



Article

Fast Optical Image Compression Based on a Hybrid D²NN–ENN Autoencoder

Sofia S. Stroganova, Viktor V. Krasnikov , Artem V. Chetvertukhin, Andrey A. Grunin and Andrey A. Fedyanin *

Faculty of Physics, Lomonosov Moscow State University, Moscow 119991, Russia;
stroganova.ss22@physics.msu.ru (S.S.S.); krasnikovvv@my.msu.ru (V.V.K.);
chetvertukhin@nanolab.phys.msu.ru (A.V.C.); grunin@nanolab.phys.msu.ru (A.A.G.)

* Correspondence: fedyanin@nanolab.phys.msu.ru

Abstract

The rapid growth of visual data in imaging and sensing motivates compression before full electronic processing. This work investigates a hybrid D²NN–ENN autoencoder for grayscale image compression, in which a passive diffractive optical encoder forms a compact latent intensity representation and an electronic neural-network decoder reconstructs the image. We evaluate optical precompression under different illumination modes, latent-region sizes, diffractive depths, mask parameterizations, and quantization levels. Increasing the latent region improves reconstruction quality only up to an intermediate range, after which the gain saturates. Coherent illumination provides a modest reconstruction-quality advantage over incoherent illumination. The best tested configurations use two or three diffractive layers. Full amplitude–phase modulation performs best, whereas phase-only modulation remains a practical lower-complexity alternative. Quantization at 8 and 4 bits preserves most reconstruction quality, while 2-bit precision causes visible degradation. Although JPEG and learned neural codecs achieve better rate–distortion performance, the proposed architecture is promising as an ultrafast optical precompression front end for specialized visible, near-infrared, and infrared sensing systems in which low encoding latency is more important than general-purpose storage efficiency.

Keywords: diffractive deep neural networks; optical image compression; hybrid optoelectronic neural networks; autoencoders; optical precompression; computational imaging

1. Introduction

The rapid growth in the volume of visual data in scientific imaging, autonomous sensing, communication systems, and multimedia applications makes image compression one of the key problems of information processing. Classical codecs such as JPEG remain important reference points because they use mature transform, quantization, and entropy-coding schemes with well-studied rate–distortion characteristics [1]. Newer image and video coding standards develop the same general idea through more complex prediction, transform selection, in-loop filtering, context modeling, and arithmetic coding; a representative example of this development in block-based hybrid coding is VVC/H.266 [2]. In parallel, learned image compression treats a codec as an end-to-end optimizable system that combines a nonlinear analysis transform, quantization or its differentiable surrogates, a probabilistic entropy model, and a nonlinear synthesis transform optimized with respect to a rate–distortion objective. This direction has already led to formal standardization initiatives, including JPEG AI [3].



Academic Editor: Shuping He

Received: 29 July 2026

Revised: 24 September 2026

Accepted: 9 October 2026

Published: 15 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

A separate research direction is associated with optical and hybrid optoelectronic neural architectures, in which part of the computation is transferred from digital electronics to the physical propagation of an optical field. The modern concept of a diffractive deep neural network (D²NN) was proposed by Lin et al.; the authors showed that a cascade of trainable passive diffractive layers can perform optical inference after fabrication [4]. The first experiments included handwritten-digit classification and lens-like wavefront transformations in the terahertz range. However, the principle is not limited to terahertz radiation: diffractive propagation, spatial modulation, and sensor readout can also be implemented in the visible and near-infrared (NIR) ranges. Such approaches are therefore relevant to NIR imaging, sensing, and machine-vision systems. The significance of the early results was not that passive optics outperformed conventional electronic neural networks, but that they demonstrated that a trained optical structure can perform inference through free-space propagation with a latency determined mainly by the optical path length. This potential for high-speed and energy-efficient processing has stimulated broad interest in photonic computing. However, its practical advantages over GPUs and TPUs depend on the physical constraints and configuration of the optical system; unified device models can identify favorable operating regimes and guide hardware design before fabrication [5]. At the same time, the accuracy of the early five-layer D²NN on MNIST remained below that of standard electronic convolutional models, motivating further studies of optical readout schemes, network depth and capacity, spatial coherence, and hybrid architectures [4,6–8].

The limitations of fully passive D²NNs are fundamental for understanding their role in compression tasks. In the absence of optical nonlinearities, a diffractive system is constrained by linear wave propagation, aperture size, diffraction, coherence, fabrication accuracy, and alignment errors. Therefore, increasing the number of diffractive layers is not directly analogous to increasing the depth of an electronic neural network. Studies of the capacity and depth of diffractive networks show that additional layers can improve trainability or reduce the required number of diffractive features only up to a certain limit; after that, the gain may saturate and does not necessarily surpass a carefully optimized shallow architecture [7,9]. In this regard, reviews and methodological studies increasingly emphasize the importance of hybrid optoelectronic schemes, in which optics performs a fast, passive, and approximately linear front-end transform, while the electronic part provides nonlinear adaptation, decoding, and robustness to hardware imperfections [8,10].

This engineering logic naturally leads to hybrid D²NN–ENN architectures. Chang et al. demonstrated one of the early hybrid optical–electronic convolutional neural networks, in which an optimized diffractive optical element served as a physical input block before an electronic neural network [11]. This formulation differs from a fully optical D²NN: the optical component is not required to solve the entire task by itself, but is trained jointly with the electronic part, which compensates for some of the representational limitations of passive optics. Subsequently, related hybrid schemes extended the use of optical processing from classification to data transmission, reconstruction, image formation, and representation learning [12–16].

The breadth of this idea is illustrated by several examples. In super-resolution displays, an electronic encoder can transform a high-resolution image into a low-resolution modulation pattern, while a trained diffractive decoder optically reconstructs an image with an increased space–bandwidth product [13]. In diffractive optical communication around opaque obstacles, a digital convolutional encoder and a passive diffractive decoder are trained jointly to recover information from waves scattered around objects that block the direct line of sight [14]. Spectrally encoded single-pixel machine vision shows that a diffractive network can encode spatial information into the spectrum of diffracted light measured by a single-pixel spectroscopic detector, with spectral channels serving as outputs

associated with individual classes [15]. In turn, information-hiding cameras demonstrate that a passive diffractive processor can mask information about an object in an intensity distribution that appears ordinary, after which a jointly trained CNN reconstructs the hidden image [16]. These works show that an optical input block can encode data spatially, spectrally, geometrically, or through task-dependent hidden representations.

For the present work, studies in which the optical module is explicitly regarded as an encoder, that is, as a physical mechanism for compression or compact representation formation, are especially important. Wang et al. proposed an integrated photonic encoder motivated by the growth of data streams in modern cameras. Their system performs analog photonic matrix–vector transformations based on random projections before full digital registration, thereby reducing the amount of data requiring electronic processing [17]. At the same time, accelerating an optical projection does not necessarily guarantee an end-to-end computational advantage. Fomin et al. analyzed several random-projection-based algorithms and showed that the projection often accounts for only a small fraction of their total computational time, making system-level bottleneck analysis essential when assessing optical processors [18]. The concept of a fully optical autoencoder proposed by Feng et al. also shows that linear diffractive processors can be trained to encode an input optical field into a latent diffractive representation and then reconstruct it [19]. The Latent Space Imaging approach develops this idea toward task-oriented compression: optical measurements are mapped into the latent space of a pretrained generative model, which makes it possible to operate at compression ratios on the order of 1:100–1:1000 and, in certain regimes, up to 1:4096 [20]. In these works, the evaluation criterion shifts from pixel-level reconstruction alone to the informativeness and application utility of the physical latent representation.

Recent hardware-oriented studies also confirm the transition from optical classification to optical encoding. Architectures of optical matrix–vector multipliers have been proposed for both encoding and decoding, including schemes adapted for image reconstruction from low-dimensional latent vectors [21]. Terahertz compressive optical–digital neural network imaging uses a diffractive optical neural network to compress amplitude–phase information into a substantially more compact intensity measurement before electronic reconstruction [22]. Integrated optical diffractive nonlinear neural networks with degenerate input and hidden layers show that integrated diffractive systems can combine classification with strong feature compression in compact photonic hardware [23]. Taken together, these results indicate the emergence of several variants of optical encoding: spectral encoding, latent-space imaging, optical matrix–vector multiplication, communication through complex physical channels, and integrated photonic diffractive processing.

Despite this progress, optical and hybrid D²NN systems are rarely evaluated as image-compression systems in the same sense as JPEG, VVC, or learned neural codecs. In traditional and learned codecs, compression is defined by the number of bits after quantization and entropy coding, while quality is usually described by rate–distortion curves, PSNR, MS-SSIM, perceptual metrics, encoding and decoding latencies, and memory requirements [1,2]. In D²NN–ENN systems, by contrast, an explicit bitstream is often absent: the optical component physically projects an image into a domain of fields, intensities, modes, spectra, or measurements, and the electronic part uses this representation for reconstruction or inference. Such systems are therefore closer to analog compressive sensing or physical feature extraction than to a full digital codec with entropy coding, unless the latent representation, sensor quantization, noise model, bit depth, and hardware constraints are explicitly specified. In this context, the size of the latent region of the proposed autoencoder can be interpreted by analogy with the quality setting in a classical codec: changing the latent representation size and the quantization level defines different operating points with different sizes of the compressed representation. For this reason,

the central comparison in this work is performed primarily in terms of the resulting file size rather than only in terms of the nominal model configuration.

This methodological gap motivates the present study. We consider a hybrid D²NN–ENN autoencoder as an image-compression and reconstruction system located at the boundary between optical neural networks and learned compression. The D²NN encoder forms a compact optical latent representation, and the electronic decoder reconstructs the image from it. The aim of the work is not only to demonstrate reconstruction, but also to evaluate how the latent-region size, encoder depth, quantization, and illumination mode (coherent or incoherent) affect reconstruction quality and computational time. The resulting trade-offs are compared with the characteristics of established codecs under explicitly specified experimental conditions. This formulation is especially relevant for application scenarios in which encoding speed is critical: if the optical input block is physically integrated with the image-acquisition system, the D²NN–ENN encoder can be useful as a fast passive precompressor, even if it is not intended to replace general-purpose codecs with entropy coding.

2. Materials and Methods

2.1. Dataset and Preprocessing

The experiments were performed on the Animal Faces-HQ (AFHQ) dataset [24]. AFHQ was selected because animal faces provide readily recognizable shapes and spatially organized features, such as the head outline, eyes, ears, muzzle, and fur texture. Consequently, degradation caused by optical compression can be assessed directly through distortions of facial geometry, loss of contour sharpness, and suppression of fine textural details. The dataset therefore enables a controlled evaluation of how well the optical latent representation preserves both coarse image structure and high-frequency content. The experimental corpus contained 16,130 images; 1500 of them were used for validation. In addition, a separate test subset of 300 images was formed from the training set and used only for final metric evaluation after training. The original 512×512 RGB images were converted to single-channel grayscale images and resized to 400×400 pixels. This preprocessing ensured consistency with the optical simulation grid used by the D²NN encoder and with the spatial resolution reconstructed by the electronic decoder.

2.2. Optical D²NN Encoder Model

The optical part of the autoencoder was modeled as a stack of diffractive propagation and modulation operators. The preprocessed 400×400 grayscale image was placed into a 600×600 computational aperture. The resulting input field was denoted by $u_0(x, y)$. For an encoder with N diffractive layers, the field evolution was written as

$$u_n = m_n \odot \mathcal{P}_{d_n}\{u_{n-1}\}, \quad n = 1, \dots, N,$$

where $\mathcal{P}_{d_n}$ is the free-space propagation operator over distance d_n , $m_n(x, y)$ is the trainable complex modulation mask of the n th diffractive layer, and $\odot$ denotes elementwise multiplication. Free-space light propagation was described by the scalar Fresnel–Huygens formula. In the paraxial approximation, for wavelength λ and wavenumber $k = 2\pi/\lambda$, the operator $\mathcal{P}_d$ can be written as

$$\begin{aligned} \mathcal{P}_d\{u\}(x, y) &= \frac{\exp(ikd)}{i\lambda d} \iint_{\Omega} u(\xi, \eta) \\ &\quad \times \exp\left[\frac{ik}{2d}\left((x - \xi)^2 + (y - \eta)^2\right)\right] d\xi d\eta. \end{aligned}$$

Here, Ω denotes the aperture of the previous plane, that is, the region covered by the discrete computational grid. Thus, the field in the next plane is the coherent sum of contributions from all points of the previous plane with a quadratic phase delay. In the numerical implementation, this linear operator was computed on a discrete grid and remained differentiable, so the gradient of the loss function could propagate through diffractive propagation to the mask parameters. Each modulation mask corresponds to an amplitude–phase diffractive element,

$$m_n(x, y) = a_n(x, y) \exp(i\phi_n(x, y)),$$

where a_n and ϕ_n are the trainable amplitude and phase, respectively. After the last diffractive layer, the field propagated to the latent plane,

$$u_{\text{lat}} = \mathcal{P}_{d_{N+1}}\{u_N\}.$$

In the detector plane, the intensity was recorded as

$$I_{\text{lat}}(x, y) = |u_{\text{lat}}(x, y)|^2.$$

The latent representation was formed by cropping the central square region with a side length of `latent_size` from this intensity map. The `latent_size` parameter specified the side length of this window in discrete detector-grid pixels; for a value s , the latent map had size $s \times s$ and, after flattening, contained s^2 scalar intensity samples. This central region was then flattened into a vector and fed to the electronic decoder. The single-layer architecture used in the initial experiments corresponds to $N = 1$, while in the layer-analysis experiments the value of N was varied.

From the training perspective, the diffractive encoder was treated as the differentiable physical part of a unified computational-graph model, which corresponds to the general idea of D²NN: the transmittance coefficients of individual points of the diffractive layers are trainable parameters and are selected by backpropagation [4]. Let the parameters of the optical block be denoted by

$$\Theta_{\text{opt}} = \{a_n(x, y), \phi_n(x, y)\}_{n=1}^N,$$

and the parameters of the electronic decoder by θ_{dec} . For an input image x_j , the complete forward pass can be written as

$$z_j = \mathcal{C}_s\{I_{\text{lat},j}\}, \quad \hat{x}_j = f_{\theta_{\text{dec}}}(z_j),$$

where $\mathcal{C}_s\{\cdot\}$ denotes a central crop with side length s followed by vectorization, s corresponds to the `latent_size` parameter, and $f_{\theta_{\text{dec}}}$ is the electronic decoder. Training was performed by minimizing the empirical risk over the training set:

$$\min_{\Theta_{\text{opt}}, \theta_{\text{dec}}} \mathcal{J}(\Theta_{\text{opt}}, \theta_{\text{dec}}) = \frac{1}{|\mathcal{B}|} \sum_{j \in \mathcal{B}} \mathcal{L}_{\text{MS-SSIM}}(\hat{x}_j, x_j),$$

where $\mathcal{B}$ is a mini-batch and $\mathcal{L}_{\text{MS-SSIM}}$ is the loss function associated with multiscale structural similarity. Because the propagation operators $\mathcal{P}_{d_n}$ are linear in the complex field and the modulation $m_n \odot u$ is an elementwise differentiable operation with respect to the mask parameters, the gradient of the loss function can be propagated from the electronic decoder back through the latent-region extraction, free-space propagation, and all diffractive layers. In the numerical implementation, complex fields were processed as pairs

of real and imaginary components, so the optical and electronic parameters were updated by a gradient optimizer according to the chain rule:

$$(\Theta_{\text{opt}}, \theta_{\text{dec}})^{(t+1)} = (\Theta_{\text{opt}}, \theta_{\text{dec}})^{(t)} - \eta \nabla_{(\Theta_{\text{opt}}, \theta_{\text{dec}})} \mathcal{J},$$

where η denotes the optimization step size. In variants with a full amplitude–phase mask, both components a_n and ϕ_n were optimized; in the amplitude-only and phase-only ablations, one of the components was fixed to evaluate the contribution of the corresponding modulation mechanism. After training was completed, the obtained masks were considered fixed, and the optical encoder in inference mode performed only passive propagation and modulation without additional iterative adjustment.

2.3. Coherent and Incoherent Illumination

Two illumination modes were modeled; they differed in how the recorded intensity in the latent plane was formed. In both cases, the input image was treated as a normalized intensity distribution $I_0(x, y)$ rather than as an arbitrary complex field. Before optical propagation, the intensity was converted into amplitude:

$$u_0(x, y) = \sqrt{I_0(x, y)}.$$

The same optical-grid parameters were used for both modes: wavelength $\lambda = 512$ nm, sampling step $\Delta x = 1.024$ μm , and propagation distance between consecutive planes $z = 1$ mm.

In the coherent mode, the field $u_0(x, y)$ propagated once through the diffractive stack with a fixed input phase. The latent representation was formed from the detected intensity rather than from the complex field:

$$I_{\text{coh}}(x, y) = |u_{\text{lat}}(x, y)|^2.$$

The central crop of this map was used as the latent representation for the electronic decoder.

In the incoherent mode, before padding and subsequent propagation, an ensemble of $R = 50$ realizations of a random phase screen was applied to the amplitude $u_0(x, y)$:

$$u_0^{(r)}(x, y) = u_0(x, y) \exp(i2\pi\phi^{(r)}(x, y)), \\ r = 1, \dots, R.$$

Here, $\phi^{(r)}$ is the normalized random phase field; the factor 2π matches how the phase factor is implemented in the code. The phase screens were not specified as independent white noise at each pixel: they were generated in the spectral domain from a random spectrum with a Gaussian mask, followed by distribution normalization. The experiments used spatial and temporal coherence parameters $l_c = 100$ μm and $\tau_c = 10$ ns. Each realization propagated through the same trained D²NN encoder; the propagation geometry and diffractive masks did not change. The incoherent output intensity was then computed by averaging the intensities of individual realizations,

$$I_{\text{inc}}(x, y) = \frac{1}{R} \sum_{r=1}^R |u_{\text{lat}}^{(r)}(x, y)|^2.$$

Averaging was performed specifically at the level of intensities $|u|^2$, not complex fields, because mutual phase relationships between realizations are not observable in this model. For the chosen short coherence time, independent realizations are used as a Monte

Carlo approximation of partially incoherent illumination. The latent region provided to the decoder was extracted from the distribution $I_{\text{inc}}(x, y)$. This ensemble formulation makes it possible to separate the effect of coherence from the effects of latent-region size, propagation geometry, and diffractive-mask parameterization.

2.4. Training Protocol

The hybrid D²NN–ENN model was trained as an autoencoder for image reconstruction. The optical encoder transformed each input image into a square latent representation, and the electronic decoder was optimized to reconstruct the corresponding target grayscale image. The side length of the latent region was selected from the set $\{15, 30, 50, 80, 100, 150, 200, 300\}$ pixels, which made it possible to evaluate the trade-off between representation compactness and reconstruction quality. Each configuration was trained for 15 epochs.

Training was carried out in the coherent and incoherent modes described above. The optimization objective was based on the MS-SSIM loss, which directly encourages perceptual structural similarity between the reconstructed and target images. Validation data were used to monitor training and select the model, and all final metrics for the proposed D²NN–ENN models and reference codecs were computed on the test subset after training was completed.

2.5. Bicubic Downsample–Quantize–Upsample Baseline

In addition to JPEG and learned codecs, we intentionally used a simple non-learned bicubic algorithm rather than another codec. The aim was to provide the most elementary control for spatial subsampling: each test image was deterministically downsampled, stored as an 8-bit low-resolution array, and bicubically upsampled using the same output resolution and quality metrics.

Let $x \in [0, 1]^{400 \times 400}$ denote a preprocessed grayscale test image. For a target compressed spatial size $W \times H$, the baseline encoder first resized the image with bicubic interpolation and then quantized the resulting low-resolution image to 8-bit integer values:

$$c = Q_8\left(\mathcal{D}_{W,H}^{\text{bic}}\{x\}\right), \quad Q_8(v) = \text{round}(255 \text{ clip}(v, 0, 1)),$$

where $\mathcal{D}_{W,H}^{\text{bic}}$ is the bicubic downsampling operator and $c \in \{0, \dots, 255\}^{W \times H}$ is the stored representation. The corresponding file size is therefore

$$B_{\text{bic}} = WH$$

bytes, or $B_{\text{bic}}/1024$ kilobytes in the rate–distortion plots. The decoder dequantized the stored image and restored it to the original spatial resolution using bicubic interpolation:

$$\hat{x}_{\text{bic}} = \mathcal{U}_{400,400}^{\text{bic}}\left\{\frac{c}{255}\right\},$$

where $\mathcal{U}_{400,400}^{\text{bic}}$ is the bicubic upsampling operator. The quality of each baseline point was evaluated with the same metrics as for the other methods,

$$\text{MSE}_{\text{bic}} = \frac{1}{P} \sum_{i=1}^P (x_i - \hat{x}_{\text{bic},i})^2, \quad \text{PSNR}_{\text{bic}} = 10 \log_{10} \left(\frac{1}{\text{MSE}_{\text{bic}}} \right),$$

and MS-SSIM($x, \hat{x}_{\text{bic}}$) was computed using the multiscale procedure described below.

2.6. Evaluation Metrics and Time Measurement

Reconstruction accuracy was evaluated using PSNR and MS-SSIM. For a target image x and reconstruction $\hat{x}$ containing P pixels, the mean squared error was defined as

$$\text{MSE}(x, \hat{x}) = \frac{1}{P} \sum_{i=1}^P (x_i - \hat{x}_i)^2.$$

Because all images were represented as normalized intensities, the maximum signal value was taken as $I_{\max} = 1$, and PSNR was computed as

$$\text{PSNR}(x, \hat{x}) = 10 \log_{10} \left(\frac{I_{\max}^2}{\text{MSE}(x, \hat{x})} \right).$$

MS-SSIM was evaluated between the grayscale target image and the corresponding reconstruction. For the reported test metrics, both images were mapped to the 8-bit intensity scale, so the dynamic range was $L = 255$. At each scale, local means, variances, and covariance were estimated with an 11×11 Gaussian window with $\sigma = 1.5$, and the luminance and contrast–structure terms were computed as

$$l(x, \hat{x}) = \frac{2\mu_x\mu_{\hat{x}} + C_1}{\mu_x^2 + \mu_{\hat{x}}^2 + C_1},$$

$$cs(x, \hat{x}) = \frac{2\sigma_{x\hat{x}} + C_2}{\sigma_x^2 + \sigma_{\hat{x}}^2 + C_2},$$

where $C_1 = (0.01L)^2$ and $C_2 = (0.03L)^2$. The multiscale score was computed over five scales. Between consecutive scales, both images were downsampled by average pooling with a factor of 2. With the standard weights

$$w = (0.0448, 0.2856, 0.3001, 0.2363, 0.1333),$$

which were derived from subjective image-quality calibration across scales in the original MS-SSIM work [25] and are used by the `pytorch-msssim` implementation, the final value was computed as follows. These weights were fixed and were neither learned nor tuned on our dataset:

$$\text{MS-SSIM}(x, \hat{x}) = \left(\prod_{j=1}^4 cs_j^{w_j} \right) l_5^{w_5} cs_5^{w_5},$$

where the luminance term is used only at the coarsest scale. The reported MS-SSIM was averaged over the test subset of 300 images. During training, the same metric was used in the loss 1 – MS-SSIM on normalized floating-point images with $L = 1$.

File size was defined as the number of bytes B required to represent the compressed signal; all plots report $B/1024$ kilobytes. For JPEG and neural codecs, B is the size of the corresponding compressed bitstream. For the bicubic baseline, the compressed representation is a downsampled grayscale image of size $W \times H$ represented with 8-bit pixels,

$$B_{\text{bic}} = WH.$$

For a $D^2\text{NN-ENN}$ latent map with side length $s = \text{latent_size}$, the unquantized representation uses 32-bit floating-point samples, hence

$$B_{32} = 4s^2.$$

After q -bit quantization and bit packing, where $q \in \{8, 4, 2\}$, the size is

$$B_q = \left\lceil \frac{qs^2}{8} \right\rceil,$$

which gives s^2 , $\lceil s^2/2 \rceil$, and $\lceil s^2/4 \rceil$ bytes for 8-, 4-, and 2-bit quantization, respectively.

The reference methods included JPEG, standard neural codecs from the CompressAI benchmark suite, and the bicubic-interpolation baseline. All neural-network training experiments and final evaluations, including the proposed D²NN–ENN models and benchmark measurements, were performed on an NVIDIA H100 GPU. For the electronic reference methods, encoding and decoding times were measured directly on the same test subset. For the proposed hybrid model, the ideal optical-encoding time was evaluated separately: in the main single-layer configuration, light passes through two 1 mm intervals, i.e., approximately 2 mm in total. For such an optical path, the propagation delay is approximately 6.7×10^{-12} s, or 6.7×10^{-9} ms. This is only the physical transit time of light through the encoder; it does not include sensor readout, data transfer, or the operation of the electronic decoder.

2.7. Latent-Representation Quantization

Quantization was applied after training to the latent optical representation. In other words, it was not taken into account during optimization of the diffractive masks and decoder parameters, but was introduced only at the testing and export stages between the optical encoder and the electronic decoder. After propagation through the optical encoder, each input image formed a latent vector or intensity map whose spatial size was determined by the selected `latent_size` value: the latent map had the shape `latent_size × latent_size`. The latent map was then flattened into a vector and fed to the electronic decoder. To assess the effect of limited latent-representation precision, it was quantized before decoding.

Uniform scalar quantization with 8-, 4-, and 2-bit precision was considered. For bit depth b , the number of quantization levels was defined as

$$L = 2^b - 1.$$

In all experiments, normalization of the latent representation before quantization was disabled. Therefore, the same scheme was applied to each latent tensor: values were first clipped to the interval $[0, 1]$ and then rounded to the nearest quantization level:

$$z_{\text{clipped}} = \text{clip}(z, 0, 1), \quad z_q = \frac{\text{round}(z_{\text{clipped}} L)}{L}.$$

The decoder received the dequantized floating-point value z_q , not the integer code itself.

The corresponding bit budget is determined by the number of latent samples and the quantization depth. Compared with the unquantized 32-bit representation, q -bit quantization changes the latent-representation size from $32s^2$ bits to qs^2 bits, or from $4s^2$ bytes to $\lceil qs^2/8 \rceil$ bytes. Thus, reconstruction degradation can be analyzed as a function of latent precision while keeping the trained optical encoder and electronic decoder fixed.

3. Results

3.1. D²NN–ENN Autoencoder Architecture

The proposed architecture is a hybrid D²NN–ENN autoencoder that comprises an optical diffractive encoder and an electronic neural-network decoder (Figure 1). The encoder performs the initial compression: it transforms the input optical field into a compact latent representation. The decoder, in turn, maps this representation back to the image domain and reconstructs the normalized intensity distribution. The key novelty is an image-

compression approach that transfers the encoding stage to a passive diffractive optical processor, enabling ultrafast physical precompression before electronic reconstruction.

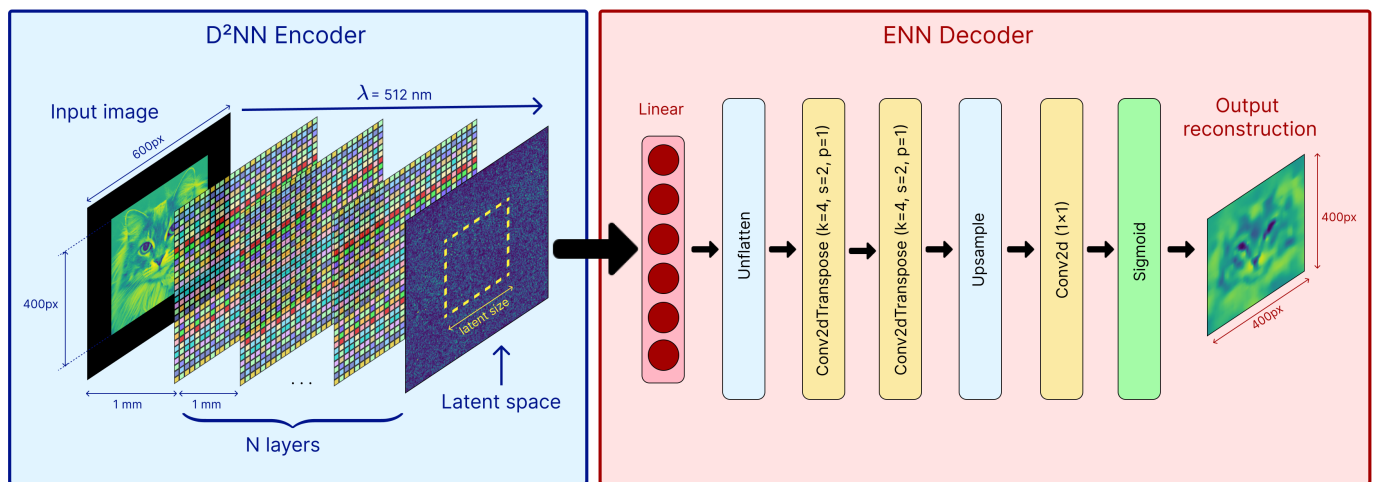


Figure 1. Architecture of the proposed hybrid D^2NN –ENN autoencoder: the diffractive optical encoder forms a compact latent intensity representation, which is reconstructed by the electronic neural-network decoder.

D^2NN encoder. The encoder input is a 400×400 -pixel grayscale image embedded into a 600×600 computational field. The zero boundary region is used to model free-space propagation and simultaneously reduces the effect of diffractive artifacts at the boundaries. The optical field propagates from the input plane to a diffractive amplitude–phase modulation layer located at a distance of 1 mm. The wavelength is 512 nm, and the physical pixel size is 1024 nm. After modulation, the field propagates another 1 mm to the detector plane, which records the latent map. At this plane, a square region of a specified size is cropped from the propagated field and used as the compressed latent representation. In this work, this block is generalized to an N -layer diffractive stack: the number of diffractive layers N , placed at equal distances, is treated as a tunable architectural parameter, and the single-layer encoder corresponds to the special case $N = 1$.

ENN decoder. The decoder receives the latent representation formed by the D^2NN encoder and reconstructs the image using electronic neural-network operations. First, the latent vector is processed by a multilayer perceptron; a ReLU nonlinearity is applied after each of its hidden layers. The resulting vector is then converted by an unflattening operation into a tensor of size $B \times 256 \times 10 \times 10$, where B denotes the batch size. The spatial resolution is increased by two transposed convolutional layers with kernel size $k = 4$, stride $s = 2$, and padding $p = 1$: these layers increase the feature-map size first to 20×20 and then to 40×40 , while reducing the number of channels from 256 to 128 and then to 64; ReLU is also applied after each transposed convolutional layer. In the final stage, bilinear interpolation expands the representation to 400×400 , a 1×1 convolution maps the feature tensor to one output channel, and a sigmoid activation constrains the reconstructed image to the normalized range of pixel intensities. Decoder variants with a larger number of convolutional layers were also tested; however, they did not yield a noticeable increase in PSNR or MS-SSIM, so a more compact architecture was used in the main scheme.

3.2. Comparison of D^2NN –ENN Compression Quality and Time with Standard Codecs

Figure 2 shows the main trade-off considered in this work: reconstruction quality and execution time are compared with the actual size of the compressed representation. The plots compare the proposed D^2NN –ENN model, JPEG, standard neural codecs, and the

bicubic-interpolation baseline. For D²NN-ENN, each point corresponds to a specific `latent_size` value, and the numerical labels next to the points denote the `latent_size` value in pixels. As the side length of the latent region increases from very compact representations to intermediate values, more optical information is preserved, improving both PSNR and MS-SSIM. In the coherent mode, quality increases from 18.22 to a maximum of 20.91 dB PSNR, while MS-SSIM increases from 0.57 to 0.68. In the incoherent mode, the same qualitative trend is observed, but the maximum values are slightly lower: for latent side lengths of 15–200 pixels, PSNR increases from 18.29 to 20.49 dB, and MS-SSIM increases from 0.56 to 0.66. Beyond intermediate sizes, the incoherent curve reaches a plateau or slightly decreases, so further enlargement of the latent region does not provide a proportional gain in quality.

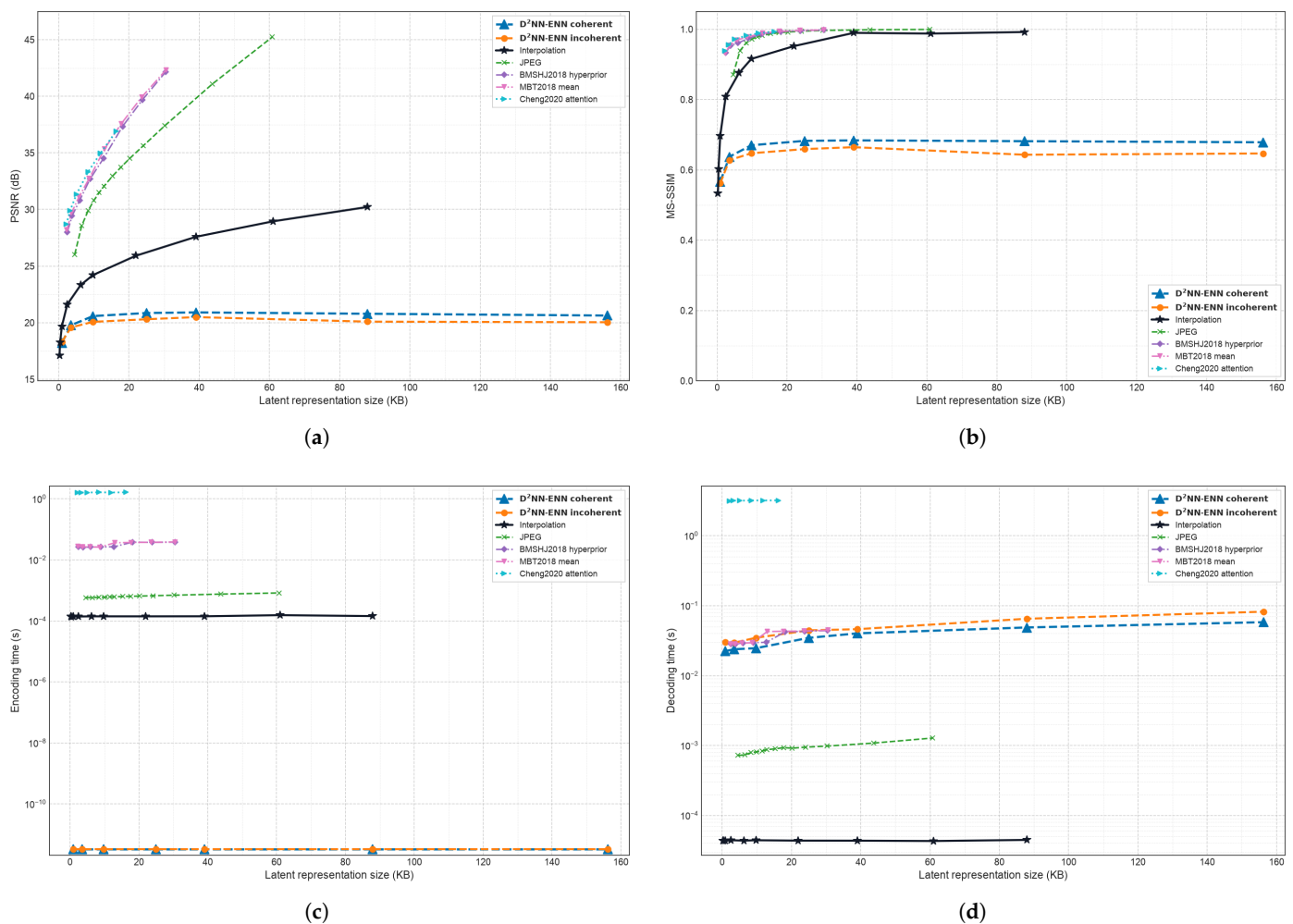


Figure 2. Comparison of reconstruction quality and computational time for D²NN-ENN, JPEG, standard neural codecs, and bicubic interpolation: (a) PSNR; (b) MS-SSIM; (c) encoding time; (d) decoding time. The numbers next to the D²NN-ENN points denote the `latent_size` value in pixels.

The image-level comparison in Figure 3 complements the aggregate results by showing the characteristic artifacts of the evaluated methods. The optical reconstructions preserve the global position and shape of the face but strongly smooth fine fur texture and local contrast. The conventional and learned codecs retain progressively more high-frequency detail as their compressed representation grows, whereas interpolation mainly reflects the loss caused by spatial subsampling. This example also illustrates that similar compressed sizes can correspond to visibly different error structures because the methods encode different types of image information.



Figure 3. Visual comparison of a representative test image reconstructed at different compression operating points using coherent and incoherent D²NN-ENN, JPEG, JPEG 2000, the BMSHJ2018, MBT2018, and Cheng2020 learned codecs, and bicubic interpolation. The labels report the compressed size, PSNR, and MS-SSIM for each reconstruction.

More detailed numerical values for the D²NN-ENN operating points are given in Tables 1 and 2. They correspond to the same latent sizes annotated in Figure 2 and show how increasing file size affects reconstruction quality and electronic decoding time in the coherent and incoherent modes. The encoding time in these tables refers to software-based numerical testing of the optical encoder; the ideal hardware latency of passive optical propagation is discussed separately below.

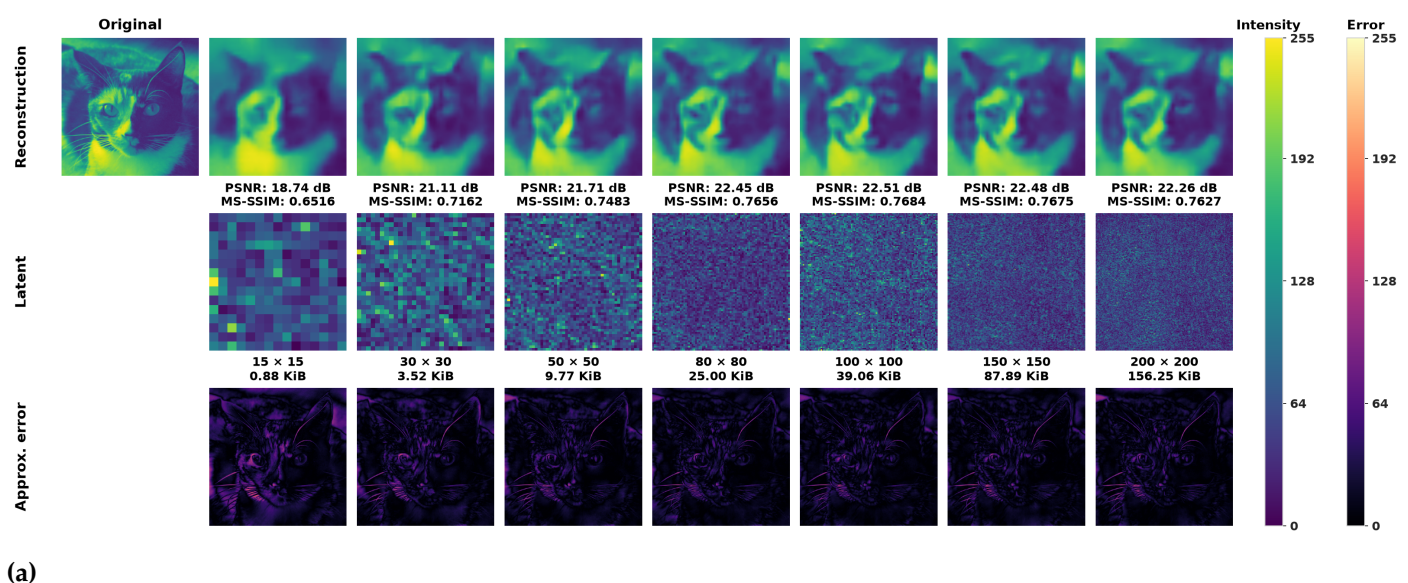
Table 1. Test results for the hybrid D²NN–ENN architecture under coherent illumination.

Latent Size (Pixels)	File Size (KB)	PSNR (dB)	MS-SSIM	Encoding Time (ms)	Decoding Time (ms)
15	0.88	18.22	0.57	29.37	22.39
30	3.52	19.78	0.64	33.02	23.81
50	9.77	20.58	0.67	33.48	24.67
80	25.00	20.85	0.68	39.33	34.66
100	39.06	20.91	0.68	35.96	40.17
150	87.89	20.79	0.68	32.54	48.78
200	156.25	20.63	0.68	31.20	58.01

Table 2. Test results for the hybrid D²NN–ENN architecture under incoherent illumination.

Latent Size (Pixels)	File Size (KB)	PSNR (dB)	MS-SSIM	Encoding Time (ms)	Decoding Time (ms)
15	0.88	18.29	0.56	1110.03	29.98
30	3.52	19.58	0.63	1109.77	29.63
50	9.77	20.07	0.65	1136.47	34.77
80	25.00	20.30	0.66	1135.83	44.38
100	39.06	20.49	0.66	1149.26	46.11
150	87.89	20.09	0.64	1109.41	65.02
200	156.25	20.04	0.65	1111.83	81.97

The reconstruction examples in Figure 4 are consistent with the quantitative results. For the smallest latent sizes, the decoder mainly preserves the coarse spatial structure of animal faces, whereas fine texture, sharp contours, and high-frequency details are smoothed. Increasing the latent region improves the transfer of overall shape and contrast, but this effect is limited: after the intermediate range, additional latent samples do not yield a proportional visual gain. Coherent illumination provides slightly sharper and more structurally accurate reconstructions, whereas incoherent ensemble averaging makes the images smoother and reduces local contrast. This visual behavior is consistent with the PSNR and MS-SSIM trends and shows that quality is limited not only by the number of latent samples, but also by the information selected by the optical transform and by the ability of the electronic decoder to invert this transform.

**Figure 4.** Cont.

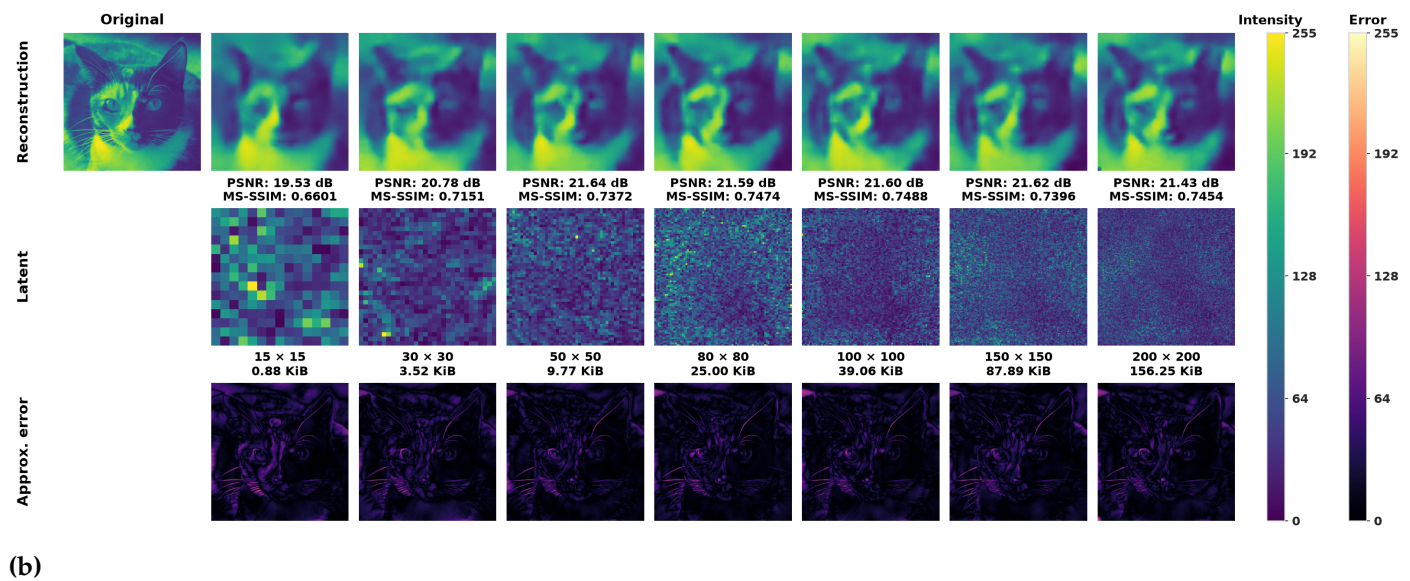


Figure 4. Examples of reconstructed images obtained in (a) coherent and (b) incoherent optical modes.

Figure 5 demonstrates that the model generalizes across different animal classes, poses, brightness distributions, and background structures. Both illumination modes preserve the dominant facial geometry, including the head outline, eyes, muzzle, and major contrast regions. The error maps show that most residual errors are concentrated around fine fur, object boundaries, and high-contrast local features. The relatively weak difference between the coherent and incoherent outputs confirms that the two modes usually produce similar coarse reconstructions, although the coherent model tends to retain slightly stronger local contrast.

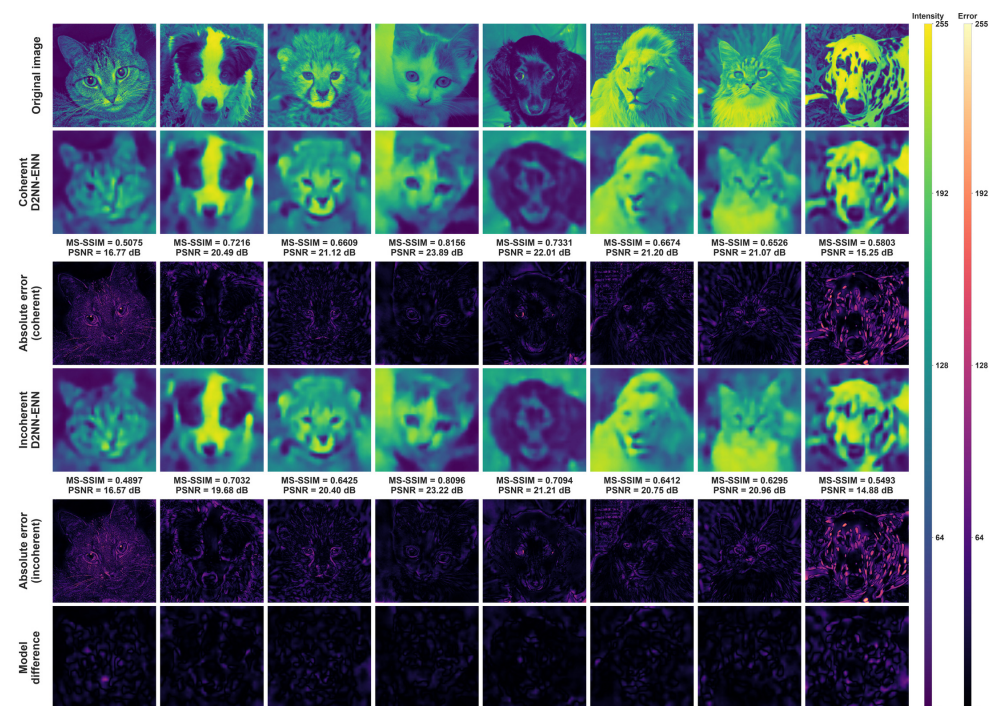


Figure 5. Representative test-set reconstructions for a 100 × 100-pixel latent region. From top to bottom, the rows show the original images, coherent D²NN-ENN reconstructions, coherent absolute errors, incoherent D²NN-ENN reconstructions, incoherent absolute errors, and the absolute difference between the coherent and incoherent reconstructions.

Most of the poorest reconstructions in Figure 6 are tigers. Their high-contrast stripes and spots create a strongly structured texture that is difficult to preserve in the compact latent representation. The error maps show that these markings contribute substantially to the reconstruction error, although the overall silhouette remains recognizable. Similar coherent and incoherent error patterns indicate that the main limitation is the information bottleneck and decoder capacity rather than illumination mode.

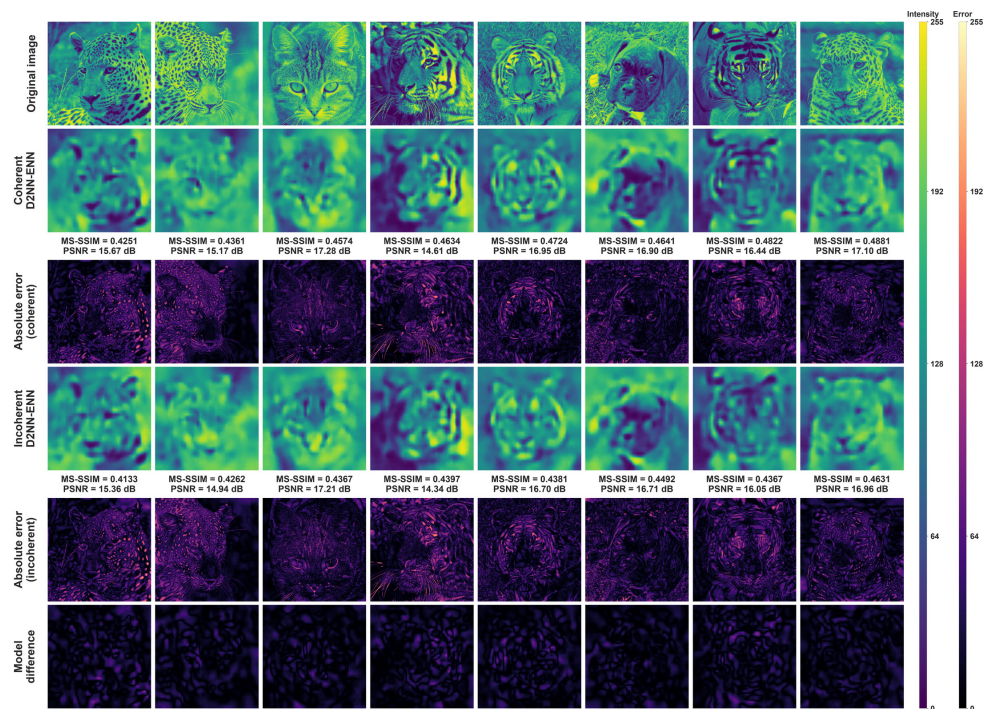


Figure 6. Representative failure cases for a 100×100 -pixel latent region. The row order is the same as in Figure 5: original images, coherent reconstructions and absolute errors, incoherent reconstructions and absolute errors, and the difference between the two reconstruction modes.

3.3. The Effect of Encoder Depth

For the same D²NN-ENN architecture, we analyzed how the number of diffractive layers affects reconstruction quality and whether both components of the complex modulation mask are indeed necessary. In this experiment, the electronic decoder and the overall training procedure remained unchanged, while the optical encoder was varied by changing the number of diffractive layers and choosing one of three parameterizations: a full amplitude–phase mask, an amplitude-only mask, or a phase-only mask.

The results in Figure 7 show that the full amplitude–phase parameterization provides the best overall performance, which is consistent with its larger optical-modulation capacity. At the same time, the phase-only configuration also remains competitive and yields comparatively high reconstruction quality. This is an important observation: it shows that a simpler diffractive element controlling only phase is still capable of forming an informative latent representation for the electronic decoder. By contrast, the amplitude-only configuration performs worse in the tested setting, indicating lower efficiency of amplitude modulation without trainable phase control for this compression and reconstruction task.

The dependence of quality on the number of layers is nonmonotonic. The best reconstruction metrics are achieved with two and three diffractive layers, whereas further increasing the depth does not provide a stable improvement. For the same encoder depth and mask parameterization, the incoherent mode follows the same qualitative trend but gives slightly lower PSNR and MS-SSIM values than the coherent mode. This indicates that, for the considered architecture and dataset, adding a small number of diffractive

transforms increases the expressiveness of the optical encoder, but excessive depth does not automatically lead to better compression and may complicate optimization.

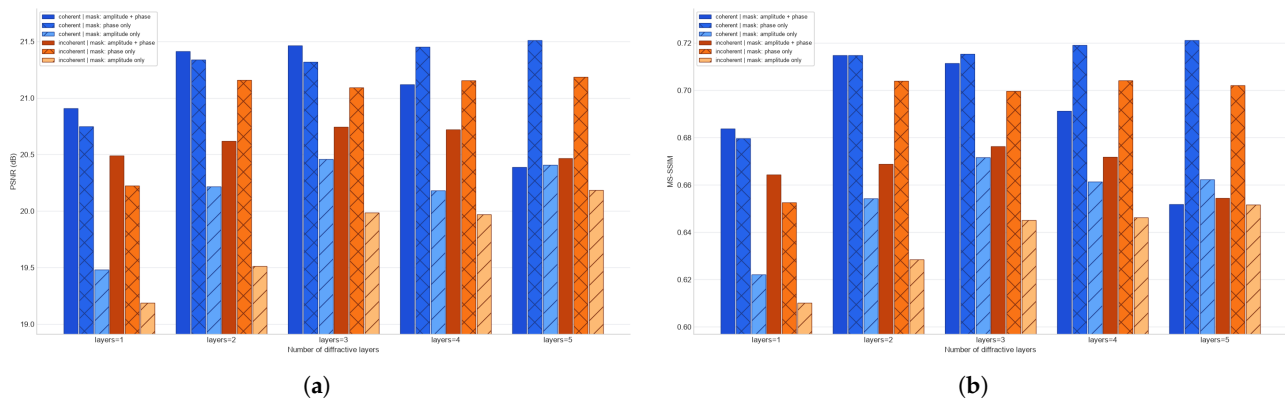


Figure 7. Effect of the number of diffractive encoder layers and modulation-mask parameterization on reconstruction quality: (a) PSNR; (b) MS-SSIM.

3.4. Robustness to Quantization

Figure 8 shows that the trained latent representation is sufficiently robust to moderate post-training quantization. The transition from the original floating-point representation to 8-bit and 4-bit precision causes only a limited decrease in PSNR and MS-SSIM. This indicates that, at these quantization levels, most of the information used by the decoder is preserved. The examples in Figure 9 show the same trend visually for a 50×50 -pixel latent region in the coherent and incoherent modes: at 8- and 4-bit precision, the overall shape and contrast of the reconstructions are largely preserved, whereas the 2-bit representation smooths details more noticeably and enhances artifacts. Consequently, the practical gain from quantizing the latent representation is limited by the transition to very low precision, while 4-bit and 8-bit representations remain comparatively stable for the tested architecture. Figure 10 also compares the 4-bit quantized D²NN-ENN with JPEG, standard neural codecs, and bicubic interpolation in the same setting as the codec-comparison section, using PSNR, MS-SSIM, encoding time, and decoding time. An important feature of this variant is that the D²NN-ENN latent representation uses 4-bit precision: one latent sample requires eight times fewer bits than in a 32-bit floating-point representation, so the strength of the method becomes a very compact compressed representation while preserving the main image structure.

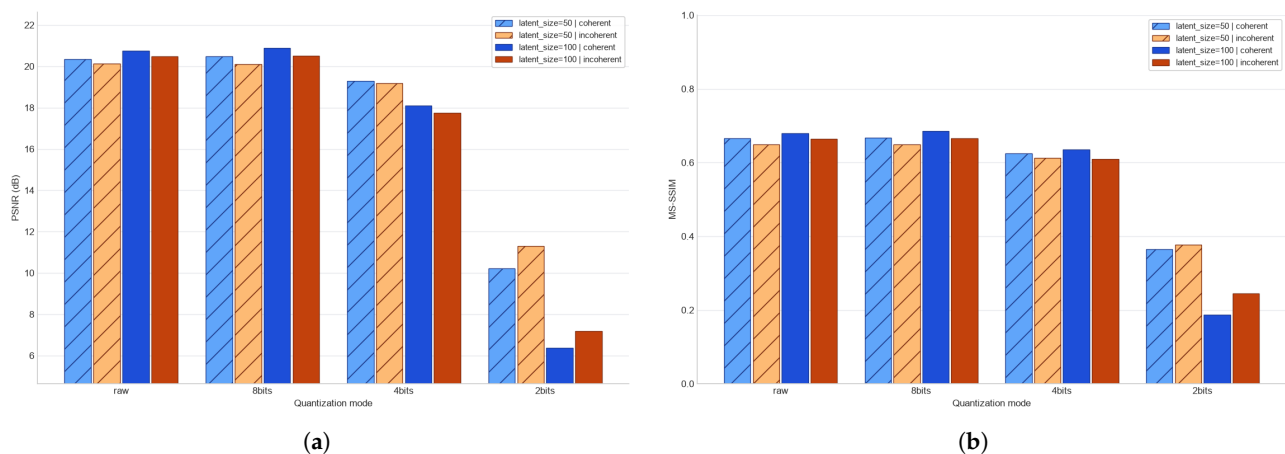


Figure 8. Robustness of reconstruction quality to latent-representation quantization: (a) PSNR; (b) MS-SSIM.

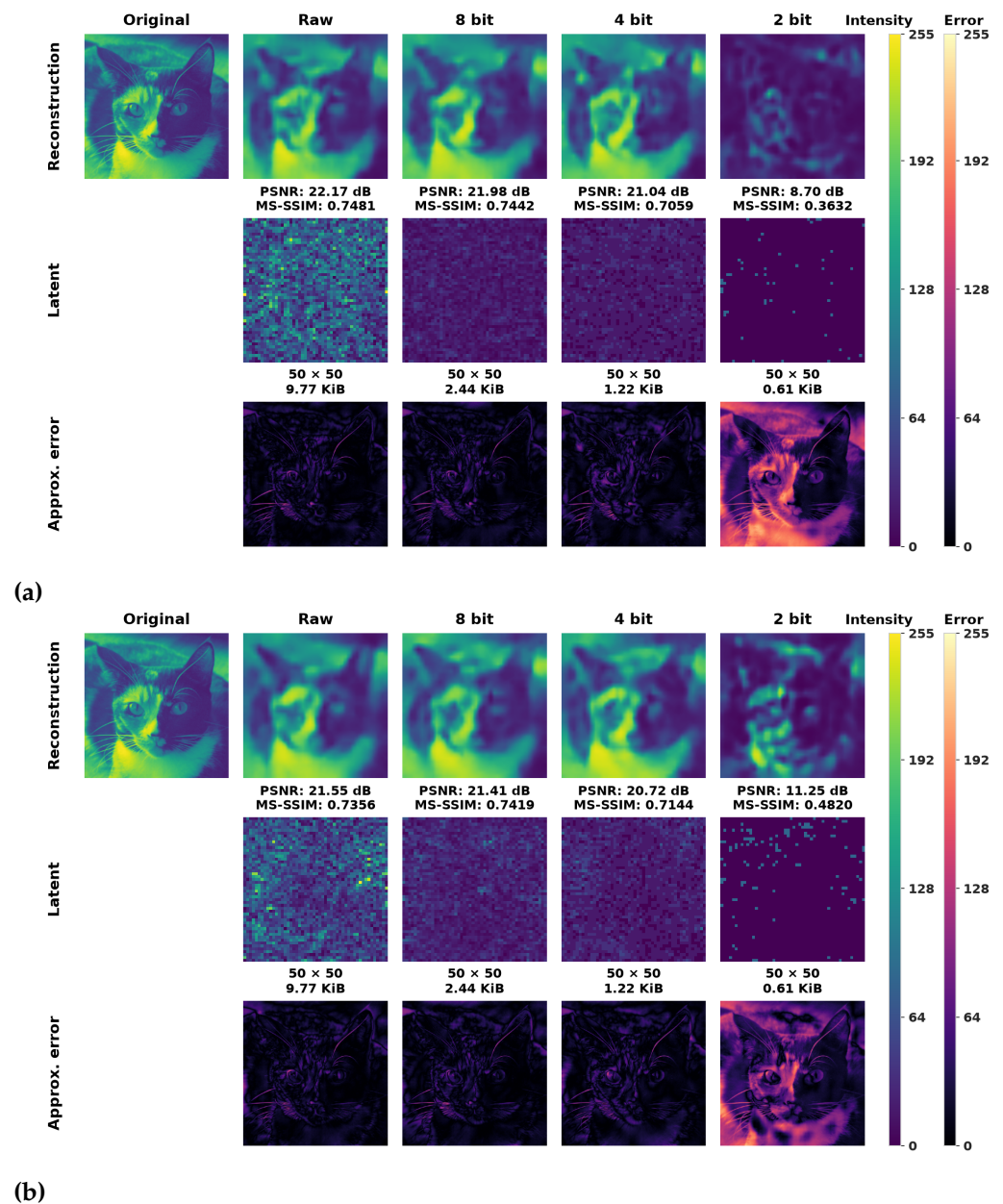


Figure 9. Examples of reconstructed images at different quantization levels of the latent representation for a 50×50 -pixel latent region: (a) coherent optical mode; (b) incoherent optical mode.

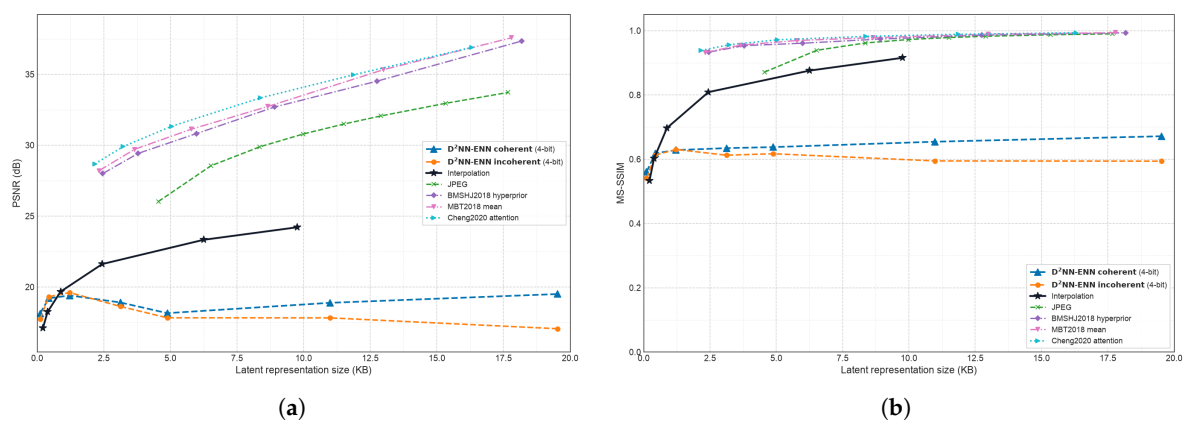


Figure 10. Cont.

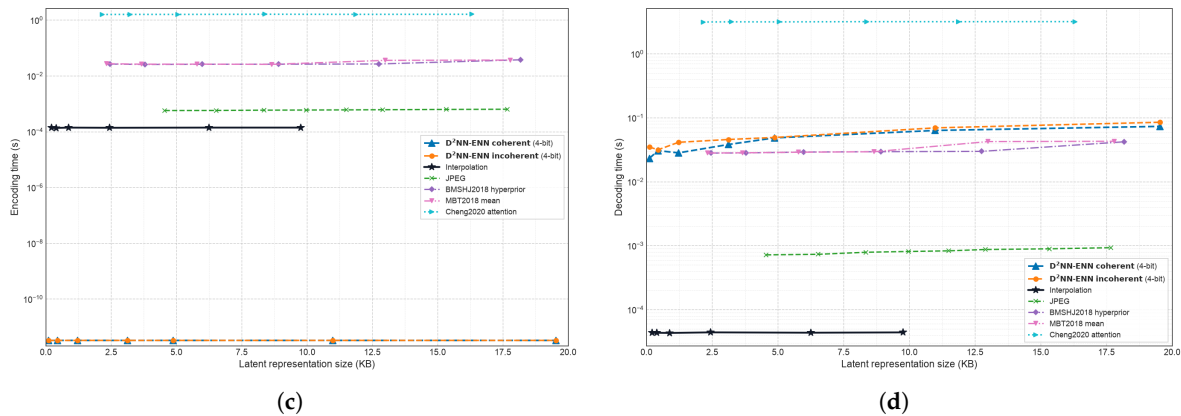


Figure 10. Comparison of the 4-bit quantized version of D²NN-ENN with JPEG, standard neural codecs, and bicubic interpolation in terms of reconstruction quality and execution time: (a) PSNR; (b) MS-SSIM; (c) encoding time; (d) decoding time. The numbers next to the D²NN-ENN points denote the latent_size value in pixels.

4. Discussion

A comparison with electronic codecs shows that these codecs provide substantially higher PSNR and MS-SSIM at comparable file sizes. This is expected: such codecs combine transform coding, quantization, and entropy coding, whereas the proposed optical encoder forms a compact discretized latent representation that is quantized only afterward. Therefore, in its current form, the D²NN-ENN architecture should not be considered a direct replacement for optimized digital codecs for general-purpose image storage. Its advantage appears in the execution-time plots: for the incoherent mode, the ideal encoding time remains on the order of 6.7×10^{-12} s for all latent sizes, whereas the fast neural codecs BMSHJ2018 and MBT2018 encode images on timescales of about 10^{-2} s, and the heavier Cheng2020 attention model operates on a timescale of about 10^0 s. Optical formation of the latent representation is therefore approximately 9–10 orders of magnitude faster than the first two neural codecs and more than 11 orders of magnitude faster than the heavier attention model. Decoding time, by contrast, is determined by the electronic decoder: in the incoherent mode, it is 29.63–81.97 ms for the considered latent sizes, i.e., of the same order as BMSHJ2018 and MBT2018 decoding, although it increases as the latent representation grows.

These observations show a clear separation between advantages and limitations. Standard electronic codecs dominate the quality–file-size curves, but the proposed hybrid model belongs to a different engineering regime: it is oriented toward applications in which the encoder is located directly at the input of the sensor system, and ultrafast formation of a compact representation is more important than achieving the best quality at a given file size. In such scenarios, especially in task-oriented imaging systems and NIR sensing, the optical encoder can be considered a passive physical precompressor whose output is then refined by an electronic decoder.

In this sense, the hybrid D²NN-ENN scheme is not a direct analogue of a classical digital codec. Its latent representation arises from a physical optical transform and can only then be quantized and stored in digital form. Therefore, comparison with electronic codecs should be interpreted not as an attempt to replace mature general-purpose codecs, but as an evaluation of an alternative operating regime in which part of the compression is performed before full electronic processing.

The main advantage of this regime is associated with the speed of latent-representation formation. If the optical encoder is integrated directly with the image-acquisition system, encoding can occur through passive light propagation with a latency close to the time of

flight through the optical system. The gain therefore arises specifically at the encoding stage, where latency is reduced by several orders of magnitude relative to digital neural codecs; the decoding stage remains electronic and is closer in time to conventional neural-network inference. At the same time, the final reconstruction quality is determined not only by the latent-region size, but also by how informative a transform the diffractive stack implements, how well the electronic decoder can invert this transform, and how realistically hardware constraints are taken into account.

The proposed approach has several limitations that define directions for future work. First, the current study is based on idealized numerical propagation and does not explicitly model sensor noise, finite detector dynamic range, wavelength variations, fabrication errors, or lateral and axial misalignment of the diffractive elements. Future implementations should therefore use hardware-aware training, in which such perturbations are introduced during optimization, followed by experimental calibration and validation of the fabricated optical encoder.

Second, scalability to higher-resolution images remains an important issue. The present experiments use 400×400 -pixel inputs, and increasing the spatial resolution also increases the computational cost of numerical propagation during training and the required resolution of the diffractive elements and detector. Tiled or hierarchical optical encoding, multiscale latent representations, and more efficient propagation models may help address this problem.

Third, the compression ratio is controlled by the size of the cropped central latent region. A future model could form a nested latent representation, allowing different crop sizes to provide different compression–quality operating points without redesigning the optical encoder.

Finally, the framework can be extended beyond grayscale images to RGB, multispectral, hyperspectral, amplitude, and phase data. In these systems, the diffractive front end could perform physical precompression directly at the sensor plane.

5. Conclusions

This work examined a hybrid D²NN–ENN autoencoder as an image-compression architecture in which the first stage of representation formation is performed optically. The main goal was not to replace mature digital codecs directly, but to evaluate whether a diffractive optical front end can form a compact latent representation that remains informative enough for electronic reconstruction.

The experiments show that the proposed system exhibits a clear trade-off between latent size and reconstruction quality. Increasing `latent_size` improves PSNR and MS-SSIM up to an intermediate range, after which the gain becomes limited. Coherent illumination gives slightly better reconstructions than incoherent illumination, with maximum values of 20.91 dB and 0.68 MS-SSIM versus 20.49 dB and 0.66 MS-SSIM. The best encoder depths in the tested configuration are two and three diffractive layers. Full amplitude–phase modulation provides the strongest performance, while phase-only modulation remains a practical lower-complexity alternative. The latent representation is also moderately robust to quantization: 8-bit and 4-bit precision preserve most of the reconstruction quality, whereas 2-bit precision introduces visible degradation.

The main significance of the approach lies in the encoding stage. In the idealized optical model, latent formation occurs through passive propagation with latency on the order of 10^{-12} s, which is many orders of magnitude shorter than digital neural-codec encoding. This makes the method most relevant for specialized sensing systems in which data should be preprocessed directly at the optical or sensor plane. Such a regime is especially natural for visible, near-infrared, and infrared imaging, because the same diffractive-processing

principle can be adapted to different wavelength ranges by choosing suitable optical materials, pixel sizes, and propagation geometries.

The present implementation also has important limitations. Its rate–distortion performance is still below that of classical and learned codecs; the model does not include an entropy model; and the simulations do not yet fully account for sensor noise, finite detector dynamic range, fabrication errors, or optical misalignment. Future work should therefore include hardware-aware training, stronger electronic decoders, realistic sensor and noise models, entropy coding of the quantized latent representation, and experimental validation with physical diffractive elements.

A further direction is secure optical preprocessing. Since the measured latent representation depends on the trained diffractive masks and is not directly interpretable as an image, the optical encoder could potentially combine compression with physical-layer obfuscation. In such a system, the diffractive structure and the matched electronic decoder would act as components of a task-specific key: without the correct optical transform and decoder, the latent measurement may be difficult to reconstruct. This should not yet be treated as proven cryptographic security, but it suggests a promising direction for optical compression, privacy-preserving sensing, and infrared imaging systems where fast preprocessing and restricted access to visual information are both important.

Author Contributions: Conceptualization, V.V.K., A.V.C. and A.A.G.; methodology, S.S.S. and V.V.K.; software, S.S.S.; validation, S.S.S.; formal analysis, S.S.S.; investigation, S.S.S.; writing—original draft preparation, S.S.S.; writing—review and editing, V.V.K. and A.A.F.; supervision, A.A.F. and A.V.C.; funding acquisition, A.A.G.; project administration, A.A.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Ministry of Economic Development of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000C313925P4H0002; grant No. 139-15-2025-012).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: A publicly available dataset was analyzed in this study. The Animal Faces-HQ (AFHQ) dataset was derived from the following resources available in the public domain: <https://www.kaggle.com/datasets/andrewmvd/animal-faces> (accessed on 28 July 2026).

Acknowledgments: The authors thank Egor Yu. Belashov for providing the code implementing the mathematical model of the D²NN diffractive layers.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

D ² NN	Diffractive deep neural network
ENN	Electronic neural network
MS-SSIM	Multiscale structural similarity index measure
NIR	Near-infrared
PSNR	Peak signal-to-noise ratio

References

1. Wallace, G.K. The JPEG still picture compression standard. *Commun. ACM* **1991**, *34*, 30–44. <https://doi.org/10.1145/103085.103089>.
2. Bross, B.; Chen, J.; Ohm, J.-R.; Sullivan, G.J.; Wang, Y.-K. Developments in international video coding standardization after AVC, with an overview of Versatile Video Coding (VVC). *Proc. IEEE* **2021**, *109*, 1463–1493. <https://doi.org/10.1109/JPROC.2020.3043399>.

3. ISO/IEC 6048-1:2025; Information Technology—JPEG AI Learning-Based Image Coding System—Part 1: Core Coding System. ISO/IEC: Geneva, Switzerland, 2025.
4. Lin, X.; Rivenson, Y.; Yardimci, N.T.; Veli, M.; Luo, Y.; Jarrahi, M.; Ozcan, A. All-optical machine learning using diffractive deep neural networks. *Science* **2018**, *361*, 1004–1008. <https://doi.org/10.1126/science.aat8084>.
5. Kimlichenko, E.O.; Batanov, D.S.; Ivanov, D.A.; Domaratskiy, K.S.; Shmyrin, N.S.; Solokhina, E.E.; Bayslanova, L.R.; Chetvertukhin, A.V.; Grunin, A.A.; Fedyanin, A.A. Optical processing units vs. GPUs and TPUs: Conditions for performance and energy advantage. *Opt. Express* **2026**, *34*, 36260–36277. <https://doi.org/10.1364/OE.604976>.
6. Li, J.; Rahman, M.S.; Ozcan, A. Class-specific differential detection in diffractive optical neural networks improves inference accuracy. *Adv. Photonics* **2019**, *1*, 046001. <https://doi.org/10.1117/1.AP.1.4.046001>.
7. Kulce, O.; Mengü, D.; Rivenson, Y.; Ozcan, A. All-optical information-processing capacity of diffractive surfaces. *Light. Sci. Appl.* **2021**, *10*, 25. <https://doi.org/10.1038/s41377-020-00439-9>.
8. Filipovich, M.; Malyshev, A.; Lvovsky, A.I. Diffractive optical neural networks with arbitrary spatial coherence. In Proceedings of the First Workshop on Machine Learning with New Compute Paradigms at NeurIPS, New Orleans, LA, USA, 16 December 2023.
9. Leonard, J.; Masson, C.; Larger, L.; Goedgebuer, J.-P.; Jacquot, M. Role of depth in optical diffractive neural networks. *Phys. Rev. Appl.* **2024**, *22*, 014048. <https://doi.org/10.1103/PhysRevApplied.22.014048>.
10. Mengü, D.; Luo, Y.; Rivenson, Y.; Ozcan, A. Analysis of diffractive optical neural networks and their integration with electronic neural networks. *IEEE J. Sel. Top. Quantum Electron.* **2020**, *26*, 3700114. <https://doi.org/10.1109/JSTQE.2019.2921376>.
11. Chang, J.; Sitzmann, V.; Dun, X.; Heidrich, W.; Wetzstein, G. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Sci. Rep.* **2018**, *8*, 12324. <https://doi.org/10.1038/s41598-018-30619-y>.
12. Chen, Y.; Zhou, T.; Yu, Y.; Fang, Z.; Johnson, S.G.; Fang, L. Unsupervised photonic variational autoencoder for high-throughput and low-latency image transmission. *Sci. Adv.* **2023**, *9*, eadf8437. <https://doi.org/10.1126/sciadv.adf8437>.
13. Isil, C.; Mengü, D.; Zhao, Y.; Li, J.; Rivenson, Y.; Ozcan, A. Super-resolution image display using diffractive decoders. *Sci. Adv.* **2022**, *8*, eadd3433. <https://doi.org/10.1126/sciadv.add3433>.
14. Rahman, M.S.S.; Gan, T.; Deger, E.A.; Isil, C.; Jarrahi, M.; Ozcan, A. Learning diffractive optical communication around arbitrary opaque occlusions. *Nat. Commun.* **2023**, *14*, 6830. <https://doi.org/10.1038/s41467-023-42556-0>.
15. Li, J.; Mengü, D.; Yardimci, N.T.; Luo, Y.; Li, X.; Rivenson, Y.; Ozcan, A. Spectrally encoded single-pixel machine vision using diffractive networks. *Sci. Adv.* **2021**, *7*, eabd7690. <https://doi.org/10.1126/sciadv.abd7690>.
16. Bai, B.; Lee, R.; Li, Y.; Gan, T.; Wang, Y.; Jarrahi, M.; Ozcan, A. Information-hiding cameras: Optical concealment of object information into ordinary images. *Sci. Adv.* **2024**, *10*, eadn9420. <https://doi.org/10.1126/sciadv.adn9420>.
17. Wang, X.; Redding, B.; Karl, N.; Long, C.; Zhu, Z.; Skowronek, J.; Pang, S.; Brady, D.; Sarma, R. Integrated photonic encoder for low power and high-speed image processing. *Nat. Commun.* **2024**, *15*, 4510. <https://doi.org/10.1038/s41467-024-48099-2>.
18. Fomin, G.V.; Batanov, D.S.; Kimlichenko, E.O.; Shmyrin, N.S.; Suetin, G.M.; Chetvertukhin, A.V.; Grunin, A.A.; Fedyanin, A.A. Promise and pitfalls of random projection algorithms for optical processors. *Opt. Express* **2026**, *34*, 24098–24113. <https://doi.org/10.1364/OE.599206>.
19. Feng, P.; Tan, Y.; Chong, M.; Li, L.; Zhang, Z.; Liu, F.; Wen, Y.; Tan, Y. All-optical autoencoder machine learning framework using linear diffractive processors. *Laser Photonics Rev.* **2025**, *19*, 2401945. <https://doi.org/10.1002/lpor.202401945>.
20. Souza, M.; Zheng, Y.; Kang, K.; Mishra, Y.N.; Fu, Q.; Heidrich, W. Latent space imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 11–15 June 2025*; IEEE: New York, NY, USA, 2025; pp. 28295–28305.
21. Kim, M.; Kim, Y.; Park, W.I. Image processing with optical matrix vector multipliers implemented for encoding and decoding tasks. *Light. Sci. Appl.* **2025**, *14*, 248. <https://doi.org/10.1038/s41377-025-01904-z>.
22. Wu, S.-H.; Latifi, S.M.; Lin, C.-W.; Yang, S.-H. Terahertz compressive optical–digital neural network imaging. *APL Photonics* **2025**, *10*, 090801. <https://doi.org/10.1063/5.0281838>.
23. Xu, W.; Wu, B.; Zhang, S.; Zhou, H.; Wang, Y.; Dong, J.; Zhang, X. On-chip input-hidden-layer-degenerate optical diffractive nonlinear neural network. *Optica* **2026**, *13*, 172–179. <https://doi.org/10.1364/OPTICA.578963>.
24. Choi, Y.; Uh, Y.; Yoo, J.; Ha, J.-W. StarGAN v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020*; IEEE: New York, NY, USA, 2020; pp. 8188–8197.
25. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multiscale structural similarity for image quality assessment. In *Proceedings of the 37th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 9–12 November 2003*; Volume 2, pp. 1398–1402. <https://doi.org/10.1109/ACSSC.2003.1292216>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.