photonics

MDPI

Article

Robust Lightweight Mode Decomposition Based on Geometric Normalization and Multi-Scale Feature Extraction

Qingqiang Xiong [1], Haoyu Wang [2], Zhiyong Liu [2], Linya Jing [1], Binlei Yang [2], Yalong Wen [2], Teng Wang [1,*] and Jiancun Zuo [1,*]

[1] School of Computer and Information Engineering, Institute for Artificial Intelligence, Shanghai Polytechnic University, Shanghai 201209, China
[2] School of Intelligent Manufacturing and Control Engineering, Shanghai Polytechnic University, Shanghai 201209, China
* Correspondence: wangteng@sspu.edu.cn (T.W.); jczuo@sspu.edu.cn (J.Z.)

Abstract

Mode Decomposition (MD) is an important tool for characterizing the modal energy composition and mode states. However, existing deep learning-based approaches have primarily focused on prediction accuracy under ideal conditions, while the stability of their performance under varying input conditions has received considerably less attention. To address non-ideal perturbations, including translation, scaling, rotation, blur, and noise, we propose a lightweight and robust MD method based on RepViT. Several robustness-oriented feature extraction modules are introduced to enhance local–global spatial, frequency-domain, multi-scale, and spatial-channel representations. Moment-Based Geometric Normalization and BlurPool are further introduced to improve the stability of the model against geometric variations and downsampling-induced sampling sensitivity. In addition, a robustness evaluation framework is established using a Clean test set and five graded composite perturbation levels, together with individual-perturbation tests, the mean composite–perturbation correlation coefficient, and the relative correlation degradation rate. Experimental results show that the proposed method achieves a Pearson correlation coefficient of $R = 0.9317$ even under the strongest composite perturbation. Across the five composite perturbation levels, the mean correlation coefficient reaches 0.9728, while the relative correlation degradation rate is only 2.66%. These results demonstrate that the proposed method substantially improves the stability of mode purity prediction under diverse perturbation conditions while maintaining a lightweight architecture, providing a practical reference for robust MD and deployment in fiber-optic systems.

Keywords: mode decomposition; deep learning; robustness; composite perturbation

Check for updates

Received: 13 October 2026
Revised: 4 November 2026
Accepted: 10 November 2026
Published: 12 November 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) license.

1. Introduction

Mode-division multiplexing improves the transmission capacity of optical communication systems by exploiting multiple spatial modes in few-mode fibers (FMFs) [1]. The modal energy distribution provides an important description of the output field state [2], making accurate mode decomposition (MD) essential for mode-state monitoring, mode-coupling analysis, and subsequent adaptive control. Previous studies have investigated mode coupling and transmission control in multimode fibers using wavefront-shaping and analytical approaches [3–5], highlighting the complexity of modal evolution during fiber propagation. Conventional mode analysis generally relies on interferometric measurement, digital holography, correlation filtering, or numerical optimization [6–8]. Although these methods

have clear physical foundations, they often require additional optical components, multiple measurements, or iterative optimization, which increases system complexity and limits real-time applicability.

With the development of deep learning, directly learning the nonlinear mapping from near-field intensity images to modal parameters has become an important direction in MD [9–18]. An et al. demonstrated CNN-based modal-parameter prediction from a single near-field intensity image in 2019 [9]. Subsequent studies improved ambiguity mitigation, supported mode number, computational efficiency, multi-task learning, and degenerate-mode analysis [10–13,15]. More recently, lightweight architectures have also been explored for efficient MD: Zhao et al. proposed a MobileNetV3-Light-based method in 2024 [19], while Wang et al. achieved model compression and real-time inference through traceable knowledge distillation in 2026 [14]. Despite these advances, most existing studies have primarily focused on prediction accuracy under ideal or predefined input conditions.

In practical imaging systems, however, beam displacement, magnification variation, coordinate rotation, defocus, and detector noise can alter the position, scale, orientation, sharpness, and local intensity distribution of near-field patterns. Consequently, a model with high accuracy under Clean conditions may still produce unstable purity estimates when acquisition conditions vary, thereby affecting mode-state identification, coupling analysis, and closed-loop control. As prediction performance under ideal conditions approaches saturation, robustness to non-ideal imaging variations becomes another important dimension for evaluating the reliability of MD.

Several studies have begun to investigate this issue. Fan et al. and Gao et al. analyzed the influence of noise and signal-to-noise ratio on MD in 2020 and 2021, respectively [10,15]. Shi et al. proposed a noise-robust method for few-mode ring-core fibers in 2024 [20]. In the same year, Yan et al. modeled beam-center offset, objective-lens variation, and coordinate misalignment as translation, scaling, and rotation, and reduced the simulation-to-experiment gap through data augmentation [21]. Further studies in 2025 considered experimental mismatch, noise augmentation, and adversarial perturbations [22,23]. Nevertheless, two limitations remain. First, most existing approaches address one or several predefined perturbations mainly through data augmentation, restoration, or perturbation-specific training, whereas structural robustness to perturbations unseen during training has received less attention. Second, perturbation types, severity ranges, and evaluation criteria vary substantially across studies, making it difficult to systematically characterize performance degradation from Clean to progressively degraded conditions or to distinguish baseline predictive capability from performance retention. Graded corruption benchmarks and anti-aliased downsampling have demonstrated the importance of evaluating robustness beyond Clean accuracy in general vision tasks [24,25], but similarly systematic studies remain limited for MD in FMFs.

To address the above issues, we propose a lightweight and robust mode purity regression method based on RepViT [26], and investigate its ability to retain performance under unseen input perturbations using only Clean data for training. The network is designed for robustness from three aspects: feature representation, input geometric stability, and internal sampling stability, aiming to reduce its sensitivity to varying perturbation conditions while maintaining low computational cost. In addition, we construct a graded composite perturbation protocol consisting of the Clean condition and five perturbation levels (A–E). Combined with controlled single-perturbation experiments and comprehensive performance-retention metrics, this protocol enables a systematic evaluation of model performance from ideal conditions to progressively more severe non-ideal perturbations. The main contributions of this work are summarized as follows.

(1) A lightweight feature extraction architecture is proposed for robust MD based on mode purity regression. Based on RepViT, the Bifocal Spatial Token Mixer (BSTM), Spatial-Frequency Gated Token Mixer (SFGTM), and lightweight Atrous Spatial Pyramid Pooling (ASPP) are introduced at different stages, together with Spatial-Channel Coupled ECA (SC-ECA), to enhance modal feature representation in terms of local–global spatial relationships, spectral information, multi-scale structures, and spatial-channel feature selection.

(2) A cooperative mechanism combining input geometric normalization and anti-aliased downsampling is developed. Moment-Based Geometric Normalization (MomentNorm2d) is used to reduce the effects of spot displacement and limited scale variations, while BlurPool suppresses aliasing during internal downsampling, thereby improving prediction stability at both the input-geometry and feature-sampling levels.

(3) A systematic robustness evaluation framework is established for mode purity prediction in FMFs. Clean and A–E graded composite perturbation test sets are constructed, together with response analyses under five individual perturbations, the mean composite–perturbation correlation coefficient, and the relative correlation degradation rate, to evaluate model robustness in terms of both absolute predictive performance and performance retention relative to the Clean baseline.

2. Method

2.1. Fiber Mode Superposition and Mode Purity Formulation

For a given few-mode fiber, the output complex optical field can be expressed as a coherent superposition of its eigenmodes [6]:

$$E(x,y) = \sum_{n=1}^{N} a_n \psi_n(x,y) e^{i\phi_n} \tag{1}$$

where $\psi_n(x,y)$ denotes the n-th energy-normalized eigenmode field, and a_n and ϕ_n represent the corresponding amplitude coefficient and relative phase, respectively. In practice, CCD or CMOS detectors generally record only the near-field intensity, and the observed image can therefore be expressed as $I(x,y) = |E(x,y)|^2$. Since a single intensity image does not contain complete phase information, simultaneously recovering the amplitudes and phases of all modes from $I(x,y)$ is generally non-unique. Rather than reconstructing the complete complex optical field, this work focuses on the normalized energy fraction of each mode in the total optical power, referred to as mode purity. Since all eigenmodes are normalized to unit energy, the mode purity of the n-th mode is defined as $p_n = \frac{|a_n|^2}{\sum_{m=1}^{N}|a_m|^2}$.

In this work, the normalized energies of the four mode groups, LP_{01}, LP_{11}, LP_{21}, and LP_{02}, are used as the supervised targets. The corresponding mode purity vector is defined as $\boldsymbol{p} = \left[p_{LP_{01}}, p_{LP_{11}}, p_{LP_{21}}, p_{LP_{02}}\right]^T$. Accordingly, mode purity prediction can be formulated as a supervised regression problem that maps a near-field intensity image to a four-dimensional mode purity vector: $\hat{\boldsymbol{p}} = f_\theta(I)$, where f_θ denotes the regression network parameterized by θ, and $\hat{\boldsymbol{p}}$ is the predicted mode purity vector. The focus of this work is not only to minimize the error between $\hat{\boldsymbol{p}}$ and $\boldsymbol{p}$ under ideal input conditions, but also to maintain stable predictions when the modal energy composition remains unchanged while the input is affected by non-ideal perturbations.

2.2. Robust Geometric Normalization and Anti-Aliased Sampling

To address these imaging-induced geometric variations, prediction stability is improved at two levels: input geometry and internal network sampling. MomentNorm2d is used to correct beam position and limited-scale variations, while BlurPool reduces the sampling sensitivity caused by residual geometric shifts during network downsampling.

2.2.1. Moment-Based Geometric Normalization

Given an input near-field intensity image $I(x, y)$, the total intensity is first calculated as $S = \sum_{x,y} I(x, y)$. The centroid of the beam pattern is then determined from the first-order moments of the intensity distribution:

$$c_x = \frac{\sum_{x,y} x\, I(x, y)}{S}, \quad c_y = \frac{\sum_{x,y} y\, I(x, y)}{S}. \tag{2}$$

Since the centroid is obtained by weighting the continuous intensity distribution, it provides a stable reference for correcting beam translation. The effective spatial scale of the beam pattern is further estimated using the second central moment:

$$r = \sqrt{\frac{\sum_{x,y} [(x - c_x)^2 + (y - c_y)^2] I(x, y)}{S}}. \tag{3}$$

Compared with threshold-based bounding regions, the second-moment radius does not rely on a manually specified segmentation threshold. Instead, it estimates the overall spatial scale directly from the continuous intensity distribution, providing better stability against weak background variations and continuous grayscale changes. Given a target radius r_t, the scale factor is calculated as $s = \mathrm{clip}\left(\frac{r_t}{r}, S_{\min}, S_{\max}\right)$. Here, $\mathrm{clip}(\cdot)$ constrains the scale factor to a predefined range, which is important for mode purity regression. In all experiments, r_t was set to 0.32, with $S_{\min} = 0.70$ and $S_{\max} = 2.20$. The second central moment is affected not only by external imaging magnification, but also by the intrinsic radial distribution, lobe structure, and combination of different modes. Therefore, the second-moment radius is not regarded as an imaging-scale parameter completely independent of the modal state. Nor are all inputs forced to be normalized to an identical spatial distribution. Instead, only a limited-range scale correction is applied to reduce the distribution shift caused by variations in imaging magnification. This improves scale stability while avoiding excessive normalization of spatial structures that are discriminative for different modes. In other words, the purpose of MomentNorm2d is not to eliminate discriminative spatial-scale differences among modes, but only to correct limited scale shifts introduced by external imaging conditions. An affine sampling transformation is then constructed according to the centroid (c_x, c_y) and the scale factor s. The normalized image is obtained by bilinear interpolation as $I_{\mathrm{norm}} = T(I, c_x, c_y, s)$, where $T(\cdot)$ denotes a differentiable sampling operation consisting of translation and limited-scale transformation.

MomentNorm2d applies the same deterministic processing during both training and inference. By reducing the distribution shift caused by variations in the absolute position of the near-field beam pattern and imaging magnification, the subsequent network can rely less on absolute beam location and overall scale, and instead focus more on morphology-related features associated with modal composition, such as lobe distributions, ring structures, radial nodes, and the locations of dark regions.

2.2.2. Anti-Aliased Downsampling with BlurPool

MomentNorm2d mainly corrects global geometric variations at the input level. However, even after approximate input alignment, stride-2 convolutions and downsampling operations within the convolutional network may still be sensitive to small positional changes. When the input features contain high-frequency components beyond the Nyquist frequency after downsampling, direct strided sampling may fold these components into lower-frequency regions, resulting in aliasing. For near-field images of FMFs, inter-lobe dark fringes, ring edges, and local interference structures all contain pronounced high-frequency responses. Small positional or interpolation variations may be progressively amplified through multiple downsampling stages, leading to noticeable changes in deep features.

To reduce this effect, BlurPool [25] is introduced at the network stem and all stride-2 downsampling locations. The basic operation is expressed as $Y = \text{Subsample}(H * X)$, where X denotes the input feature, H is a fixed low-pass filter kernel, $*$ represents the convolution operation, and $\text{Subsample}(\cdot)$ denotes downsampling with a stride of 2. In this work, a length-3 binomial low-pass kernel is adopted, and the two-dimensional low-pass filtering is implemented using separable filtering along the horizontal and vertical directions. The 3-tap kernel provides moderate anti-aliasing while limiting excessive smoothing of discriminative local modal structures and maintaining low computational overhead. This design introduces almost no additional learnable parameters.

The purpose of BlurPool is not to remove all high-frequency modal information, but to appropriately suppress high-frequency components that may cause aliasing before reducing the spatial resolution, thereby making the network features vary more smoothly with changes in input position. In this way, MomentNorm2d and BlurPool operate at two different levels: MomentNorm2d reduces the distribution shift caused by global positional and limited magnification variations at the input level, whereas BlurPool mitigates the amplification of residual geometric variations during multi-stage downsampling within the network. Together, they form the input geometric stabilization and internal sampling stabilization mechanism used in this work.

3. Network Architecture

3.1. Overall Architecture and Optimization Objective

3.1.1. Overall Architecture

RepViT [26] is adopted as the lightweight backbone, and the overall architecture is shown in Figure 1. The network is designed to improve robustness from three complementary aspects: input geometry, internal sampling, and feature representation. The input near-field intensity image is first processed by MomentNorm2d to reduce beam position and limited-scale variations. The normalized image is then fed into the modified RepViT backbone, where BlurPool is introduced at the stem and all stride-2 downsampling locations to reduce aliasing and sampling sensitivity. For feature extraction, different Token Mixers are assigned to different stages according to feature resolution and semantic level. BSTM is used in Stages 1 and 4, SFGTM in Stage 2, and lightweight ASPP in Stage 3. In addition, the original SE modules are replaced with SC-ECA for spatial-channel feature recalibration. Finally, the Stage 4 output is mapped to a four-dimensional mode purity vector through global average pooling and a mode-purity-constrained regression head, where Softmax enforces non-negativity and unit-sum constraints. The output size, number of channels, and main module of each stage are summarized in Table 1. Detailed designs of the Token Mixers, SC-ECA, and regression head are presented in Sections 3.2 and 3.3.

Table 1. Output sizes, channel numbers, and main modules of the proposed network.

Stage	Output Size	Channels	Main Module
Input	224 × 224	1	resize and normalization
Geometry norm	224 × 224	1	MomentNorm2d
Stem	56 × 56	40	anti-aliased stem
Stage 1	56 × 56	40	BSTM
Stage 2	28 × 28	80	SFGTM
Stage 3	14 × 14	160	lightweight ASPP
Stage 4	7 × 7	320	BSTM
Head	4	–	Mode-Purity-Constrained Regression Head

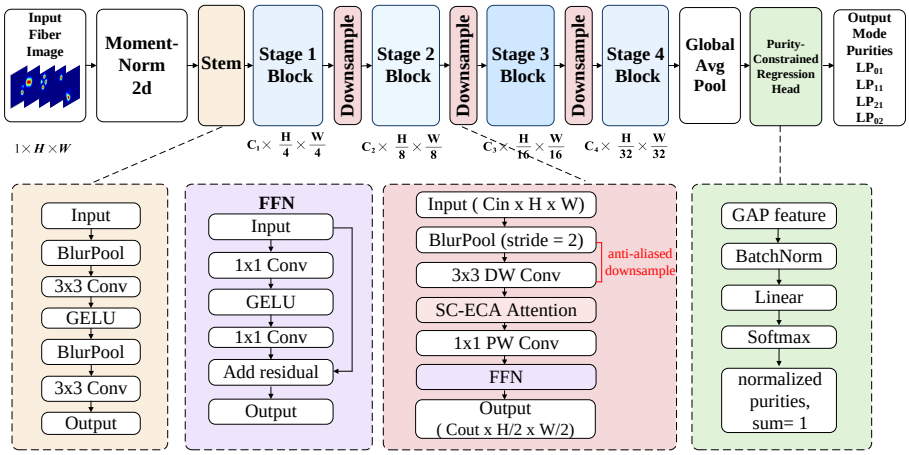


Figure 1. Overall architecture of the proposed model.

3.1.2. Optimization Objective

In addition to the architectural design, the model is trained end-to-end using the Smooth L1 loss, with $\beta = 0.01$. The prediction error is defined as $\boldsymbol{e} = \hat{\boldsymbol{p}} - \boldsymbol{p}$. The corresponding element-wise loss for each scalar error component is

$$\ell(e) = \begin{cases} \frac{1}{2}\frac{e^2}{\beta}, & |e| < \beta, \\ |e| - \frac{1}{2}\beta, & |e| \geq \beta. \end{cases} \tag{4}$$

Averaging over all samples and mode components in a batch gives the overall training loss:

$$L = \frac{1}{BK}\sum_{i=1}^{B}\sum_{k=1}^{K}\ell\left(\hat{p}_{i,k} - p_{i,k}\right) \tag{5}$$

where B denotes the batch size and K is the number of mode components. For small residuals, Smooth L1 applies a quadratic penalty to preserve fine-grained regression accuracy. For larger residuals, the gradient becomes constant in magnitude, reducing the excessive influence of a small number of large errors on parameter updates. In this work, Smooth L1 is mainly used to stabilize the optimization of mode purity regression, while robustness to imaging perturbations is primarily achieved through the geometric normalization, anti-aliased sampling, and robust feature representation structures described above.

3.2. Robust Multi-Scale Spatial-Frequency Token Mixing

After input geometric correction and anti-aliasing processing, the backbone network still needs to extract discriminative features from the near-field intensity images that can stably represent the modal energy composition. Near-field images of FMFs contain local lobe boundaries, long-range inter-lobe relationships, radial and ring-shaped structures, as well as global spectral information. A single receptive field or a single spatial-domain Token Mixer is insufficient to capture these complementary features simultaneously. Accordingly, instead of repeatedly using the same Token Mixer across all stages, BSTM, SFGTM, and lightweight ASPP are configured according to the feature resolution and semantic level of each stage, as shown in Figure 2.

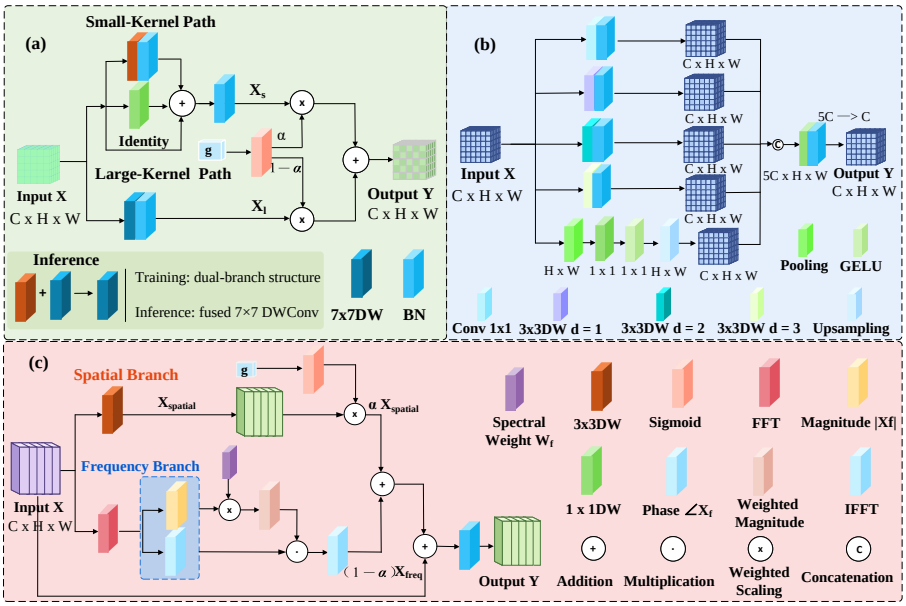


Figure 2. Structures of the three Token Mixers. (**a**) BSTM; (**b**) lightweight ASPP; (**c**) SFGTM.

3.2.1. Bifocal Spatial Token Mixer

The discriminative information in near-field mode images exhibits clear differences in spatial scale. Lobe boundaries, local intensity gradients, and dark fringes correspond to relatively small-scale structures, whereas the overall beam symmetry, spatial relationships among different bright lobes, and ring-shaped contours require a larger effective receptive field. Using only small convolution kernels may limit long-range spatial modeling, while excessive reliance on large receptive fields may weaken fine-grained local information. To address this issue, BSTM is constructed with two parallel spatial branches to extract complementary features at local and large spatial ranges. Given an input feature X, the outputs of the two branches are expressed as $X_s = T_s(X), \quad X_l = T_l(X)$, where T_s denotes the RepViT-style depthwise Token Mixer for preserving local edges and fine-grained structures, while Tl employs depthwise convolution with a larger effective receptive field to capture long-range spatial dependencies. The two types of features are adaptively fused through a learnable channel-wise gating mechanism:

$$Y = g \odot X_s + (1 - g) \odot X_l, \quad g = \sigma(\alpha) \tag{6}$$

where α is the trainable gating parameter, $\sigma(\cdot)$ denotes the Sigmoid function, and $\odot$ represents element-wise multiplication. BSTM is employed in both Stage 1 and Stage 4, but serves different purposes at the two stages. Stage 1 retains a relatively high spatial resolution and mainly preserves the lobe boundaries and local geometric details of the original beam pattern. In contrast, Stage 4 contains higher-level semantic representations, where BSTM focuses more on integrating overall symmetry and long-range modal structures. By performing local–global joint modeling at both early and late stages, the network reduces its dependence on features at a single spatial scale.

3.2.2. Spatial-Frequency Gated Token Mixer

Spatial convolution is effective at extracting local neighborhood features, but the differences among mode groups are not limited to local morphology. The number of lobes, ring-shaped distributions, and radial nodes also affect the overall spatial-frequency distribution. However, relying solely on local spatial convolution may be insufficient to fully characterize these global structural patterns. To address this issue, SFGTM is introduced in Stage 2 to

construct complementary feature representations in both the spatial and frequency-domains. The spatial branch employs depthwise convolution: $X_s = \mathrm{DWConv}_{3\times3}(X)$.

For the frequency-domain branch, the spectrum is first obtained using a two-dimensional real Fourier transform. Its amplitude response is then adaptively rescaled by learnable parameters while preserving the original phase information. This process can be compactly expressed as $X_f = F^{-1}\left[\left(W_f \odot |F(X)|\right)e^{j\angle F(X)}\right]$, where $F(\cdot)$ and $F^{-1}(\cdot)$ denote the two-dimensional Fourier transform and its inverse, respectively. For an input feature $X \in \mathbb{R}^{B\times C\times H\times W}$, the one-sided spectrum has a size of $B \times C \times H \times (\lfloor W/2 \rfloor + 1)$, and the learnable spectral weight is defined as $W_f \in \mathbb{R}^{1\times C\times H\times(\lfloor W/2\rfloor+1)}$. In Stage 2, where $C = 80$ and $H = W = 28$, W_f has a size of $1 \times 80 \times 28 \times 15$. The spectral magnitude is reweighted by W_f, while the original phase is retained during inverse Fourier reconstruction. Finally, the spatial and frequency branches are fused through a gating mechanism: $Y = g_f \odot X_s + (1 - g_f) \odot X_f$.

The spatial branch preserves local geometric structures, while the frequency-domain branch provides spectral statistics with global receptive characteristics. In this work, frequency-domain representations are not regarded as inherently translation- or scale-invariant; instead, their global modeling capability is used to reduce the network's dependence on a single type of local texture pattern. SFGTM is placed in Stage 2, where the feature resolution has already been substantially reduced from that of the original input, thereby controlling the computational cost of the FFT while retaining sufficient spatial structure of the modal patterns. Compared with introducing frequency-domain modeling at shallower layers, this placement reduces the influence of low-level noise on spectral responses; compared with introducing it at deeper stages, it avoids excessive loss of spectral details caused by repeated downsampling.

3.2.3. Lightweight ASPP

Although MomentNorm2d reduces global scale variations, local modal structures may still appear at different spatial extents. Therefore, multi-scale receptive fields are introduced in Stage 3 using a lightweight ASPP module.

ASPP extracts multi-scale features through parallel atrous convolution branches [27]. Inspired by this design, the proposed lightweight ASPP combines a 1×1 projection branch, multiple depthwise dilated convolution branches, and a global pooling branch. Its output is formulated as

$$Y = \mathrm{Proj}\Big(\mathrm{Concat}[B_0(X), D_{r_1}(X), \ldots, D_{r_m}(X), P(X)]\Big) \tag{7}$$

where $B_0(\cdot)$ denotes the local projection branch, $D_{r_m}(\cdot)$ the depthwise dilated convolution branch with dilation rate r_m, $P(\cdot)$ the global pooling branch, and $\mathrm{Proj}(\cdot)$ the channel projection after concatenation.

Compared with the original ASPP, depthwise dilated convolution reduces the additional parameter and computational overhead of parallel multi-scale processing. Since the Stage 3 feature map is 14×14, dilation rates $r = \{1, 2, 3\}$ are used to enlarge the receptive field while limiting excessive boundary sampling.

3.3. SC-ECA Feature Recalibration and Mode-Purity-Constrained Regression

3.3.1. Spatial-Channel Coupled ECA

After multi-scale spatial-frequency feature extraction, different spatial locations and channels contribute unequally to mode purity prediction. To emphasize informative modal responses while suppressing perturbation-sensitive features, SC-ECA is introduced to replace the SE modules in the original RepViT, as shown in Figure 3.

Given an input feature X, LayerNorm is first applied to obtain the normalized feature Z. SC-ECA then processes Z through parallel spatial and channel branches. The spatial branch

extracts local responses using depthwise convolution: $X_s = \mathrm{DWConv}_{3\times3}(Z)$. The channel branch adopts Efficient Channel Attention (ECA) [28] to generate channel weights: $w_c = \sigma\left[\mathrm{Conv1D}_k\left(\mathrm{GAP}(Z)\right)\right]$, where k denotes the kernel size of the one-dimensional convolution. Compared with the SE module used in the original RepViT, ECA avoids explicit channel reduction and restoration through fully connected layers, and instead models local cross-channel interactions using lightweight 1D convolution. Combined with depthwise convolution in the spatial branch, SC-ECA introduces spatial-channel recalibration with a more compact parameter structure than the original SE design. The spatial and channel responses are then multiplicatively coupled and combined with the input through a residual connection: $Y = X + X_s \odot (w_c \odot Z)$. This design enhances features that exhibit both strong spatial responses and high channel importance while preserving the original representation through the residual path. To control model complexity, SC-ECA is used only to replace the original SE modules rather than being additionally inserted into every block.

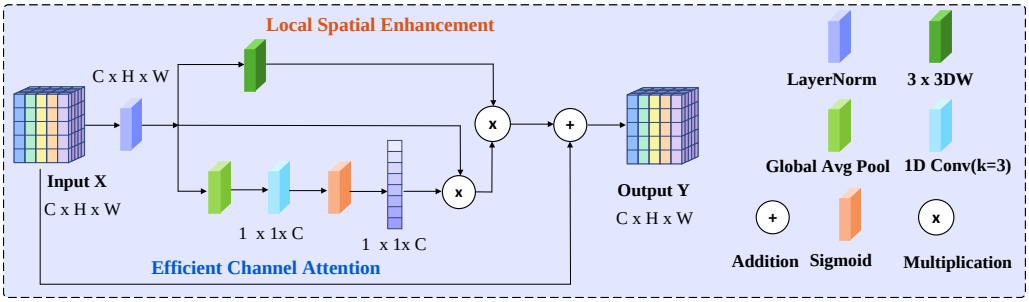


Figure 3. Structure of the SC-ECA module.

3.3.2. Mode-Purity-Constrained Regression Head

After Stage 4, the network produces a high-level feature map of size 7 × 7 with 320 channels. To map these features into the mode-purity space, global average pooling is first applied, followed by BatchNorm1d and a linear layer to generate four-dimensional logits. The predicted mode purity vector is then obtained through Softmax $\hat{p} = \mathrm{Softmax}\left(W\,\mathrm{BN}\left(\mathrm{GAP}(X_4)\right) + b\right)$. As a result, the network output naturally satisfies $\hat{p} \geq 0, \quad \sum_{k=1}^{4} \hat{p}_k = 1$.

This constraint is consistent with the non-negativity and normalization definitions of mode purity, preventing non-physical predictions such as negative mode purities or mode components whose sum deviates from unity. Consequently, no additional non-negativity or normalization penalty terms are required.

In summary, the proposed model incorporates robustness-oriented designs at multiple levels for mode purity regression under non-ideal input perturbations. MomentNorm2d reduces task-irrelevant translation and limited scale variations at the input level, while BlurPool suppresses aliasing and sampling sensitivity during multi-stage downsampling. BSTM, SFGTM, and lightweight ASPP enhance modal feature representation from three complementary perspectives: local–global spatial relationships, global spectral information, and multi-scale structures. SC-ECA further improves the selection of informative spatial–channel responses, and the mode-purity-constrained regression head ensures that the final predictions satisfy the basic physical constraints of modal energy composition. These modules operate at different levels and serve distinct functions, jointly forming a lightweight and robust mode purity regression framework for non-ideal input conditions.

4. Experimental Results

All experiments were conducted on a workstation equipped with an Intel Core i9-14900KF CPU, 64 GB RAM, and an NVIDIA GeForce RTX 5080 GPU, using Python 3.11 and

PyTorch 2.11. Except for differences in network architecture, all comparison models used the same data split, input resolution, and training protocol to ensure a fair comparison.

4.1. Experimental Setup and Robustness Evaluation Protocol

4.1.1. Dataset and Implementation Details

Simulated data were generated based on step-index fiber theory. The fiber core radius was set to 10 μm, with a core refractive index of 1.45, a numerical aperture of 0.132, and an operating wavelength of 1550 nm. Based on these parameters, the eigenmode fields of LP_{01}, LP_{11}, LP_{21}, and LP_{02} were solved and normalized. Near-field intensity images were then generated from random mode combinations and resized to 224 × 224 single-channel inputs.

A total of 40,000 samples were generated and divided into training, validation, and test sets at a ratio of 8:1:1, containing 32,000, 4000, and 4000 samples, respectively. A fixed random seed of 42 was used for all stochastic processes. The model was trained for 150 epochs with a batch size of 128 using AdamW, with an initial learning rate of 3×10^{-4}, weight decay of 0.05, and a 20-epoch warm-up. The model converged after approximately 100 epochs. Smooth L1 loss with $\beta = 0.01$ was used as the training objective.

All models were trained and validated exclusively on Clean data. The A–E composite perturbation sets and individual-perturbation sets were used only for testing, without retraining, fine-tuning, or parameter selection. All comparison models followed the same data split, input resolution, and training protocol to ensure a fair robustness comparison.

4.1.2. Construction of the Graded Composite Perturbation Benchmark

Performance degradation in practical fiber imaging systems is usually caused by multiple factors rather than a single source. Relative displacement between the fiber facet and camera, variations in optical magnification, changes in coordinate orientation, defocus, and sensor noise may occur simultaneously, jointly affecting the position, scale, edge sharpness, and local intensity distribution of the near-field beam pattern. Evaluating a model under only a single perturbation is therefore insufficient to fully reflect the composite degradation encountered in practical acquisition environments. To address this issue, a composite perturbation protocol with five progressively increasing severity levels, denoted A, B, C, D, and E, is constructed, where A represents mild perturbation and E represents the strongest perturbation. Let the five composite perturbation levels form the set $\mathcal{L} = \{A, B, C, D, E\}$. For any perturbation level $\ell \in \mathcal{L}$, the corresponding parameter combination is defined as $\theta_\ell = (\lambda_\ell, \alpha_\ell, \tau_\ell, \sigma_\ell, \gamma_\ell)$, $\ell \in \mathcal{L}$, where λ_ℓ denotes the scaling factor, α_ℓ the rotation angle, τ_ℓ the maximum translation amplitude, σ_ℓ the standard deviation of the Gaussian blur kernel, and γ_ℓ the Gaussian noise level characterized by the signal-to-noise ratio (SNR). For each perturbation level, the scaling factor is uniformly sampled within the interval listed in Table 2. The rotation angle is uniformly sampled from the corresponding symmetric range, while the horizontal and vertical translation offsets are sampled independently within the allowable range. Gaussian blur and Gaussian noise are then applied using the level-specific blur standard deviation and SNR. Given an original near-field intensity image I, the five perturbations are applied sequentially in a fixed order:

$$I_\ell = N_{\gamma_\ell} \circ B_{\sigma_\ell} \circ T_{\tau_\ell} \circ R_{\alpha_\ell} \circ S_{\lambda_\ell}(I), \quad \ell \in \mathcal{L} \tag{8}$$

where S_{λ_ℓ}, R_{α_ℓ}, T_{τ_ℓ}, B_{σ_ℓ}, and N_{γ_ℓ} denote scaling, rotation, safe translation, Gaussian blur, and Gaussian noise operations, respectively. The perturbations are applied in the order of scaling, rotation, and translation so that the feasible translation range can be determined according to the effective beam boundary after the geometric transformations. Blur and noise are then applied to simulate the simultaneous presence of geometric deviations and imaging-quality degradation. Safe translation does not perform unconstrained displacement

within a preset range; instead, the actual displacement is limited according to the effective region of the transformed beam pattern, ensuring that the main beam remains within the imaging field of view after translation. This avoids unrealistic distortions caused by extensive cropping of the beam pattern and ensures that the increasing difficulty across perturbation levels mainly results from changes in imaging conditions rather than complete loss of valid modal information.

Levels A–E correspond to progressively increasing composite perturbations from mild to severe, with the scaling, rotation, translation, blur, and noise parameters strengthened simultaneously across levels. The parameter ranges were selected to represent typical variations that may occur in practical fiber imaging while providing a controlled and graded robustness benchmark. The specific parameter ranges for the five composite perturbation levels are listed in Table 2, while the original test set without any perturbation is denoted as Clean. As the perturbation level increases, the beam pattern exhibits more pronounced scale reduction, positional shift, orientation change, edge smoothing, and noise contamination, while the main modal structure remains preserved. This design avoids artificially constructing extreme test samples through severe cropping or complete destruction of valid information, so that A–E primarily evaluate model robustness under progressively degraded imaging conditions in which useful modal information is still retained.

Thus, A–E are not five independent sets of random augmentations, but instead form a graded robustness test sequence with monotonically increasing perturbation severity. By tracking model performance from Clean to E, the response to mild perturbations, the limiting performance under severe perturbations, and the overall degradation trajectory can all be examined. For each perturbation level, all compared models are evaluated using exactly the same pre-generated test samples, thereby avoiding interference from differences among randomly generated perturbation instances in the robustness comparison.

Table 2. Parameter settings for the five graded composite perturbation levels.

Test Set	Scaling Factor	Rotation Angle (°)	Maximum Translation (%)	Blur Standard Deviation (Pixel)	SNR (dB)
Clean	1.00	0	0	0	–
A	0.90–0.96	±3°	2–4%	0.3	45
B	0.82–0.90	±5°	5–8%	0.5	40
C	0.74–0.82	±10°	9–12%	0.8	35
D	0.64–0.74	±20°	13–17%	1.0	30
E	0.55–0.64	±30°	18–22%	1.3	25

In addition to the graded composite perturbations, the intensity of each individual perturbation, including noise, translation, scaling, rotation, and blur, is varied separately while all other conditions are kept unchanged. These experiments are used to characterize the model's performance response to increasing severity of each individual imaging perturbation and to further analyze its robustness under specific perturbation conditions. The results are presented in Section 4.3.

4.1.3. Basic Regression Metrics

Mean absolute error (MAE) and the Pearson correlation coefficient are used to evaluate mode purity regression performance from two complementary perspectives: absolute prediction error and consistency of variation. Suppose the test set contains N samples and

$K = 4$ mode groups. For the i-th sample, the ground-truth and predicted purities of the k-th mode group are denoted by p_{ik} and $\hat{p}_{ik}$, respectively. The mean absolute error is defined as:

$$\mathrm{MAE} = \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} |p_{ik} - \hat{p}_{ik}| \tag{9}$$

MAE directly measures the average deviation between the predicted and ground-truth mode purities, with equal weighting assigned to all mode components. A smaller MAE indicates a lower absolute regression error. For the k-th mode group, the Pearson correlation coefficient is defined as:

$$R_k = \frac{\sum_{i=1}^{N} (p_{ik} - \bar{p}_k)(\hat{p}_{ik} - \bar{\hat{p}}_k)}{\sqrt{\sum_{i=1}^{N} (p_{ik} - \bar{p}_k)^2} \sqrt{\sum_{i=1}^{N} (\hat{p}_{ik} - \bar{\hat{p}}_k)^2}} \tag{10}$$

where $\bar{p}_k$ and $\bar{\hat{p}}_k$ denote the sample means of the ground-truth and predicted purities of the k-th mode group, respectively. A value of R closer to 1 indicates higher linear consistency between the predicted and ground-truth mode purities across samples. Compared with MAE, R responds differently to overall and scale deviations: MAE measures the numerical deviation of the predictions from the ground truth, whereas R focuses more on whether the model preserves the relative variation of mode purity across different samples. Thus, both metrics are reported to avoid drawing conclusions from a single error- or correlation-based measure.

4.1.4. Composite Perturbation Robustness Metrics

The performance at individual points under the Clean and A–E conditions reflects the model's predictive capability at specific perturbation levels, but it is difficult to summarize the overall stability across the entire perturbation range using a single value. Accordingly, in addition to the basic regression metrics, this work introduces the mean composite–perturbation correlation coefficient $\bar{R}_{\mathrm{cp}}$ and the relative correlation degradation rate $\mathrm{RDR}_{\mathrm{cp}}$, which measure the overall predictive capability under composite perturbations and the degree of performance retention relative to the Clean baseline, respectively.

Let the five composite perturbation levels form the set $\mathcal{L} = \{A, B, C, D, E\}$. If the Pearson correlation coefficient of the model at perturbation level ℓ is denoted by R_ℓ, the mean composite–perturbation correlation coefficient is defined as:

$$\bar{R}_{\mathrm{cp}} = \frac{1}{5} \sum_{\ell \in \mathcal{L}} R_\ell = \frac{R_{\mathrm{A}} + R_{\mathrm{B}} + R_{\mathrm{C}} + R_{\mathrm{D}} + R_{\mathrm{E}}}{5} \tag{11}$$

Equal weight is assigned to the A–E perturbation levels, so that $\bar{R}_{\mathrm{cp}}$ reflects the average prediction consistency over the entire perturbation range. This metric is averaged only over the five perturbed test sets, excluding the Clean result, to avoid artificially inflating the robustness evaluation with near-saturated performance under unperturbed conditions. A larger $\bar{R}_{\mathrm{cp}}$ indicates better overall predictive performance under composite perturbations.

Different models may exhibit different baseline performance under Clean conditions. Therefore, comparing only $\bar{R}_{\mathrm{cp}}$ cannot fully distinguish between "higher baseline accuracy" and "stronger performance retention". To address this issue, the relative correlation degradation rate is further defined as:

$$\mathrm{RDR}_{\mathrm{cp}} = \frac{R_{\mathrm{Clean}} - \bar{R}_{\mathrm{cp}}}{R_{\mathrm{Clean}}} \times 100\% \tag{12}$$

where R_{Clean} denotes the correlation coefficient of the model on the original test set. This metric normalizes the average correlation degradation caused by composite perturbations

relative to the model's own Clean performance, thereby measuring the relative performance loss when the input distribution shifts from the unperturbed condition to composite perturbation conditions. A smaller RDR_{cp} indicates stronger performance retention across different imaging conditions.

In summary, the Clean condition is used to characterize baseline regression performance, while A–E describe the performance degradation trajectory as composite perturbations increase. The mean composite–perturbation correlation coefficient $\bar{R}_{cp}$ and the relative correlation degradation rate RDR_{cp} are used to quantify the average predictive capability over the full perturbation range and the relative performance retention with respect to the Clean baseline, respectively, thereby forming a multi-level robustness evaluation framework.

4.2. Comparison with Existing Models

To evaluate the performance of the proposed method under composite imaging degradation, DenseNet, EfficientNet, MobileNetV3_Small, ResNet18, ResNet50, Swin Transformer (Swin), Vision Transformer (ViT), and VGG16 are selected as representative comparison networks. All models use the same data split, training strategy, and perturbation testing protocol. Table 3 reports the Pearson correlation coefficient R, the mean composite–perturbation correlation coefficient $\bar{R}_{cp}$, and the relative correlation degradation rate RDR_{cp} under the Clean and A–E conditions. Figure 4 further shows the variations in R and MAE.

Table 3. Comparison of different models under graded composite perturbations.

Model	Clean	A	B	C	D	E	$\bar{R}_{cp}$ ↑	RDR_{cp} ↓
DenseNet	0.9999	0.8213	0.6235	0.4551	0.2843	0.1266	0.4622	53.78%
EfficientNet	0.9999	0.8267	0.5377	0.2581	0.0448	0.0308	0.3396	66.03%
MobileNetV3_Small	0.9999	0.8851	0.6747	0.4389	0.2048	0.0417	0.4490	55.09%
ResNet18	0.9999	0.7947	0.4540	0.2370	0.0774	0.0706	0.3267	67.32%
ResNet50	0.9999	0.8071	0.5992	0.4038	0.2085	0.2104	0.4458	55.42%
Swin Transformer	0.9999	0.9002	0.6831	0.3951	0.1966	0.1367	0.4623	53.76%
ViT	0.9999	0.8842	0.5592	0.1884	0.1135	0.1914	0.3873	61.26%
VGG16	0.9996	0.9761	0.9128	0.8153	0.7252	0.6226	0.8104	18.93%
Ours	0.9993	**0.9963**	**0.9907**	**0.9806**	**0.9645**	**0.9317**	**0.9728**	**2.66%**

Note: ↑ and ↓ indicate that higher and lower values are better, respectively.

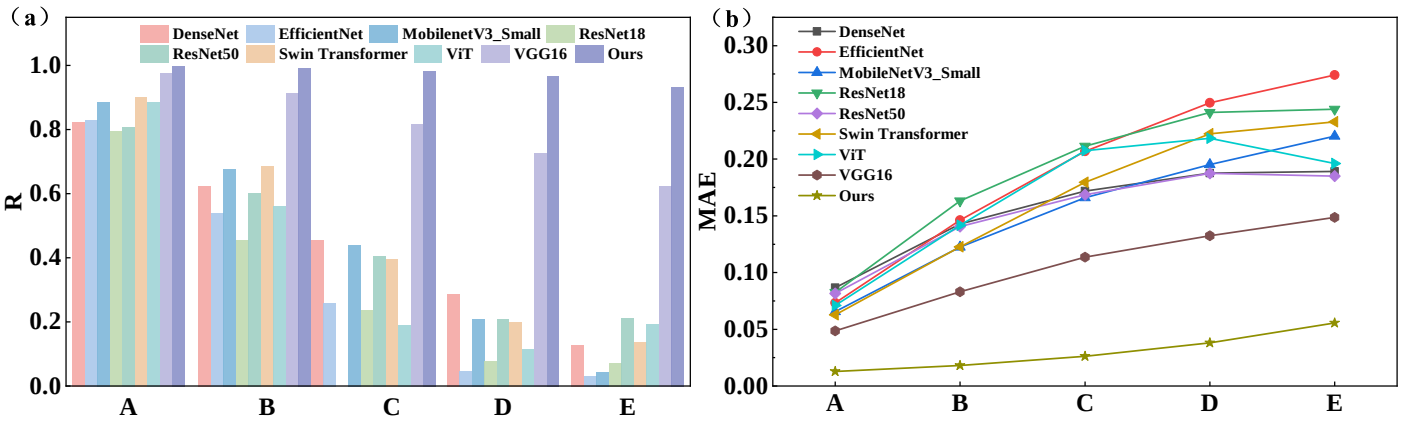


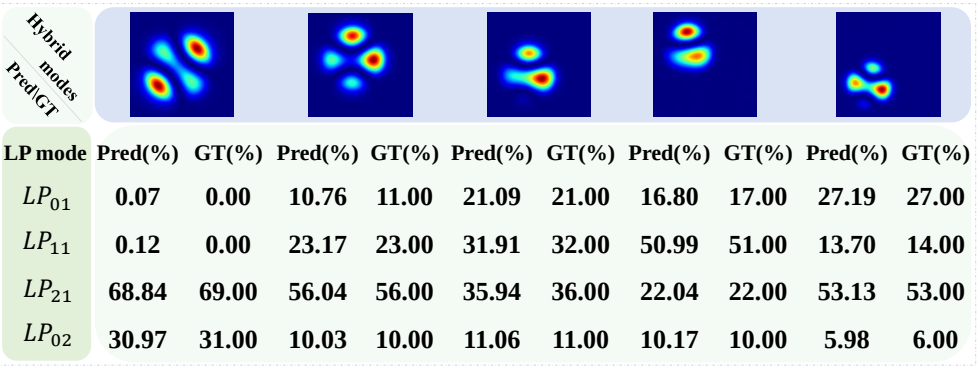
Figure 4. Comparison of model performance under different perturbation levels. (**a**) Correlation coefficient R comparison. (**b**) MAE comparison.

4.2.1. Performance Under Composite Perturbations

Under Clean conditions, all models achieve near-saturated Pearson correlation coefficients, with R ranging from 0.9993 to 0.9999. The proposed model obtains $R = 0.9993$, which is comparable to the other models. As the perturbation severity increases from A to E, most comparison models exhibit substantial performance degradation. Under level E, the R values of most baselines decrease to below 0.21, while VGG16 remains the most robust comparison model with $R = 0.6226$. In contrast, the proposed model still achieves $R = 0.9317$. Across the full A–E range, the proposed model obtains $\bar{R}_{cp} = 0.9728$, exceeding the 0.8104 of VGG16, while its RDR_{cp} is only 2.66%, substantially lower than the 18.93% of VGG16.

The MAE results in Figure 4 exhibit the same trend. As the perturbation severity increases from A to E, the MAE of the proposed model increases from 0.0129 to 0.0555, whereas that of VGG16 increases from 0.0487 to 0.1486. Over the five perturbation levels, the proposed model achieves an average MAE of 0.0301, compared with 0.1053 for VGG16. Thus, the robustness advantage is consistently reflected in both correlation-based and absolute-error metrics.

To provide an intuitive illustration of the prediction performance of the proposed model, Figure 5 presents one randomly selected sample and its prediction results for each of the five perturbation levels A–E. Although the inputs are affected by different degrees of translation, scaling, rotation, blur, and noise, the predicted purities of the four mode groups, LP_{01}, LP_{11}, LP_{21}, and LP_{02}, remain highly consistent with the ground-truth values, with a maximum absolute deviation of only 0.30 percentage points. These sample-level results further support the statistical findings reported in Table 3 and Figure 4.



LP mode	Pred(%)	GT(%)	Pred(%)	GT(%)	Pred(%)	GT(%)	Pred(%)	GT(%)	Pred(%)	GT(%)
LP_{01}	0.07	0.00	10.76	11.00	21.09	21.00	16.80	17.00	27.19	27.00
LP_{11}	0.12	0.00	23.17	23.00	31.91	32.00	50.99	51.00	13.70	14.00
LP_{21}	68.84	69.00	56.04	56.00	35.94	36.00	22.04	22.00	53.13	53.00
LP_{02}	30.97	31.00	10.03	10.00	11.06	11.00	10.17	10.00	5.98	6.00

Figure 5. Randomly selected samples and corresponding prediction results under the five perturbation levels A–E.

4.2.2. Model Complexity and Efficiency Analysis

In addition to predictive performance under perturbations, model complexity is also an important consideration for the practical application of MD. Figure 6 compares the number of parameters and FLOPs of the different models. The proposed model contains 3.80 M parameters and requires 1.367 G FLOPs. Among the comparison models, VGG16 shows the strongest robustness, with 14.85 M parameters and 30.73 G FLOPs. In comparison, the proposed model reduces the parameter count and FLOPs by approximately 74.41% and 95.55%, respectively, while increasing the mean composite–perturbation correlation coefficient over A–E from 0.8104 to 0.9728. MobileNetV3_Small has the lowest parameter count and computational cost, yet its R under level E is only 0.0417. EfficientNet also requires fewer FLOPs than the proposed model, but achieves only $R = 0.0308$ under level E.

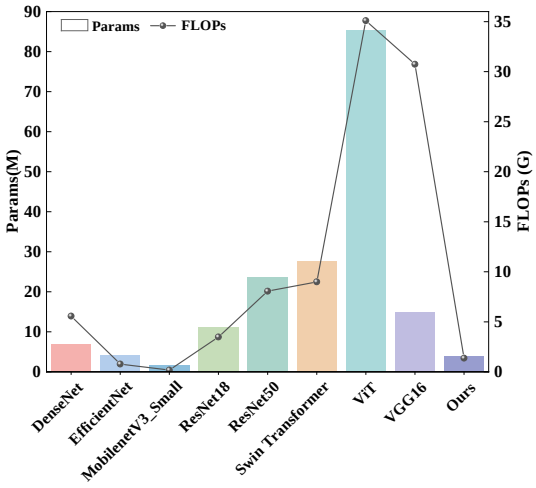


Figure 6. Comparison of model parameters and computational complexity. Bars represent the number of parameters, while the line represents FLOPs.

4.3. Robustness Under Individual Perturbations

Composite perturbation experiments can evaluate the overall robustness of a model under multi-factor degradation, but they do not clearly reveal how the model responds to increasing severity of a specific perturbation. To further analyze the robustness of the proposed model under different types of input perturbations, controlled single-perturbation experiments are conducted in addition to the composite perturbation tests. Specifically, only one perturbation type is varied at a time, while all other perturbations are disabled, and the Pearson correlation coefficient R is calculated using the same test samples. The investigated perturbations include Gaussian noise, translation, scaling, rotation, and Gaussian blur. Because these five perturbation types have different physical quantities and value ranges, their severities are independently normalized to $[0, 1]$ within their respective test ranges only for result visualization in a common coordinate system. Normalized severities of 0 and 1 correspond to the weakest and strongest tested conditions for each perturbation type, respectively, and do not imply physically equivalent degradation across different perturbations at the same normalized severity. The results for all five individual perturbations are shown together in Figure 7.

The geometric perturbation experiments show that the proposed model remains generally stable against variations in beam position, orientation, and scale, although the effects of different geometric perturbations are not identical. When the translation amplitude increases from 0% to 24%, R remains at 0.9752. When the rotation angle increases to 45°, R is 0.9810. In comparison, scaling causes a more pronounced performance decrease: when the scale factor is reduced from 1.0 to 0.5, R decreases to 0.9186. A further resolution comparison shows that increasing the input size from 224 × 224 to 448 × 448 improves the correlation coefficient at 0.5 scaling from 0.9186 to 0.9394. The rotation curve remains relatively stable overall, with $R = 0.9789$ and $R = 0.9810$ at 30° and 45°, respectively. Compared with geometric perturbations, noise and blur exhibit distinctly different performance trends. Under Gaussian blur, as the blur standard deviation σ increases from 0 to 2.0 pixels, R remains within the range of 0.9965–0.9993. In contrast, Gaussian noise reveals a more pronounced performance boundary. As the SNR decreases from 40 dB to 20 dB, R remains high, decreasing only from 0.9964 to 0.9892. However, when the SNR is further reduced to 15, 10, and 5 dB, R drops to 0.9479, 0.8244, and 0.6329, respectively.

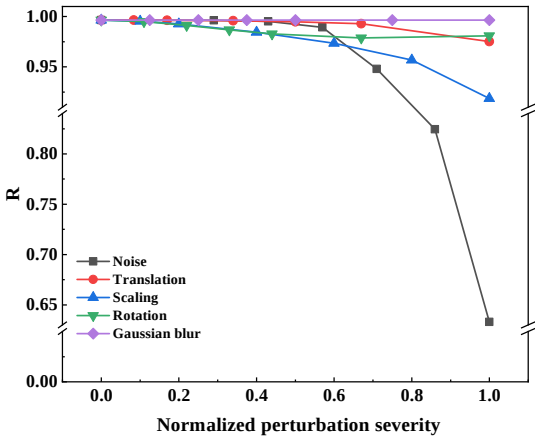


Figure 7. Robustness profiles under five individual perturbations.

Overall, the five single-perturbation experiments characterize the performance evolution of the proposed model under different imaging variations as perturbation severity increases. The single-perturbation tests are mainly used to analyze the robustness trend and possible performance limits under specific perturbations, whereas the A–E graded composite perturbation experiments evaluate the overall robustness of the model when multiple imaging degradations occur simultaneously. These two types of experiments provide complementary robustness evidence from different perspectives and jointly verify the performance retention of the proposed method under diverse input perturbations.

4.4. Ablation Study

To evaluate the contribution of each robustness-oriented component, RepViT is adopted as the baseline, and BSTM, SFGTM, lightweight ASPP, SC-ECA, BlurPool, and MomentNorm2d are progressively introduced. A cumulative ablation strategy is used, and all configurations are evaluated on the medium-severity composite perturbation set C using the Pearson correlation coefficient R, as summarized in Table 4.

Progressively modifying the feature representation modules consistently improves robustness. Replacing the original Token Mixers with BSTM increases R from 0.7079 to 0.7299, while introducing SFGTM and lightweight ASPP further raises it to 0.7445 and 0.8003, respectively. Replacing SE with SC-ECA increases R to 0.8073 while reducing the parameter count from 3.96 M to 3.80 M.

Table 4. Ablation results of different architectural components under composite perturbation C.

S1	S2	S3	S4	Attention	BlurPool	MomentNorm2d	$R\uparrow$	Params (M)↓
RepViT	RepViT	RepViT	RepViT	SE	×	×	0.7079	2.17
BSTM	BSTM	BSTM	BSTM	SE	×	×	0.7299	2.27
BSTM	SFGTM	BSTM	BSTM	SE	×	×	0.7445	2.39
BSTM	SFGTM	lightweight ASPP	BSTM	SE	×	×	0.8003	3.96
BSTM	SFGTM	lightweight ASPP	BSTM	SC-ECA	×	×	0.8073	3.80
BSTM	SFGTM	lightweight ASPP	BSTM	SC-ECA	✓	×	0.8577	3.80
BSTM	SFGTM	lightweight ASPP	BSTM	SC-ECA	✓	✓	**0.9806**	3.80

The largest gains are obtained from the sampling- and geometry-oriented modules. BlurPool increases R from 0.8073 to 0.8577, corresponding to an absolute gain of 0.0504, while introducing MomentNorm2d further raises R to 0.9806, with a gain of 0.1229. Overall,

the complete model improves R from 0.7079 to 0.9806 while increasing the parameter count only from 2.17 M to 3.80 M.

5. Discussion

5.1. Robustness Mechanism and Perturbation Response

The experimental results show that the proposed model maintains substantially better performance retention than the comparison models as the perturbation severity increases. These differences may be related to the feature formation mechanisms of different network architectures. Conventional networks are exposed only to a fixed imaging distribution during training and may therefore rely on statistical features strongly correlated with the training distribution, such as absolute position and local texture. When the input undergoes translation, scaling, or rotation, this can lead to feature mismatch between the training and test distributions. In addition, strided convolution and pooling are sensitive to sampling positions, and small shifts may be amplified through successive downsampling stages.

Although ViT and Swin Transformer provide stronger global modeling capabilities, such capability does not guarantee invariance to geometric transformations, and both models exhibit substantial degradation under the graded perturbations. VGG16, in contrast, retains relatively higher robustness under severe composite perturbations, which may be related to its hierarchical convolutional structure and progressive receptive-field expansion that help preserve part of the local modal structures under strong input degradation. Geometric data augmentation substantially improves the robustness of the tested baseline models. Nevertheless, the proposed model, trained exclusively on Clean data, still maintains markedly better performance under the A–E composite perturbations. This result further suggests that the structural design of the proposed method contributes to improved performance stability under unseen perturbations. The proposed model combines MomentNorm2d for input centroid and scale correction, BlurPool for reducing downsampling sensitivity, and BSTM, SFGTM, and lightweight ASPP for improving local–global, spectral, and multi-scale feature representation. These structural designs reduce the influence of task-irrelevant variations and contribute to the slower degradation observed under the A–E perturbations.

The individual-perturbation results further clarify the different sources of robustness degradation. Translation and rotation mainly alter the position and orientation of the modal structure in the image coordinate system without directly reducing its spatial content. Strong downscaling, however, compresses the beam pattern into fewer effective pixels and gradually removes details such as local lobe boundaries, radial nodes, and dark-region widths. The improvement from $R = 0.9186$ to $R = 0.9394$ when the input resolution is increased from 224 × 224 to 448 × 448 at 0.5 scaling supports this interpretation, showing that denser spatial sampling can partially alleviate the information loss caused by strong downscaling. Nevertheless, the performance does not fully recover to the unscaled level, indicating that strong scale reduction still imposes an information limit even after geometric normalization and multi-scale feature modeling.

The relatively stable rotation response is also related to the definition of the supervised targets. In this work, the prediction targets are defined at the mode-group level and do not depend on a specific spatial orientation basis. Rotation therefore changes the spatial orientation of the near-field pattern without changing the true purity of the corresponding mode group, reducing the need for the network to rely strongly on orientation-specific textures. The slight increase in R from 30° to 45° is regarded as a local numerical fluctuation rather than a systematic improvement.

Noise and blur exhibit different degradation characteristics. Gaussian blur primarily suppresses high-frequency local structures while leaving the overall lobe distribution, ring-

shaped contours, and radial structures largely preserved within the tested range, which is consistent with the small variation in R. In contrast, strong Gaussian noise progressively disrupts local intensity relationships, dark regions, and edge information, resulting in a clear performance boundary at low SNR. Under extremely low-SNR conditions, strong background noise may also disturb the intensity distribution used for centroid and second-moment estimation in MomentNorm2d, thereby reducing the reliability of geometric normalization.

5.2. Robustness–Complexity Trade-Off

The comparison of model complexity and robustness indicates that stronger robustness is not obtained simply by increasing network size or computational cost. VGG16 requires substantially more parameters and FLOPs than the proposed model, but its mean composite–perturbation correlation coefficient remains lower. Conversely, MobileNetV3_Small has the lowest parameter count and computational cost among the comparison models, but exhibits considerable degradation under severe perturbations. EfficientNet shows a similar pattern, requiring relatively low computational cost while providing limited robustness under Level-E perturbation.

These observations indicate that lightweight design and robustness do not exhibit a simple positive correlation. Compared with merely increasing or compressing network capacity, targeted structural designs for input geometric variations, downsampling-induced sampling sensitivity, and modal feature representation are more important for maintaining prediction stability under changing imaging conditions. The proposed model therefore achieves its robustness advantage through the complementary effects of geometric normalization, anti-aliased sampling, and multi-scale spatial-frequency feature extraction rather than through increased model capacity. The ablation results further support this interpretation. Replacing SE with SC-ECA improves the correlation coefficient while reducing the parameter count, indicating that spatial-channel recalibration can enhance feature selection without increasing model complexity. More substantial gains are obtained after introducing BlurPool and MomentNorm2d, suggesting that input geometric variation and downsampling-induced sampling sensitivity are important sources of performance degradation under composite perturbations. These results further confirm that the robustness improvement arises from multiple structural components rather than from a single module.

Overall, the proposed architecture provides a favorable balance between perturbation robustness and computational complexity. The complete model requires only 3.80 M parameters and 1.367 G FLOPs while maintaining $\bar{R}_{\mathrm{cp}} = 0.9728$ across the A–E perturbation levels, supporting its suitability for lightweight and robustness-oriented mode decomposition.

6. Conclusions

To address the performance degradation of mode purity prediction in FMFs under non-ideal input conditions, this work has proposed a lightweight and robust mode purity regression method based on RepViT, with targeted designs at three levels: input geometry, internal sampling, and multi-scale spatial-frequency feature representation. Meanwhile, a robustness evaluation protocol has been established, consisting of the Clean condition, A–E graded composite perturbations, and five individual perturbations, together with comprehensive performance metrics to evaluate performance retention under perturbed conditions. Experimental results have shown that, although all models achieve near-saturated performance under Clean conditions, the proposed method exhibits a slower performance degradation trend as perturbation severity increases. It still achieves $R = 0.9317$ under the strongest level-E perturbation, with a mean composite–perturbation correlation coefficient of 0.9728 over the A–E range and a relative correlation degradation rate of only 2.66%.

The model contains only 3.80 M parameters and requires 1.367 G FLOPs, indicating that it achieves strong perturbation stability while maintaining a lightweight architecture.

This work extends the research focus from prediction accuracy under ideal conditions to performance retention under varying imaging conditions, highlighting the importance of robustness in evaluating the reliability of MD models. The proposed lightweight robust prediction framework and graded perturbation evaluation method provide a practical approach for extending few-mode fiber mode analysis from ideal simulation conditions to dynamic and non-ideal imaging environments, and offer a reference for stable mode purity estimation in subsequent mode-state monitoring, mode-coupling diagnosis, and adaptive mode control. The present method focuses on mode-group purity estimation from a single near-field intensity image and does not address intermodal phase recovery or intra-group degenerate-state discrimination. Extension to these tasks would require additional phase-sensitive information or supervision, together with corresponding modifications to the geometric normalization and anti-aliasing operations.

Author Contributions: Conceptualization, T.W.; methodology, Q.X., H.W. and Z.L.; software, H.W.; validation, Q.X., H.W., Z.L. and B.Y.; formal analysis, Q.X. and L.J.; investigation, Q.X., Z.L., L.J., B.Y. and Y.W.; data curation, Q.X., H.W. and Y.W.; visualization, Q.X.; writing—original draft preparation, Q.X.; writing—review and editing, T.W.; supervision, T.W. and J.Z.; project administration, T.W. and J.Z.; funding acquisition, T.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (62305203), "Chenguang Program" supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission (22CGA79), and Shanghai Sailing Program (22YF1414200).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

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

References

1. Richardson, D.J.; Fini, J.M.; Nelson, L.E. Space-division multiplexing in optical fibres. *Nat. Photonics* **2013**, *7*, 354–362. [CrossRef]
2. Lin, T.; Liu, A.; Zhang, X.; Li, H.; Wang, L.; Han, H.; Chen, Z.; Liu, X.; Lü, H. Analyzing OAM mode purity in optical fibers with CNN-based deep learning. *Chin. Opt. Lett.* **2019**, *17*, 100603. [CrossRef]
3. Cao, H.; Čižmár, T.; Turtaev, S.; Tyc, T.; Rotter, S. Controlling light propagation in multimode fibers for imaging, spectroscopy, and beyond. *Adv. Opt. Photonics* **2023**, *15*, 524–612. [CrossRef]
4. Xiong, W.; Ambichl, P.; Bromberg, Y.; Redding, B.; Rotter, S.; Cao, H. Spatiotemporal control of light transmission through a multimode fiber with strong mode coupling. *Phys. Rev. Lett.* **2016**, *117*, 053901. [CrossRef] [PubMed]
5. Savović, S.; Djordjevich, A. Solution of mode coupling in step-index optical fibers by the Fokker–Planck equation and the Langevin equation. *Appl. Opt.* **2002**, *41*, 2826–2830. [CrossRef] [PubMed]
6. An, Y.; Li, J.; Huang, L.; Li, L.; Leng, J.; Yang, L.; Zhou, P. Numerical mode decomposition for multimode fiber: From multi-variable optimization to deep learning. *Opt. Fiber Technol.* **2019**, *52*, 101960. [CrossRef]
7. Manuylovich, E.S.; Dvoyrin, V.V.; Turitsyn, S.K. Fast mode decomposition in few-mode fibers. *Nat. Commun.* **2020**, *11*, 5507. [CrossRef] [PubMed]
8. Wang, C.; Zhang, J.; Yan, B.; Mi, S.; Fan, G.; Wang, M.; Zhang, P. Block-based mode decomposition in few-mode fibers. *Photonics* **2025**, *12*, 66. [CrossRef]
9. An, Y.; Huang, L.; Li, J.; Leng, J.; Yang, L.; Zhou, P. Learning to decompose the modes in few-mode fibers with deep convolutional neural network. *Opt. Express* **2019**, *27*, 10127–10137. [CrossRef] [PubMed]

10. Fan, X.; Ren, F.; Xie, Y.; Zhang, Y.; Niu, J.; Zhang, J.; Wang, J. Mitigating ambiguity by deep-learning-based modal decomposition method. *Opt. Commun.* **2020**, *471*, 125845. [CrossRef]
11. Rothe, S.; Zhang, Q.; Koukourakis, N.; Czarske, J. Intensity-only mode decomposition on multimode fibers using a densely connected convolutional network. *J. Light. Technol.* **2021**, *39*, 1672–1679. [CrossRef]
12. Tian, Z.; Pei, L.; Wang, J.; Zheng, J.; Hu, K.; Chang, Y.; Li, Z.; Li, J.; Ning, T. High-precision mode decomposition for few-mode fibers based on multi-task deep learning. *J. Light. Technol.* **2022**, *40*, 7711–7717. [CrossRef]
13. Yan, B.; Zhang, J.; Wang, M.; Jiang, Y.; Mi, S. Degenerated mode decomposition with convolutional neural network for few-mode fibers. *Opt. Laser Technol.* **2022**, *154*, 108287. [CrossRef]
14. Wang, T.; Wang, H.; Liu, Z.; Xiong, Q.; Wen, Y.; Zuo, J. Lightweight fiber mode decomposition enabled by traceable knowledge distillation. *Chin. Opt. Lett.* **2026**, *24*, 061104. [CrossRef]
15. Gao, H.; Chen, Z.; Zhang, Y.-X.; Zhang, W.-G.; Hu, H.-F.; Yan, T.-Y. Rapid mode decomposition of few-mode fiber by artificial neural network. *J. Light. Technol.* **2021**, *39*, 6294–6300. [CrossRef]
16. Xu, M.; Hou, M.; Luo, X.; Xu, J.; Chen, W.; An, Y.; Zeng, X.; Li, J.; Huang, L. Multi-order hybrid vector mode decomposition in few-mode fibers with DL-based SPGD algorithm. *Opt. Laser Technol.* **2023**, *167*, 109795. [CrossRef]
17. Li, J.-J.; Zhang, R.; Wen, F.; Yang, F.; Wu, B.-J.; Qiu, K. Optical vector-eigenmode decomposition for few-mode fibers through deep learning networks. *Opt. Commun.* **2024**, *564*, 130602. [CrossRef]
18. Sun, Q.; Pei, L.; Wang, J.; Xu, L.; Gao, H.; Hu, K.; Zheng, J.; Li, J.; Ning, T. High-precision mode decomposition with low data requirements using deep learning. *IEEE Photonics Technol. Lett.* **2025**, *37*, 793–796. [CrossRef]
19. Zhao, J.; Chen, G.; Bi, X.; Cai, W.; Yue, L.; Tang, M. Fast mode decomposition for few-mode fiber based on lightweight neural network. *Chin. Opt. Lett.* **2024**, *22*, 020604. [CrossRef]
20. Shi, Z.; Mao, B.; Wang, Z.; Liu, Y.-G. Accurate and noise-robust mode decomposition by analyzing the intensity distribution of a few-mode ring core fiber. *J. Light. Technol.* **2024**, *42*, 360–370. [CrossRef]
21. Yan, B.; Zhang, J.; Mi, S.; Fan, G.; Wang, C.; Wang, M.; Yan, F. Simulation and experimental validation of the JCR-CNN model for hybrid modes recognition and decomposition in few-mode fibers. *Opt. Laser Technol.* **2024**, *175*, 110780. [CrossRef]
22. Yan, B.; Zhang, J.; Mi, S.; Wang, M.; Wang, C.; Fan, G.; Zhang, P. MT-SCUNet: A hybrid neural network for enhanced mode decomposition in optical fibers. *Opt. Fiber Technol.* **2025**, *93*, 104196. [CrossRef]
23. Jiang, M.; Lu, Z.; An, Y.; Zhi, F.; Li, J.; Huang, L.; Su, R.; Ling, Y.; Yang, J.; Zhou, P. On the robustness of mode decomposition deep neural network to adversarial perturbations. *Proc. SPIE* **2025**, *13544*, 135441T. [CrossRef]
24. Hendrycks, D.; Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
25. Zhang, R. Making convolutional networks shift-invariant again. In *Proceedings of the 36th International Conference on Machine Learning*; Proceedings of Machine Learning Research: Cambridge, MA, USA, 2019; Volume 97, pp. 7324–7334.
26. Wang, A.; Chen, H.; Lin, Z.; Han, J.; Ding, G. RepViT: Revisiting mobile CNN from ViT perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024; pp. 15909–15920.
27. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
28. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2020; pp. 11534–11542.

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.