number of intersections during the trajectory of light, determining whether the GRIN lens produces a real or virtual image. Furthermore, when the number of intersections is even, a real image is formed, such as in lens 1 with 4 intersections and lens 7 with 2 intersections. When the number of intersections is odd, a virtual image is formed, as seen in lenses 5 and 6 with 3 intersections. The green circles in Fig. 1 denote the focusing degree of the light received by the CMOS surface. Higher convergence leads to higher clarity in imaging while decreasing convergence significantly reduces the fidelity of the imaging.

In Sec. 3, a concise description is provided for the restoration process using deep learning. The training and restoration process of this CNN is illustrated in the flowchart presented in Fig. 2(a). A CNN based on U-Net was trained to restore images captured through the 7 different lengths of GRIN lenses, with the specific network architecture detailed in Sec. 2.2. In this context, we consider direct imaging without GRIN lenses using the objective lens (OL) as the ground truth (GT) for deep learning. The images obtained through the 7 different GRIN lenses are considered as disturbed images, input into the network. Ultimately, a total of 772 GT images and 772 × 7 disturbed

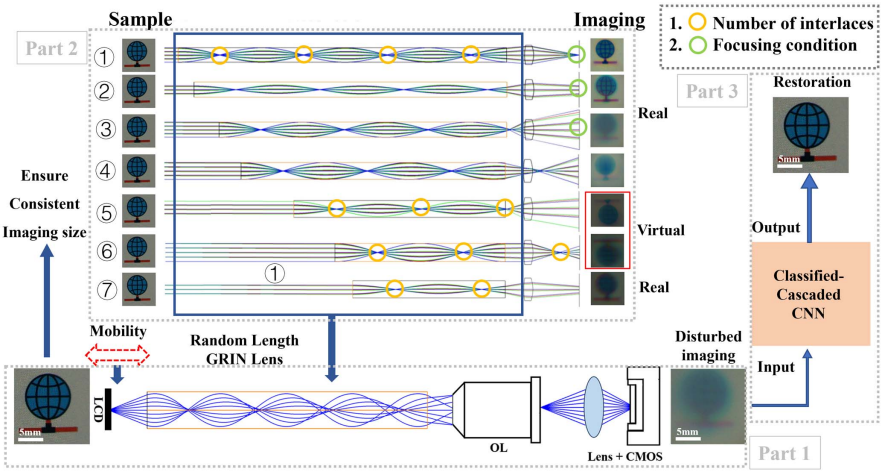

Fig. 1 Random-length GRIN lens image transmission system.

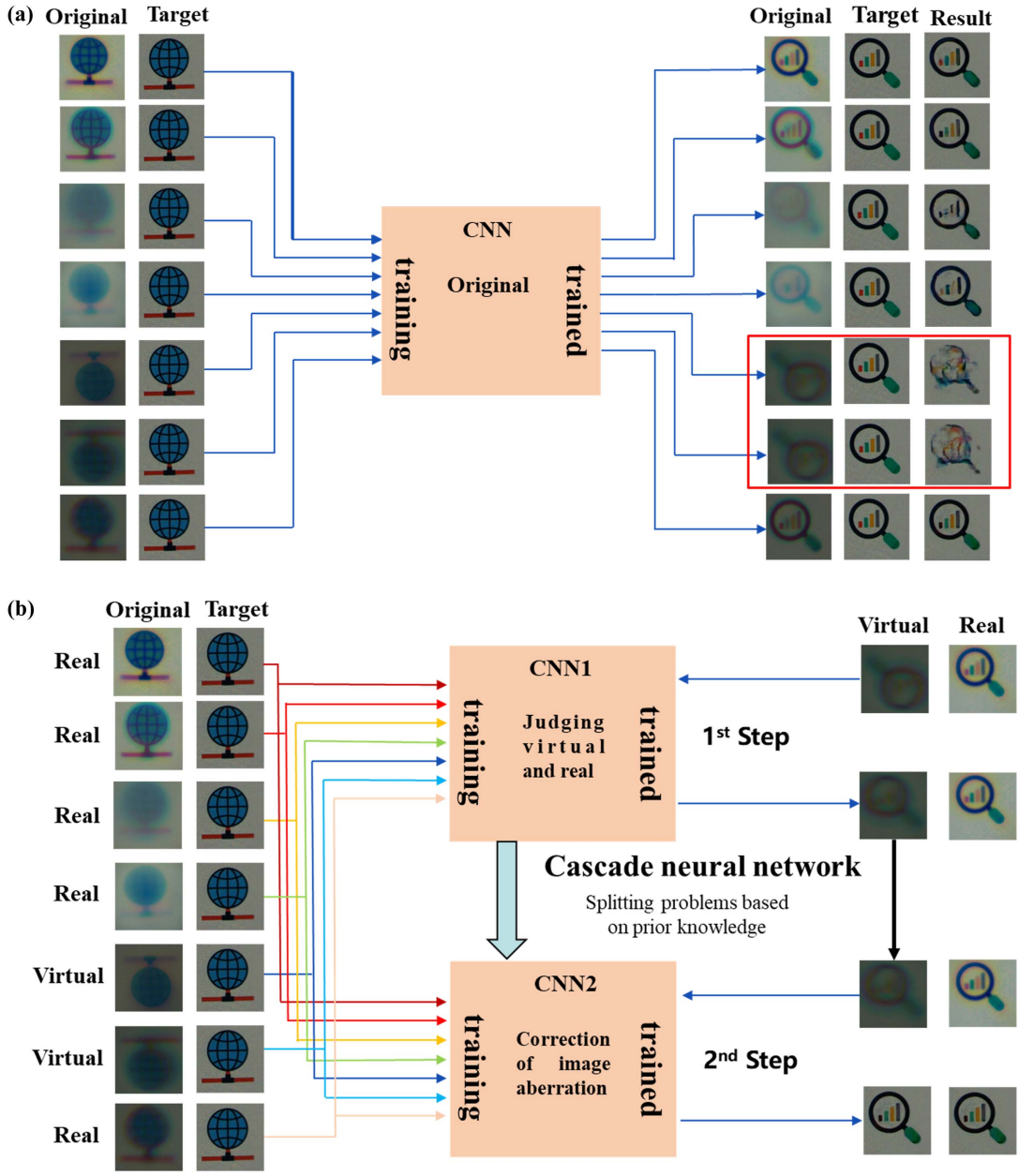

Fig. 2 Image restoration by (a) the single CNN and (b) the CC-CNN.

images were collected. Due to the invariance of GT, one GT will be reused 7 times to achieve pairing with 7 sets of distorted images. We designated 622 random sets of these image sets as the training set, while the remaining 150 sets were designated as the testing set. During the training phase, 622×7 distorted images and GT images will be used as input data for network training. Subsequently, during the testing phase, 150×7 distorted images will be used as input data to obtain network output, which will be compared with the corresponding 150×7 GT images to evaluate the quality of network recovery. All 7 sets of data will be matched with the GT simultaneously and used for training and testing, without labeling which GRIN lens they come from. Therefore, this CNN will have multitasking capability, which means it can simultaneously restore images from seven different GRIN imaging systems. Comparing the restored results with the GT images, it becomes clear that the network demonstrates favorable restoration for the five sets of real images. This is attributed to the lesser degradation induced by GRIN lenses compared to common waveguides, such as optical fibers, under this training regime. However, the restoration results for the two sets of virtual images fell short of expectations. Despite this, the graphical representation in the results section still indicates that the CNN is capable of correcting inverted virtual images to their upright counterparts. This ability, albeit not optimal for virtual images, showcases CNN's potential in addressing image inversion issues.

The subpar performance of a single CNN in restoring virtual images highlights the need for the capability to discriminate between real and virtual images. Traditional approaches might attempt to resolve this issue by enlarging the training dataset; however, this method is not cost-effective. Given that CNNs rely on pixel-to-pixel mapping relationships for image restoration,

accurately recovering inverted virtual images presents a significant challenge. Effective restoration of such images would require an impractically large training dataset, which is often unfeasible in real medical scenarios. Consequently, our focus is on exploring solutions that can address this issue without necessitating an expansion of the training dataset.

Given that a single CNN has demonstrated substantial potential in image restoration but performs inadequately when confronted with the need for direction correction in inverted images, a new CC-CNN has been designed, with the specific network architecture detailed in Sec. 2.2. Optical principles suggest that the differing real-virtual problem in images results from varying lengths of GRIN lenses. After convergence by the lens before the CMOS, the direction of the rays for real images is the same as the light emitted from the LCD, while for virtual images, it is in the opposite direction. Based on this prior knowledge, a reasonable assumption is posited: the real-virtual characteristic of the image is unrelated to its distortion effects. With this assumption, the complex task of simultaneously correcting the orientation of images and restoring distortions is split into two independent, simpler problems. Initially, a classified CNN is employed to ascertain the real-virtual problem of the image and adjust its orientation accordingly. Subsequently, another classified CNN is tasked with restoring the image's degradation. The decoupling of these two issues, real-virtual discrimination, and aberration restoration, serves to reduce the complexity of the problem and cascade to solute it. The schematic diagram of this network is illustrated in Fig. 2(b).

The training and testing datasets for the CC-CNN are the same as those in Fig. 2(a) for the single CNN. When a degraded image produced by a GRIN lens is input into the network, CNN1 based on ResNeSt50 first discriminates whether it is a real or virtual image. If it is a virtual image, CNN1 will flip it by 180°; for a real image, no adjustment will be made. The output from CNN1 is then used as input for CNN2 to correct aberrations, which is based on a U-Net architecture, similar to the single CNN. Ultimately, without augmenting the training dataset, the CC-CNN leveraging prior knowledge achieves effective restoration for both real and virtual images.

In the next section, we will introduce the network structures and algorithmic flows of the single CNN and CC-CNN in detail, and recovery results and comparison graphs of the two networks will be shown in Sec. 3.

2.2. Image restoration based on CNNs

With the advent of the information age and the rapid development of computer hardware, CNNs have become prominent in various fields, such as autonomous driving and medical image processing. CNNs utilize convolutional operations to filter the input image, extracting pivotal features such as edges and textures. This feature map is then processed through pooling and fully connected layers to perform tasks like image processing. In image restoration, CNNs proficiently learn the mapping relationship between input and output images at corresponding pixel coordinates, adeptly handling noisy inputs and restoring clear images based on predefined constraints. In this study, the variable lengths of GRIN lenses result in images that encompass both real and virtual representations, posing challenges for a single CNN to perform satisfactorily. Hence, we have adopted a cascaded network approach to restore images from random-length GRIN. Initially, an image classification network discerns whether the GRIN images are real or virtual, standardizing them into real images (i.e., leaving real images unchanged, and inverting virtual images). Subsequently, these processed images are fed into an image restoration network for recovery. Within the cascaded network, we have chosen ResNeSt50 as the classification network and an enhanced version of U-Net for image restoration.

Image classification is a pivotal task in deep learning, aimed at assigning input images to predefined categories. Our chosen architecture for this purpose is ResNeSt50, as illustrated in Fig. 3(a). This architecture builds upon the ResNet[25] framework proposed by He *et al*. ResNet revolutionized traditional convolution by introducing a residual module, effectively mitigating the gradient vanishing problem in deep neural network training through shortcut connections. Diverging from the conventional ResNet architectures, ResNeSt50 integrates the novel split-attention mechanism. This enhancement boosts the network's feature learning efficiency by dividing channels into separate groups, each with distinct weights. Such an architectural improvement enables the model to selectively focus on various facets of the input data, significantly augmenting its capability to identify and prioritize different features during the learning process. In handling our unlabeled, self-collected dataset, we categorize real images with the label "one" and virtual images with "zero." The implementation of multi-fold cross-validation further bolsters the network's classification performance on our dataset. With a total of 25.4 million parameters and 28.3 billion floating-point operations (FLOPs), our network achieves an exceptional 99% classification accuracy. This highlights the robustness and effectiveness of ResNeSt50 in adeptly tackling challenges in image classification.

The high accuracy of the network in distinguishing between virtual and real images can be mainly attributed to two factors. First, as a binary classification task, there is only one decision boundary, which simplifies the issue and reduces the likelihood of classification errors. Deep learning methods can automatically extract multi-level features from raw data, capturing complex nonlinear characteristics in the pixel distribution of images, thereby successfully classifying the images. The determination of whether an image is a virtual or real image in experiments depends on which GRIN lens it was imaged through, as GRIN lenses of different lengths possess markedly distinct imaging properties. Consequently, the judgment of virtuality or reality is not based on the abstract concept of orientation, but on analyzing the aberrations, in other words, the pixel distribution characteristics of the input image, making it a problem solvable by a computer. Second, we have employed regularization techniques such as Dropout[26] and Batch Normalization[27] to achieve higher accuracy and stronger generalization capabilities.

CNNs have demonstrated remarkable versatility in image-to-image applications, excelling particularly in the realm of image restoration. This task is centered around enhancing image quality by restoring clear, high-fidelity images from their blurry, noisy, or low-resolution counterparts. The U-Net-type CNNs, developed by Ronneberger *et al*.[28], are extensively employed for image segmentation in biomedical applications[29] and have also been explored in recent years for fiber image reconstruction[30–32]. U-Net, with its nearly symmetric architecture, is designed with a unique structure that combines down-sampling convolutional layers for extracting contextual features and up-sampling deconvolutional layers for image information

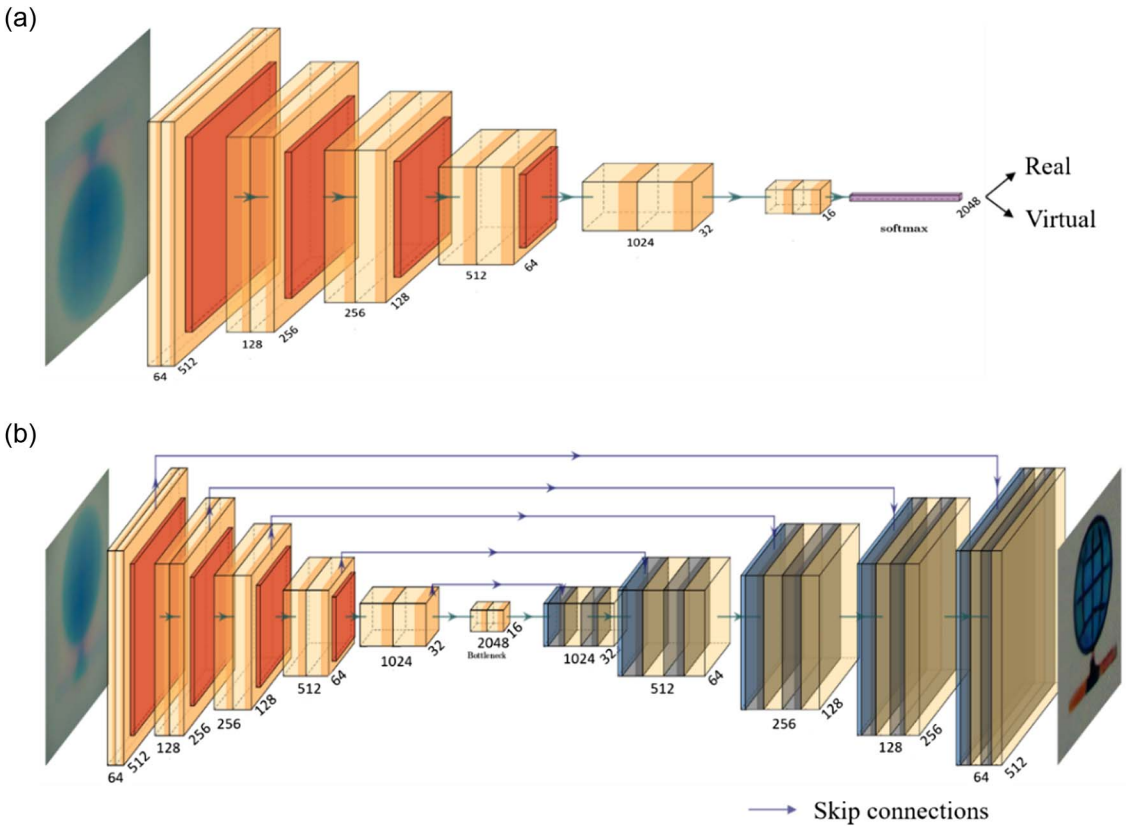

Fig. 3 Details of the implemented networks. (a) ResNeSt50 image classification. (b) U-Net image restoration.

recovery. This symmetry enables the effective capture of both local and global features in the input data, making it particularly suited for tasks like image segmentation and restoration. The down-sampling layers methodically reduce spatial dimensions to grasp hierarchical features, while the up-sampling layers reconstruct spatial information to generate high-resolution outputs. Its innate ability to balance feature extraction with spatial fidelity renders it a powerful tool in various image-to-image applications. To augment network performance, we have made some improvements. First, an additional down-sampling layer is introduced to adjust the spatial resolution of feature maps, capturing broader contextual information, as depicted in Fig. 3(b). Second, a lightweight channel attention mechanism[33] is integrated, offering finer control over the information flow between channels. This synergistic strategy enhances the network's adaptability, promoting more efficient and effective feature extraction. The network comprises 59.6 million parameters and 41.8 billion FLOPs, facilitating high-fidelity image recovery. The networks are trained for a maximum of 400 epochs and implemented on a single NVIDIA GeForce GTX 3080 graphics processing unit using the PyTorch framework. In the presented examples, we have achieved image reconstruction of a 256 pixel $\times$ 256 pixel self-curated dataset using the cascaded network on a single GPU at 26 frames per second (FPS), indicating impressive operational speed conducive to practical applications. The cascaded network details can be found in Sec. 3 of Supplementary Material.

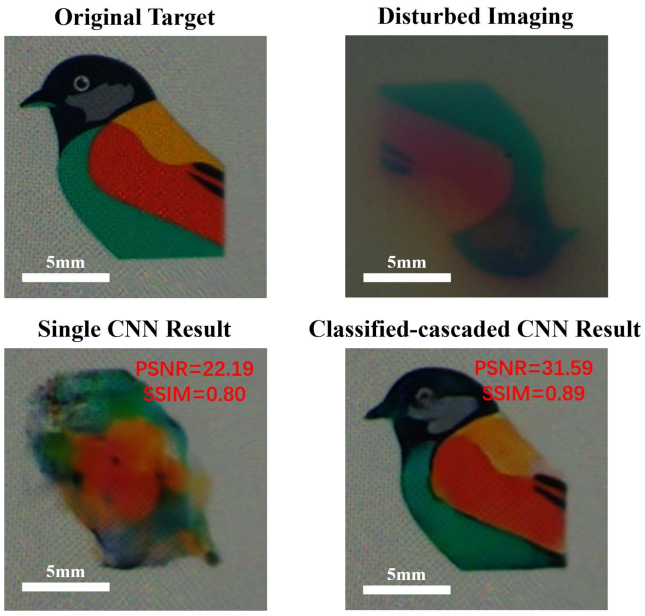

Fig. 4 Comparison of restoration results between the single CNN and the CC-CNN for virtual imaging.

3. Results

3.1. CNN restoration performance assessment

To rigorously assess the recovery outcomes and provide a comprehensive depiction of the reconstruction performance, we

Table 1 Comparisons between Single U-net and CC-CNN in All Images.

	Network Type	PSNR	SSIM
Random length	U-Net	27.74	0.76
GRIN	CC-CNN	30.97	0.81

calculate the average peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) for the test set, as detailed in Table 1. PSNR and SSIM serve as widely employed metrics for evaluating the effectiveness of image reconstruction methods. The PSNR is a measure of image quality, calculated by comparing the difference between the original image and the distorted image. It is the ratio between the maximum possible power of the original image and the power of the distorted image. The SSIM is a more advanced image quality assessment method that considers the structural information of the image and is closer to the human eye's perception of image quality. Overall, the SSIM can better reflect the human eye's perception of image quality, while the PSNR mainly depends on the error of pixel values, reflecting the machine's perception of image quality[34].

PSNR is defined as

$$\mathrm{PSNR} = 10 \times \lg \frac{255^2}{\mathrm{MSE}}, \tag{1}$$

$$\mathrm{MSE} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} [X(i,j) - Y(i,j)]^2. \tag{2}$$

In Eqs. (1) and (2), MSE denotes the mean square error between the current image X and the reference image Y, where $X(i,j)$ and $Y(i,j)$ denote the pixel values at corresponding coordinates. H and W represent the height and width of the image, respectively, and the image has a grayscale range of 256 levels. The unit of PSNR is decibels (dB). It is evident that a higher PSNR corresponds to a smaller MSE, signifying a closer resemblance between the original and restored images.

The structural similarity method evaluates the image quality from the following aspects: brightness, contrast, and structure. The concrete expression of SSIM between X and Y is

$$\mathrm{SSIM}(X,Y) = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{(\mu_X^2 + \mu_Y^2 + C_1)(\sigma_X^2 + \sigma_Y^2 + C_2)}, \tag{3}$$

where C_1, C_2, and C_3 are constants to avoid the denominator to be 0 and maintain stability. Usually $C_1 = (K_1 L)^2$, $C_2 = (K_2 L)^2$, $C_3 = C_2/2$, and generally $K_1 = 0.01$, $K_2 = 0.03$, $L = 255$ (is the dynamic range of pixel values, generally taken as 255),

$$\mu_X = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} X(i,j), \tag{4}$$

$$\mu_Y = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} Y(i,j), \tag{5}$$

$$\sigma_X = \left\{ \frac{1}{M \times N - 1} \sum_{i=1}^{M} \sum_{j=1}^{N} [X(i,j) - \mu_X]^2 \right\}^{\frac{1}{2}}, \tag{6}$$

$$\sigma_Y = \left\{ \frac{1}{M \times N - 1} \sum_{i=1}^{M} \sum_{j=1}^{N} [Y(i,j) - \mu_Y]^2 \right\}^{\frac{1}{2}}, \tag{7}$$

$$\sigma_{XY} = \left\{ \frac{1}{M \times N - 1} \sum_{i=1}^{M} \sum_{j=1}^{N} [X(i,j) - \mu_X][Y(i,j) - \mu_Y] \right\}, \tag{8}$$

where structural similarity value ranges from 0 to 1. With the larger value, the image similarity is higher.

3.2. Comparison of image recovery results

In Sec. 2, we devised two distinct CNNs. Following their training through deep learning, we employed them to restore the same set of test datasets to showcase their respective efficacy. Both CNNs exhibited remarkable capabilities in restoring images degraded by real images, with noticeable distinctions emerging when faced with virtual images. A comparison of their performance when dealing with virtual images is depicted in Fig. 4. Additionally, representative test results are presented in Fig. 5, illustrating their individual restoration performances. Besides, more testing results are presented in Sec. 4 of Supplementary Material.

Due to the significant difference in performance between single CNN and CC-CNN in handling virtual images, we prioritize presenting results for this scenario. Figure 4 presents a more comprehensive comparison of the restoration results for inverted images between the single CNN and the CC-CNN. Clearly, the CC-CNN exhibits superior performance. Quantitatively, the CC-CNN achieved a 6.20 dB improvement in the PSNR parameter compared to the single CNN. Additionally, there was a 0.25 dB improvement in the SSIM parameter. The definitions and calculation processes of these parameters will be described in Sec. 3.2. This can be attributed to the high accuracy of the real-virtual discrimination network within the CC-CNN, further confirming the validity of the prior assumption that the nature of inversion and aberrations is independent.

From Fig. 5, it becomes apparent that in imaging scenarios involving GRIN lenses of varying lengths, the resulting images appear blurred, displaying dispersion and ghosting effects, yet they remain recognizable. In this typical distortion scenario, the single CNN effectively restores the results for all five sets of real images. However, when it comes to virtual images, significant distortion is observed, underscoring the increased complexity involved in simultaneously addressing image orientation and aberration restoration. The current training dataset proves inadequate for this task, leading to overfitting in the case of inverted images.

In comparison, the CC-CNN not only addresses blurring, dispersion, and ghosting in degraded images, much like the single CNN, but it also effectively resolves the overfitting problem the single CNN encounters with virtual images. The CC-CNN delivers restoration results for virtual images that are on par with those for real images. This success is largely due to the high accuracy of the real-virtual discrimination network within the

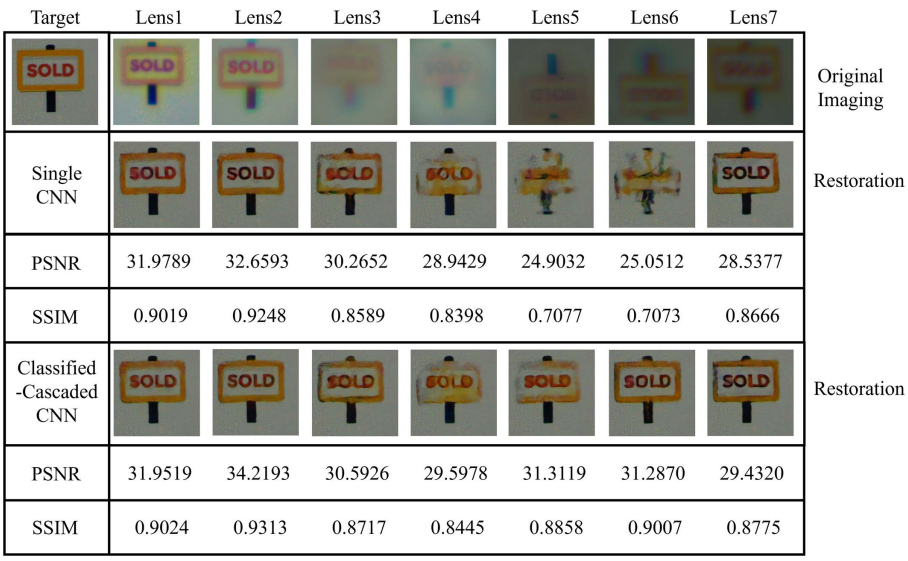

Fig. 5 Imaging restoration for each GRIN lens by the single CNN and the CC-CNN.

Table 2 Comparisons between Single U-Net and CC-CNN in the Virtual Images.

	Network Type	PSNR	SSIM
Random length	U-Net	22.98	0.70
GRIN	CC-CNN	31.00	0.81

CC-CNN. This confirms the relative validity of our assumption, based on prior knowledge, that the nature of image inversion and aberrations is mutually independent.

Tables 1 and 2 present the average PSNR and SSIM values for images captured using the single U-Net and CC-CNN under random lengths of GRIN. Table 1 includes both real and virtual images, while Table 2 focuses solely on virtual images. The results indicate a notable enhancement in both SSIM and PSNR metrics for images processed by the cascaded network compared to those obtained from a single U-Net network. This improvement is further corroborated by the research findings depicted in Fig. 5. The cascaded network exhibits remarkable effectiveness in the image recovery process for GRIN imaging of arbitrary lengths. This phenomenon is likely due to the inclusion of both real and virtual images in the GRIN-acquired data, leading to significant variations in the input. As a result, this disparity disrupts the pixel mapping relationships within the U-Net network, underscoring the cascaded network's superior adaptability in addressing the challenges posed by GRIN imaging with varying lengths.

4. Discussion and Conclusion

4.1. Discussion

Hopkins endoscopes, as mainstream rigid endoscopes, have been widely utilized for over a century. However, the accumulation of aberrations caused by rod lenses and the complex assembly structure poses challenges for Hopkins endoscopes, including difficulties in image correction, bulky system design, and higher risks of injury to patients. The integration of GRIN waveguides and CNN image restoration in endoscope architecture is gaining attention as a promising solution for the future[35–38]. The article acknowledges the potential of GRIN endoscopes while highlighting their main constraint—the significant intrinsic deterioration they face in real medical settings. In endoscopes based on narrow-diameter GRIN lenses, minimal machining errors and mode dispersion corresponding to different colored light sources can lead to changes in their optical lengths, thereby severely impacting imaging performance. Traditional CNNs typically rely on extremely large datasets to address this issue, which could be impractical in the medical imaging domain. Therefore, achieving imaging restoration for intrinsic deterioration in GRIN imaging systems is not only essential for developing next-generation endoscopes but also holds significant implications for the broader application of CNNs in realistic imaging restoration fields[39–41].

Specifically, in this experiment, we selected seven randomly sized GRIN lenses for simulation and imaging experiments. The results indicate that when the structural parameters of the imaging system change, the introduced distortions undergo significant nonlinear variations, making it challenging to discern patterns[42,43]. More critically, the imaging system's production of virtual and real images can also vary, implying that images captured at a fixed receiver may appear either upright or inverted. This further complicates distortion correction. Here, we employed both a traditional single network based on U-Net and a CC-CNN comprising ResNeSt50 and U-Net, training and testing them on a small dataset of 772 image sets. Results reveal that while traditional CNNs perform comparably to CC-CNN in distortions restoration for real images, the latter exhibits unparalleled superiority in virtual image scenarios. It is worth emphasizing that these experimental findings hold significant scientific value both in technical and scientific aspects.

In the technical aspect, we successfully achieved the robust restoration of intrinsic deteriorations in GRIN waveguide imaging. Different structural parameters of GRIN waveguides exhibit markedly different imaging mechanisms, significantly complicating the task of imaging restoration. Fortunately, the

CC-CNN effectively addressed this challenge with a small batch of datasets. This signifies the identification of a feasible, cost-effective distortion correction solution for GRIN waveguide imaging, offering technical feasibility for the realization of narrow-diameter GRIN endoscopes.

In the scientific aspect, we have proposed a potentially inspirational approach to the intriguing physical question of how to extend imaging restoration methods, which are mostly exclusive to static optical systems, to systems where structural parameter changes occur. Real-world optical imaging systems are often subject to intrinsic deterioration in their structural parameters, rather than being fixed. Traditional networks tackling this issue often require extensive training data, which may be unattainable and lead to overfitting. Through experimentation, we have observed that intrinsic deterioration manifests in imaging degradation as a superposition of various disturbances from different sources, such as the coupling of aberration changes with uncertainties in image fidelity. In traditional solutions, these two factors are not adequately separated, resulting in degradation effects akin to multiplication. However, in the CC-CNN, by decomposing and gradually restoring the two types of disturbances based on prior physical knowledge, the degradation effects resemble addition, significantly reducing the complexity of the problem. Thus, high-quality image restoration is achieved without increasing the volume of data.

In conclusion, we hope that this research not only provides better insights for the design of the next generation of GRIN endoscopes but also inspires more researchers to combine machine learning with optical imaging. Often, optical system aberrations stem from multiple sources. If we can rationally decompose their origins based on physical models and train multiple targeted network combinations, it could effectively reduce the complexity of issues faced by CNNs.

4.2. Conclusion

In conclusion, we proposed the CC-CNN, which is a combination of ResNeSt50 and U-Net, to achieve high-definition color image degradation correction for a random-length GRIN lens imaging system. The low dependency on data volume and high robustness to variations in GRIN waveguide length make the manufacturing of endoscopes centered around GRIN lenses feasible. Additionally, we addressed a critical scientific question, demonstrating that complex problems can be decoupled into combinations of simpler subproblems through reasonable prior knowledge. Subsequently, multiple networks were designed for each subproblem, and a cascaded network was employed to address the overarching problem. While this may increase the time required for CNN image restoration, it significantly alleviates the challenges associated with supplementing training datasets inspiringly, particularly in fields with high imaging costs such as medical imaging.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62125503 and 62261160388), the Natural Science Foundation of Hubei Province of China (No. 2023AFA028), the Key R&D Program of Hubei Province of China (Nos. 2020BAB001, 2020BAA007, and 2021BAA024), and the Innovation Project of Optics Valley Laboratory (No. OVL2021BG004). The authors thank Putian Huayue Electronics Co. Ltd. for providing customized GRIN lenses for this work.

Authors' Contributions

Yang Yang, Jinwei Zeng, and Jian Wang conceived the idea. Zaipeng Duan analyzed the data and conducted the imaging reconstruction. Ruiqi Zhou, Jie Ma, Jiong Xiao, Zihang Liu, and Feifei Hao supported the GRIN optical imaging experiment and helped provide the imaging materials. All authors read and approved the final manuscript.

References

1. R. Weissleder, "A clearer vision for in vivo imaging," *Nat. Biotechnol.* **19**, 316 (2001).
2. W. Yang and R. Yuste, "*In vivo* imaging of neural activity," *Nat. Methods* **14**, 349 (2017).
3. B. Huang, H. Babcock, and X. Zhuang, "Breaking the diffraction barrier: super-resolution imaging of cells," *Cell* **143**, 1047 (2010).
4. Z. Wen *et al.*, "Single multimode fibre for in vivo light-field-encoded endoscopic imaging," *Nat. Photonics* **17**, 679 (2023).
5. W. Choi *et al.*, "Flexible-type ultrathin holographic endoscope for microscopic imaging of unstained biological tissues," *Nat. Commun.* **13**, 4469 (2022).
6. B. T. Petersen *et al.*, "Multisociety guideline on reprocessing flexible GI endoscopes: 2016 update," *Gastrointest. Endosc.* **85**, 282 (2017).
7. R. Loddenkemper, "Thoracoscopy–state of the art," *Eur. Respir J.* **11**, 213 (1998).
8. S. A. Rodeo, R. A. Forster, and A. J. Weiland, "Neurological complications due to arthroscopy," *J. Bone Jt. Surg.* **75**, 917 (1993).
9. F. J. Gerges, G. E. Kanazi, and S. I. Jabbour-khoury, "Anesthesia for laparoscopy: a review," *J. Clin. Anesth.* **18**, 67 (2006).
10. T. E. Linder, D. Simmen, and S. E. Stool, "Revolutionary inventions in the 20th century: the history of endoscopy," *Arch. Otolaryngol. Head Neck Surg.* **123**, 1161 (1997).
11. G. Liu *et al.*, "Bendable long graded index lens microendoscopy," *Opt. Express* **30**, 36651 (2022).
12. C. R. Steffens *et al.*, "CNN based image restoration: adjusting Ill-exposed sRGB images in post-processing," *J. Intell. Robot Syst.* **99**, 609 (2020).
13. X. Hu *et al.*, "High-quality color image restoration from a disturbed graded-index imaging system by deep neural networks," *Opt. Express* **31**, 20616 (2023).
14. S. A. Ponomarenko, "Self-imaging of partially coherent light in graded-index media," *Opt. Lett.* **40**, 566 (2015).
15. S. Sivankutty *et al.*, "Ultra-thin rigid endoscope: two-photon imaging through a graded-index multi-mode fiber," *Opt. Express* **24**, 825 (2016).
16. B. Rasti *et al.*, "Image restoration for remote sensing: overview and toolbox," *IEEE Geosci. Remote Sens. Mag.* **10**, 201 (2022).
17. A. Wali *et al.*, "Recent progress in digital image restoration techniques: a review," *Digit. Signal Process* **141**, 104187 (2023).
18. J. Su, B. Xu, and H. Yin, "A survey of deep learning approaches to image restoration," *Neurocomputing* **487**, 46 (2022).
19. B. Rahmani *et al.*, "Multimode optical fiber transmission with a deep learning network," *Light Sci. Appl.* **7**, 69 (2018).
20. C. Zhu *et al.*, "Image reconstruction through a multimode fiber with a simple neural network architecture," *Sci. Rep.* **11**, 896 (2021).
21. H. Zhang *et al.*, "ResNeSt: split-attention networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2022), p. 2735.
22. O. Oktay *et al.*, "Attention U-Net: learning where to look for the pancreas," arXiv:1804.03999 (2018).
23. G. Yabre, "Comprehensive theory of dispersion in graded-index optical fibers," *J. Lightwave Technol.* **18**, 166 (2000).

24. M. O. F. Rasel, A. Yamauchi, and T. Ishigure, "Error-free three-dimensional multimode crossover graded-index polymer waveguides for board-level optical circuitry," *J. Lightwave Technol.* **40**, 1 (2022).
25. K. He *et al.*, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), p. 770.
26. N. Srivastava *et al.*, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.* **15**, 1929 (2014).
27. S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning* (2015), p. 448.
28. N. Navab *et al.*, "Medical image computing and computer-assisted intervention–MICCAI 2015," in *18th International Conference* (Springer International Publishing, 2015).
29. T. Falk *et al.*, "U-Net: deep learning for cell counting, detection, and morphometry," *Nat. Methods* **16**, 67 (2019).
30. J. Zhao *et al.*, "High-fidelity imaging through multimode fibers via deep learning," *J. Phys. Photonics* **3**, 015003 (2021).
31. C. Zhu *et al.*, "Image reconstruction through a multimode fiber with a simple neural network architecture," *Sci. Rep.* **11**, 896 (2021).
32. L. Zhang *et al.*, "High definition images transmission through single multimode fiber using deep learning and simulation speckles," *Opt. Lasers Eng.* **140**, 106531 (2021).
33. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018).
34. Z. Wang *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.* **13**, 600 (2004).
35. F. Liu *et al.*, "Deeply seeing through highly turbid water by active polarization imaging," *Opt. Lett.* **43**, 4903 (2018).
36. K. Wang *et al.*, "On the use of deep learning for phase recovery," *Light Sci. Appl.* **13**, 4 (2024).
37. B. Lin, X. Fan, and Z. Guo, "Self-attention module in a multi-scale improved U-net (SAM-MIU-net) motivating high-performance polarization scattering imaging," *Opt. Express* **31**, 3046 (2023).
38. X. Wang *et al.*, "Real-time vision through haze based on polarization imaging," *Appl. Sci.* **9**, 142 (2019).
39. Y. Zhao *et al.*, "Accurate calculation of computer-generated holograms using angular-spectrum layer-oriented method," *Opt. Express* **23**, 25440 (2015).
40. Z. He *et al.*, "Progress in virtual reality and augmented reality based on holographic display," *Appl. Opt.* **58**, A74 (2019).
41. X. Shao *et al.*, "An improved infrared dim and small target detection algorithm based on the contrast mechanism of human visual system," *Infrared Phys. Technol.* **55**, 403 (2012).
42. T. Treibitz and Y. Y. Schechner, "Active polarization descattering," *IEEE Trans. Pattern Anal. Mach. Intell.* **31**, 385 (2009).
43. N. K. Soni, R. V. Vinu, and R. K. Singh, "Polarization modulation for imaging behind the scattering medium," *Opt. Lett.* **41**, 906 (2016).