IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 February 2025, accepted 2 March 2025, date of publication 10 March 2025, date of current version 27 March 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3549673

RESEARCH ARTICLE

DSI-ResCNN: A Framework Enhancing the Error-Tolerance Capacity of DNA Storage for Images

CIHAN RUAN[1], (Graduate Student Member, IEEE), LIANG YANG[2], RONGDUO HAN[2], (Graduate Student Member, IEEE), SHAN GAO[3], HAOYU WU[4], QIMING YUAN[1], (Student Member, IEEE), YANTING GUO[5], AND NAM LING[1], (Life Fellow, IEEE)

[1]Department of Computer Science and Engineering, Santa Clara University, Santa Clara, CA 95053, USA
[2]College of Software, Nankai University, Tianjin 300071, China
[3]College of Life Sciences, Nankai University, Tianjin 300071, China
[4]Roku, Inc., San Jose, CA 95110, USA
[5]School of Mathematics and Statistics, Yunnan University, Kunming, Yunnan 650504, China

Corresponding author: Nam Ling (nling@scu.edu)

ABSTRACT Traditional silicon-based storage technologies have reached a bottleneck with the dramatic increase in global data storage demands. High energy consumption, limited natural resources, and environmental concerns are some of the key reasons. DNA storage emerges as a promising alternative method, offering relatively high density and long-term stability. However, this storage method still faces considerable challenges. One such challenge pertains to errors introduced during the DNA storage process, which require the development of information subsystems with enhanced error-tolerance capabilities. In the present study, we propose the integration of two modules into information subsystems to enhance their error-tolerance capacities for DNA storage of images. These two modules, namely the residual convolutional neural network (ResCNN) module and the DNA sequence dropout control (SDC) module, collectively constitute a framework, named DNA storage of images with residual CNN (DSI-ResCNN). DSI-ResCNN enhances the fidelity of images recovered from DNA sequences generated under error-prone conditions, thereby substantiating its potential as an effective solution for mitigating the adverse impacts of errors.

INDEX TERMS DNA storage, deep learning, CNN, high-throughput sequencing, synthetic biology.

I. INTRODUCTION

The annual volume of the global datasphere will increase to 175 ZB by 2025, as projected by the International Data Corporation [1]. This exponential growth has led to a surge in the demand for data storage and maintenance. Consequently, it has led to a range of challenges, including the depletion of high-purity silicon resources, a significant increase in energy consumption, and the exacerbation of environmental waste associated with conventional data storage methodologies [2].

Among this dramatically increasing data volume, a substantial proportion, comprising archived images and videos, is typically classified as "cold data", which are rarely accessed. DNA storage represents a promising alternative for the storage of such cold data. This storage method boasts several notable advantages, such as its high density and long-term stability. Although advances in DNA synthesis technologies [3] and high-throughput sequencing [4] have substantially propelled the development of DNA storage, it still faces considerable challenges. These challenges include the high cost associated with the process, the relatively slow input/output (I/O) speeds, limited random access [5], [6], and the inherent susceptibility to errors induced during the DNA storage process [7], [8].

To address the challenge associated with errors, various biological solutions have been adopted, including the use

The associate editor coordinating the review of this manuscript and approving it for publication was Faissal El Bouanani.

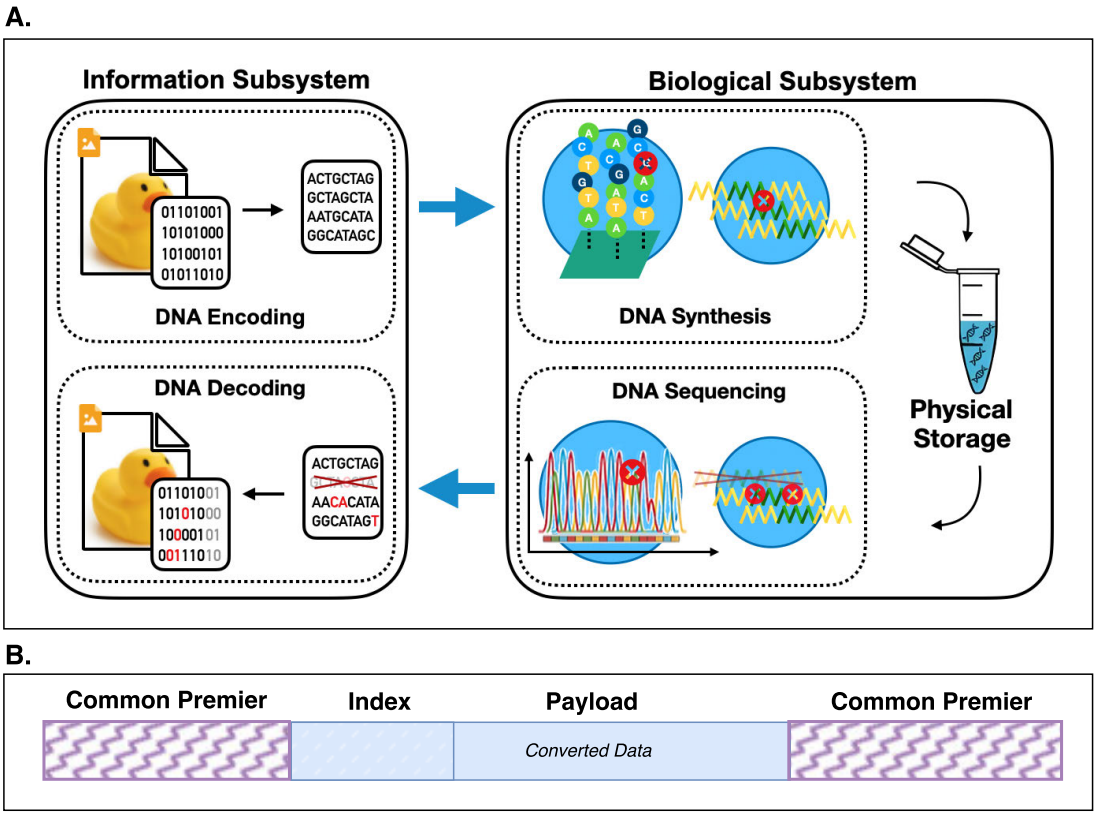


FIGURE 1. DNA Storage System Overview and Data Structure
(A) A multitude of errors are induced during the DNA storage process. These errors can be broadly categorized into two distinct types (indicated by red cross icons), which are detailed in the context. (B) Structure of a short DNA sequence used for data storage.

of biological redundancy. These biological solutions, while effective in reducing the errors, incur higher costs and lower efficiency. In contrast to biological solutions, several components, such as those using error correction code (ECC) [9], strand overlapping [10], and duplicate & voting [11], have been developed to improve the error tolerance capacities of the information subsystems for DNA storage. However, these components have been designed, based on a general model designed to handle multiple types of data, without specifically considering the properties of image data. Consequently, when applied to image storage, they can result in low fidelity or high redundancy [12].

The objective of this study was to develop an information subsystem with enhanced error-tolerance capabilities for DNA-based image storage. In our previous work [13], we proposed a decoder based on residual CNN (ResCNN) for image DNA storage, which applies its denoising and upsampling capabilities to improve both error tolerance and compression performance. This approach demonstrated promising results in handling certain types of error commonly encountered in DNA storage.

However, ResCNN alone is insufficient to address all types of errors in DNA storage, particularly when data experiences more complex or severe distortions. Moreover, in our previous study, we mainly used Gaussian noise to simulate errors, which does not comprehensively capture the full spectrum of errors that occur in DNA storage. These limitations motivated us to explore further improvements in error tolerance and a more comprehensive evaluation of DNA storage errors.

In this study, we introduce an improved framework that incorporates a new design to enhance the system's ability to handle challenging error conditions. Additionally, we conduct a broader analysis of error types, extending beyond Gaussian noise, to provide a more realistic assessment of error tolerance in DNA-based image storage. Through extensive simulations, we demonstrate that our improved system significantly enhances the fidelity of recovered images under various error-prone conditions.

The remainder of the paper is organized as follows. Section II presents the basic knowledge and related work related to DNA storage research. Section III introduces the methodology of DSI-ResCNN. Section IV details the results of this experiment and discusses the effectiveness of our approach in achieving reliable storage of image data on DNA. Finally, Section V concludes with the key findings of our research and discusses the implications of our proposed DNA decoding strategy on image storage.

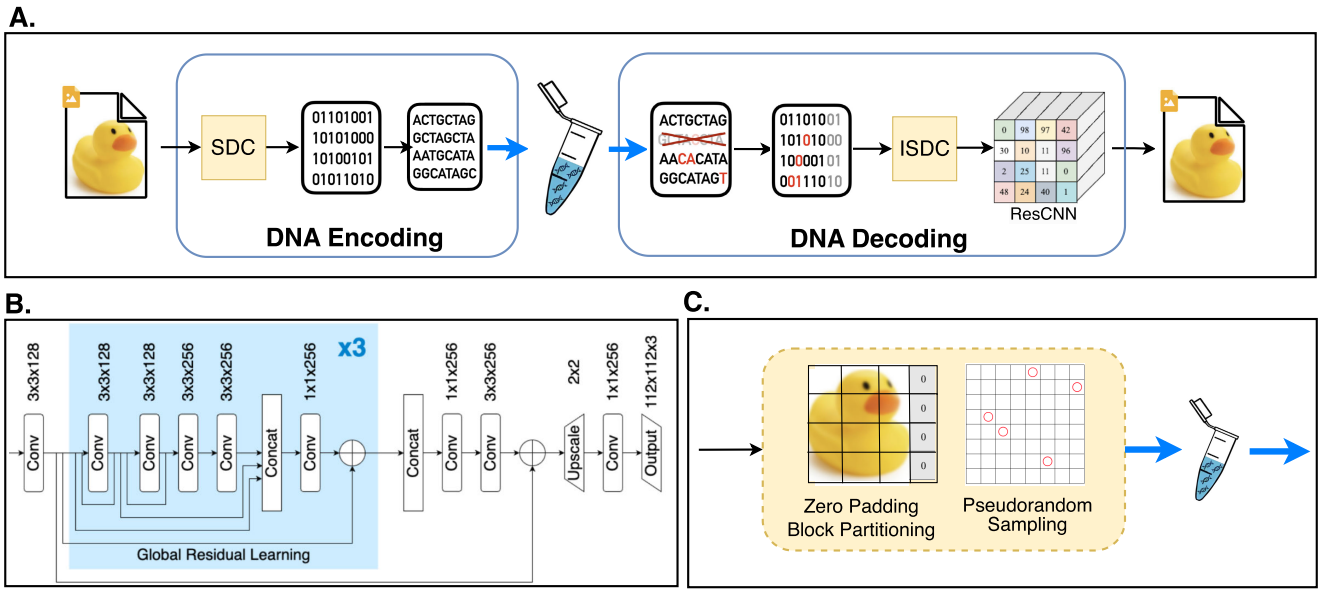


FIGURE 2. **Incorporation of DSI-ResCNN into a DNA storage system$A) Overview of DSI-ResCNN integration into a DNA storage system. The framework consists of two key modules: the sequence dropout control (SDC) module for handling Type II errors during encoding, and the residual convolutional neural network (ResCNN) module for addressing Type I errors during decoding (Fig. 1). (B) ResCNN was first proposed in our previously study [13] (C) The SDC module has two parts and its encoding part (in blue box) which is responsible for zero-padding block partitioning and pseudorandom sampling.**

II. BASIC KNOWLEDGE

A. DNA STORAGE

A DNA storage system typically comprises a biological subsystem and an information subsystem (Fig. 1(A)). The biological subsystem is responsible for a series of biological processes, notably DNA synthesis, physical storage, and DNA sequencing. Automated DNA synthesizers perform DNA synthesis to create DNA fragments. The fragments are then kept under low temperature and dry conditions to ensure optimal physical storage. DNA sequencing technologies are used to read DNA fragments and generate DNA sequences.

The information subsystem is responsible for all aspects of data processing and analysis, with particular emphasis on data encoding, decoding, and error correction. This facilitates efficient reading, writing, and storage of data in DNA sequences. Specifically, the data encoding process converts binary data into DNA sequences, while data decoding performs the reverse process. In addition, the information subsystem includes database management components for metadata storage and automation components that improve the efficiency of sample and data processing.

In an actual DNA storage system, image data are usually partitioned and packaged into short DNA sequences of approximately 150 - 300 base pairs (bp) in length (Fig. 1(B)). Generally, such a DNA sequence consists of three fundamental components: common primers, address, and data payload. Common primers are short DNA sequences, typically around 18 - 25 bp, that function as initiation sites for the synthesis or amplification of target DNA sequences. The address within a DNA sequence serves as a unique identifier, specifying the position or ordinal number of the data payload, thereby facilitating the organization of the partitioned image data. The data payload represents the actual image data that have been encoded into the DNA sequence, corresponding to the binary encoded image.

B. ERRORS DURING DNA STORAGE

Despite substantial advancement, DNA storage still faces considerable challenges (as discussed in the introduction). One such challenge pertains to errors introduced during the DNA storage process, which involve a series of stages, notably DNA synthesis, physical storage, and DNA sequencing. In addition, errors can arise from improper experimental practices, such as suboptimal reagent use [14], poor handling conditions [15], and improper equipment adjustment [16].

A multitude of errors are induced during the DNA storage process, and the underlying causes are complex and multifaceted (Fig. 1 (A)). These errors can be broadly categorized into two different types. Type I errors originate from small-scale mutations, specifically insertions, deletions, and substitutions (IDS) within DNA fragments. Many existing error-tolerant mechanisms, such as the design of a coding scheme with logical redundancy and error correction codes [9], [17], [18], are effective in mitigating these errors.

Type II errors manifest themselves as sequence dropout, where entire DNA sequences are lost due to various factors. Unlike Type I errors, sequence dropout leads to complete information loss rather than small-scale modifications, making it particularly challenging to recover lost data. As a result, developing effective error-tolerant strategies to handle sequence dropout has become a critical and enduring research challenge in the field of DNA storage. To address these

challenges, many biological solutions have been adopted, including optimizing synthesis or sequencing protocols, improving purification techniques, and employing biological redundancy [19], [20]. In contrast to biological solutions, significant contributions can also be made by the development of information subsystems with enhanced error-tolerance capabilities. Therefore, minimizing the adverse impacts of both types of errors, particularly sequence dropout, remains a key focus of ongoing research.

Although previous studies have explored various error-tolerant mechanisms, they often fail to fully address sequence dropout or rely on oversimplified error models. This study aims to bridge this gap by introducing a more comprehensive framework that enhances error tolerance and provides a more realistic evaluation of DNA storage errors.

III. METHODOLOGY

A. OVERALL ARCHITECTURE OF DSI-ResCNN

DSI-ResCNN consists of two modules: a residual convolutional neural network (ResCNN) module and a sequence dropout control (SDC) module (Fig. 2). While ResCNN (introduced in our previous study [13]) addresses Type I errors, SDC (proposed in this study) handles Type II errors, which typically cause more severe impacts on image recovery. Compared to IDS (type I errors), sequence dropout (type II errors) is more likely to result in severe adverse impacts on image recovery. Sequence dropouts corresponding to an individual or a small number of pixels can potentially be recovered through information-based methodologies, such as ResCNN. However, sequence dropouts corresponding to pixels that are concentrated in an area are unrecoverable.

Specifically considering the properties of image data, the SDC module is designed to circumvent the potential loss of pixels that are concentrated in an area in the event of sequence dropout. SDC has two parts: its encoding part (Fig. 2(C)) randomly reallocates the partitioned data of a block into different blocks prior to their encoding into DNA sequences, while its decoding part serves to recover the original order of the partitioned data for image recovery by ResCNN.

B. THE ResCNN MODULE

The ResCNN module, first introduced in our previous study [13], serves as the primary component for handling Type I errors in our framework. Based on the architecture of residual dense networks (RDN) [21], ResCNN was specifically adapted for DNA storage applications, where it demonstrated a strong capability in noise reduction and enhancement of image fidelity.

The network's effectiveness in handling IDS errors stems from its dense connectivity structure, where feature information flows directly between layers through carefully designed pathways. To optimize performance in the DNA storage context, we streamlined the architecture by pruning redundant connections and simplifying residual blocks, resulting in a balanced design that maintains error correction capability while minimizing computational overhead.

Algorithm 1 SDC for Dropout Recovery

Require: Input image I, Block size B, Dropout rate r, Chain length L, Seed s

Ensure: Processed image with applied dropout and SDC

1: Divide the image I into non-overlapping blocks of size $B \times B$
2: **for** each block b in I **do**
3: Randomly sample L pixels from block b using seed s
4: Apply dropout with rate r to sampled pixels
5: **end for**
6: Reconstruct the image $I_{\text{reconstructed}}$ from processed blocks
7: **return** processed image $I_{\text{reconstructed}}$

C. SEQUENCE DROPOUT CONTROL (SDC) STRATEGY

The challenge of sequence dropout in DNA storage presents a unique problem that traditional error correction methods struggle to address. Although ResCNN effectively handles Type I errors, sequence dropout can result in complete loss of contiguous data blocks, making recovery through neural network approaches alone insufficient. Existing solutions like ECC, while effective for general error correction, often rely on redundancy and backup copies, increasing storage costs without fundamentally addressing the concentrated data loss problem.

Our proposed SDC module takes a fundamentally different approach: instead of trying to recover lost sequences, it preemptively transforms potential concentrated losses into distributed patterns that are more amenable to recovery. This strategy is particularly effective for image data, where spatial correlations can be leveraged for reconstruction.

1) SDC IMPLEMENTATION DETAILS

The key innovation of SDC lies in its pseudorandom sampling strategy, which deliberately disperses spatially adjacent pixels across different DNA sequences. This approach ensures that, even if entire sequences are lost, the impact on any particular image region remains limited. The process is systematic yet flexible, allowing for different block sizes and sampling rates to accommodate varying storage requirements.

The complete implementation of SDC can be formalized as Algorithm 1, which details the step-by-step process to transform concentrated potential losses into distributed patterns. The algorithm takes as input the original image along with key parameters that control the distribution strategy, including block size, dropout rate, chain length, and a seed for reproducible randomization.

Fig.2 illustrates how SDC integrates into the DNA storage pipeline, replacing localized data loss with statistically manageable noise. The process involves the following steps:

1) *Block Partitioning*
Given an input image I of size $M \times N$, we first apply zero padding to ensure that the dimensions of the image are multiples of the size of the block b. The padded image I' is defined as:

$$I' \in \mathbb{R}^{M' \times N'} \tag{1}$$

where $M' = \text{ceil}(M/b) \times b$ and $N' = \text{ceil}(N/b) \times b$. After padding, the image is partitioned into $B = \frac{M' \times N'}{b^2}$ blocks of size $b \times b$.

2) *Pseudorandom Sampling with Index-Based Seed* After partitioning, the key step is to establish a reproducible random sampling pattern. For each block B_k ($k = 1, \ldots, B$), we use its index k as a seed to generate a pseudorandom sequence using a linear congruential generator (LCG). This sequence determines the sampling positions within the block.
Then, for each block B_k, we sample s pixels based on the generated sequence, forming a sampled set S_k:

$$S_k = \{p[i,j] \mid (i,j) \in P_k, p[i,j] \in B_k\} \tag{2}$$

where P_k is the set of s pseudorandomly selected pixel positions within block B_k, and $p[i,j]$ represents the pixel value at position (i,j).

3) *DNA Encoding and Storage*
The sampled values need to be converted into DNA sequences using a quaternary encoding scheme. Our structure is compatible with any binary-to-quaternary encoding scheme that satisfies biological constraints of DNA sequences. Here, we apply our previously proposed Dynamic DNA Fountain Code [22], as it provides built-in error resistance and achieves high coding efficiency while maintaining sequence balance:

$$\text{DNA}_k = \text{Encode}(S_k) \tag{3}$$

This choice adopts three key advantages of Dynamic DNA Fountain Code: inherent robustness to synthesis errors through its fountain code properties, flexibility in sequence generation that complements our SDC distribution strategy, and enhanced compression efficiency through its advanced encoding mechanisms. These characteristics naturally align with and further strengthen our error-tolerant framework.

4) *DNA Decoding and Image Recovery*
The decoding process first establishes the inverse mapping from DNA sequences back to their original spatial locations using the stored address information. Each recovered sequence contributes to reconstructing specific block positions in the final image. For missing or corrupted sequences, ResCNN uses both local and global context for recovery. This process can be formalized as:

$$I''[i,j] = \begin{cases} \text{Decode}(\text{DNA}_k[p]), & \text{If}(i,j) \in P_k \text{ and DNA}_k \text{ is available} \\ \text{ResCNN}(I''), & \text{Otherwise} \end{cases} \tag{4}$$

where *Decode* is the inverse function of *Encode*, and ResCNN denotes the residual convolution neural network that estimates the missing values by taking as input the spatial information of the surrounding pixels. ResCNN works effectively with lost data to reconstruct the dropped values with regard to both the local and global context. In fact, this approach has been proven to support relatively high dropout rates while the visual fidelity of the restored image maintains a high level of visual fidelity, making it highly compatible with subsequent processing tasks in DNA-based image storage.

5) *Addressing Principle*
To ensure accurate image reconstruction, our approach restores the spatial ordering of the encoded DNA sequences using a permutation-based mapping strategy. The encoded DNA structure includes the following key components:

- **Address**: The permutation-mapped addressing ensures that each DNA sequence can be mapped back to its original relative spatial position. Unlike traditional sequential storage, this method prevents catastrophic data loss by distributing data redundantly across multiple locations.
- **Random Seed**: The random seed governs the sampling pattern within each block during encoding, ensuring that data dispersal can be replicated during decoding. This improves robustness against data loss and enhances reconstructability.
- **Channel Identification**: Since image data consists of multiple color channels, our addressing strategy explicitly encodes the channel index within the DNA sequence. This ensures proper separation and reconstruction of individual color channels during decoding, with each channel being stored in separate but traceable DNA sequences.
- **Data Payload**: Contains the actual image data sampled using the specified random seed pattern. The payload is organized to maintain optimal balance between storage efficiency and error resilience, with the sampling pattern ensuring even distribution of spatial information.

The overall DNA oligo is structured as follows:

$$\textit{Address} - \textit{Random Seed} - \textit{Channel ID} - \textit{Data Payload} \tag{5}$$

Using this structured addressing, the system follows a two-stage reconstruction process. First, the DNA sequences are reordered based on their address information, establishing the correct spatial arrangement of

blocks. Then, within each block, the random seed is used to reconstruct the original sampling pattern. If dropout occurs, missing data points can be recovered through a combination of spatial redundancy and ResCNN-based restoration, using both local and global image context.

2) THEORETICAL ANALYSIS OF SDC

The effectiveness of SDC can be analyzed through information theory principles, particularly focusing on how information loss patterns affect recovery. In traditional storage systems without SDC, when data loss occurs, it typically affects contiguous blocks of information. If a block of size k is lost, the loss of information is:

$$L_{block} = k \cdot H_{local} \tag{6}$$

where H_{local} represents the entropy of the lost region. This concentrated loss pattern eliminates all contextual information within the affected area, making recovery extremely difficult.

In contrast, SDC distributes losses across multiple regions, preventing entire areas from becoming unrecoverable. The total distributed loss is:

$$L_{distributed} = \sum_{j=1}^{n} (k/n) \cdot H_j \tag{7}$$

where H_j represents the entropy of each affected subregion. Because natural images exhibit high spatial correlation, neighboring regions share significant contextual information. Consequently, we observe:

$$L_{distributed} < L_{block} \tag{8}$$

This improvement arises because of the following.

1) **Spatial Correlation Preservation**: SDC ensures that even when data is lost, some portion of the local information remains intact, making it easier for neural network-based recovery methods to interpolate missing values.
2) **Context Distribution**: Instead of losing entire sections of an image, the SDC disperses the losses, ensuring that no single region loses all its information, reducing severe visual degradation.

These theoretical advantages directly inform our parameter selection in Algorithm 1. The size of the block B is optimized to balance the distribution of information and recoverability: larger values B spread errors more widely but reduce the retention of the local structure, while smaller values B limit the effectiveness of the distribution. Similarly, the sample rate, controlled by the length of the chain L, is set to maximize distribution while maintaining storage efficiency.

SDC's distributed storage approach helps mitigate the impact of sequence dropout by transforming concentrated losses into scattered patterns. By preserving partial information across different locations, the system maintains a better spatial context for image recovery, reducing the adverse effects of sequence loss on reconstruction quality.

3) IMPLEMENTATION CONSIDERATIONS

The practical implementation of SDC requires careful design to balance storage efficiency, error resilience, and computational feasibility. Our approach enables flexible parameter tuning based on specific operational requirements.

- **Storage Density Optimization**: The sampling rate can be adjusted to achieve an optimal balance between error resilience and storage efficiency. Higher sampling rates improve error tolerance but reduce storage density, whereas lower rates maximize storage capacity at the cost of increased sensitivity to data loss.
- **Error Pattern Adaptability**: Type II errors can manifest in various patterns. The SDC adaptive control mechanism allows parameter adjustment to address different dropout scenarios, from random losses to systematic sequence failures.
- **Recovery Priority Control**: For applications demanding rapid local recovery, the distribution pattern can be modified to maintain stronger spatial coherence while preserving error protection. This design ensures that adjacent blocks retain sufficient correlated information for effective ResCNN-based reconstruction, significantly improving recovery success rates.

This flexibility in implementation, combined with the computational efficiency of random sampling, enables SDC to enhance error tolerance without introducing significant overhead. Furthermore, by operating in the encoding stage, the SDC maintains full compatibility with existing DNA storage pipelines, facilitating practical deployment in real-world systems.

D. DATA AUGMENTATION AND NETWORK TRAINING

Type I and II errors in DNA data storage are highly dependent on the synthesis technology, sequencing platform, and environmental conditions. Unlike traditional computer storage systems where the error patterns are well defined and reproducible, DNA storage errors vary widely between different experimental setups. As a result, publicly available DNA storage datasets are not only limited in number but also lack standardization for large-scale training of deep learning models. In addition, the limited availability and high cost of DNA synthesis and sequencing further restrict access to datasets. Although some simulation tools have attempted to model DNA storage errors, their underlying training data are aggregated from different independent studies and literature sources, resulting in significant heterogeneity and inconsistency.

To overcome these challenges, we propose a noise-based error simulation method to approximate the DNA storage error distribution, allowing us to systematically train and evaluate fault-tolerant neural networks in a controlled and reproducible manner. Although SDC modifies the information distribution at the encoding stage, our training strategy focuses on developing a network that can handle various forms of data degradation encountered in DNA storage.

To improve the robustness of ResCNN, we apply the following augmentation strategies:

- **Structured Noise Perturbations**: We introduce controllable noise distortions to approximate the effects of DNA storage errors. Frequency domain analysis suggests that the distortions induced by SDC exhibit statistical properties similar to specific noise patterns, particularly Gaussian and uniform distributions. The selection of these noise models is motivated by their ability to capture different aspects of DNA storage errors; details are described in Section IV.
- **Progressive Multi-level Training**: To improve resilience to different error severity levels, the network is trained with gradually increasing noise intensities. This ensures that the model remains robust under varying levels of data degradation.
- **Spatial Data Augmentation**: Standard image augmentation techniques, such as random crops, flips, and rotations, are used to enhance robustness and ensure that the model generalizes well across different spatial error configurations.

Since our primary objective is to develop a neural network that is robust to image data degradation, our approach focuses on building feature extraction capabilities that are invariant to a wide range of perturbations. Experimental validation demonstrates that our trained model generalizes well to DNA storage scenarios and can effectively restore images under various error conditions.

IV. EXPERIMENTAL RESULTS

A. EXPERIMENTAL RESULTS ON SDC

1) OBJECTIVE

The primary goal of this experiment is to evaluate the effectiveness of SDC in mitigating Type II error in DNA storage. Specifically, our aim is to demonstrate that:

- SDC effectively disperses Type II error, preventing highly localized missing regions.
- The SDC-induced dropout distribution statistically resembles Gaussian and uniform noise, making it more manageable for recovery algorithms.
- SDC is robust across different input resolutions, ensuring applicability in various DNA image storage settings.

2) EXPERIMENTAL SETUP

We conducted the following key experimental steps:

- **Data Preparation**: We used images from the Set-14 dataset, which were processed into nonoverlapping blocks and encoded with SDC.
- **Block Sizes**: We tested partition block sizes of 4×4, 8×8, 12×12, and 16×16, evaluating their effect on Type II error dispersion.
- **Sequence Lengths**: We simulated Type II error across sequences ranging from 80 to 240 nucleotides to evaluate SDC under different constraints.
- **Dropout Rates**: We tested dropout rates from 5% to 20%, covering different levels of sequence loss.

3) PERFORMANCE EVALUATION

To comprehensively assess the performance of SDC, we measured:

- **Peak Signal-to-Noise Ratio (PSNR)**: To quantify how well image information is preserved post-dropout and recovery.
- **Loss Distribution Visualization**: To confirm that SDC transforms structured dropout into noise-like distributions.
- **Noise Density Consistency**: To evaluate how evenly SDC disperses missing data across different settings.

4) RESULTS AND ANALYSIS

a: EFFECT OF SDC ON IMAGE QUALITY

SDC demonstrates significant robustness in mitigating Type II error by effectively redistributing data loss. As shown in Fig. 3, SDC disperses data loss throughout the image, preventing localized missing regions and improving recoverability. In contrast, sequential encoding leads to highly concentrated dropout, as evident from Loss Maps, making recovery significantly more challenging. The distributed nature of SDC ensures that missing pixels can be estimated from neighboring structures rather than entire regions being lost.

To further quantify this impact, we analyze PSNR, loss density, and autocorrelation patterns:

- **PSNR Performance Analysis**: Under a dropout rate of 0.2, both encoding methods initially exhibit low PSNR values (10 dB), as expected due to significant pixel removal. However, during recovery, a major difference emerges: basic interpolation recovery (without neural networks) improves PSNR for SDC-processed images by +7.61 dB, reaching 21.84 dB, whereas sequential encoding only improves to 14.23 dB. This substantial gain validates the core principle of SDC: transforming concentrated data loss into stochastic dispersed noise significantly enhances recovery potential.
- **Loss Density Map (Bottom Row, Fig. 3)**: The sequential encoding Loss map exhibits highly localized structured dropout, causing large portions of the image to be completely erased. In contrast, the SDC Loss Map shows a more uniform distribution of missing pixels, ensuring that no single region is disproportionately affected.
- **Channel-wise Loss Maps (Middle Columns, Fig. 3)**: In sequential encoding, dropout exhibits a strong structured pattern, forming horizontal stripes that cause severe localized information loss. In SDC, the Type II error pattern is significantly more randomized, resembling natural noise distributions. This is confirmed by the following uniformity metrics: SDC achieves higher uniformity values (0.16 across all channels) compared to sequential encoding (0.025, 0.018, 0.011).
- **Auto-correction Map (Rightmost Columns, Fig. 3)**: The auto-correction map of the SDC exhibits a centralized and radially symmetric pattern, indicating evenly

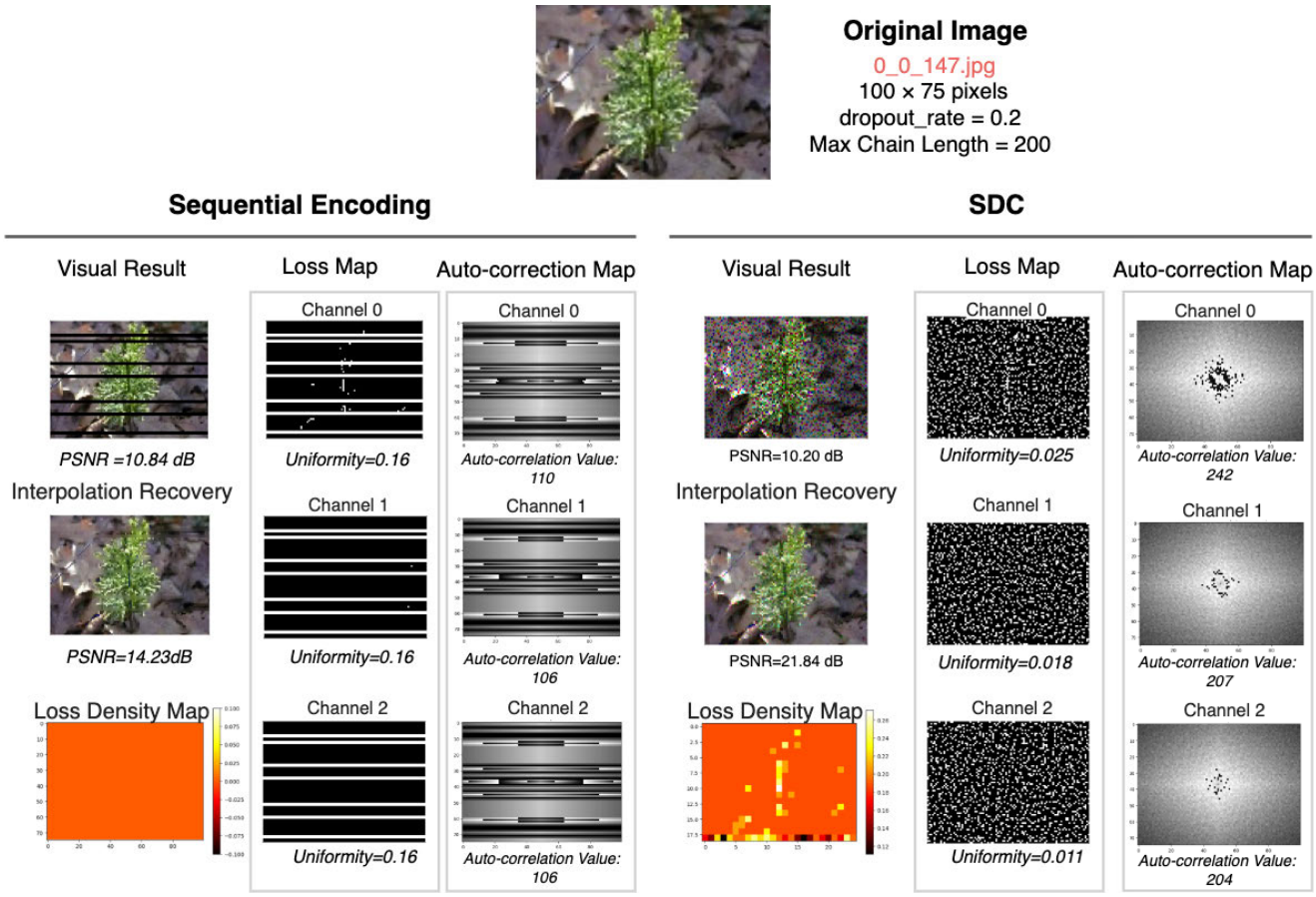


FIGURE 3. Comparison of sequential Encoding and SDC on Dropout Loss Dispersion. The left panel shows sequential encoding with a more localized dropout pattern, while the right panel illustrates SDC's more dispersed dropout distribution. Key metrics such as PSNR, uniformity, and auto-correlation demonstrate the improved potential for recovery under SDC, with a more evenly distributed loss across the image, which improves the conditions for interpolation recovery. The experiments were carried out with a sequence length of 200 nt (nucleotide units) and a dropout rate of 0.2.

distributed dropout, making it easier for recovery models to interpolate missing data. In contrast, sequential encoding shows distinct horizontal clusters, introducing correlated dropout patterns that are more difficult to restore. Quantitatively, SDC achieves significantly higher auto-correlation values (242, 207, 204) compared to sequential encoding (110, 106, 106), demonstrating better preservation of spatial structure.

The evaluation demonstrates that SDC significantly improves dropout distribution without increasing redundancy, making it highly effective for DNA-based image storage. The improved uniformity, PSNR, and autocorrelation properties ensure that missing regions can be better estimated, reducing image degradation. These results highlight SDC's core advantage: transforming catastrophic structured loss into noise-like patterns that are more recoverable using standard image restoration techniques.

b: EFFECT OF DROPOUT RATE ON IMAGE QUALITY

To evaluate how SDC mitigates the effects of dropout rate, we conducted controlled experiments with dropout rates ranging from 0.02 to 0.2. Fig. 4 presents the results on the input images after SDC processing, followed by a simple recovery based on interpolation. The key observations are as follows:

- **Dropout Distribution and Density**: As the dropout rate increases, the loss density maps reveal a clear pattern: sequential encoding produces large contiguous missing regions, while SDC ensures a more uniform dispersion of dropout loss. Even at high dropout rates (e.g. 18. 85%), SDC prevents catastrophic information loss in any single region.
- **PSNR Stability**: Despite increasing dropout rates, SDC maintains a relatively stable PSNR trend. For moderate dropout rates (5%-10%), the PSNR remains above 16 dB. At extreme dropout rates (20%), PSNR remains at 21.06 dB after basic interpolation, demonstrating that SDC preserves recoverability even under severe loss conditions.
- **Mean vs. Max Density**: As shown in the loss density maps, the mean density gradually increases with the dropout rate, reflecting the increasing probability of missing pixels. However, the maximum density remains well distributed, indicating that no single region

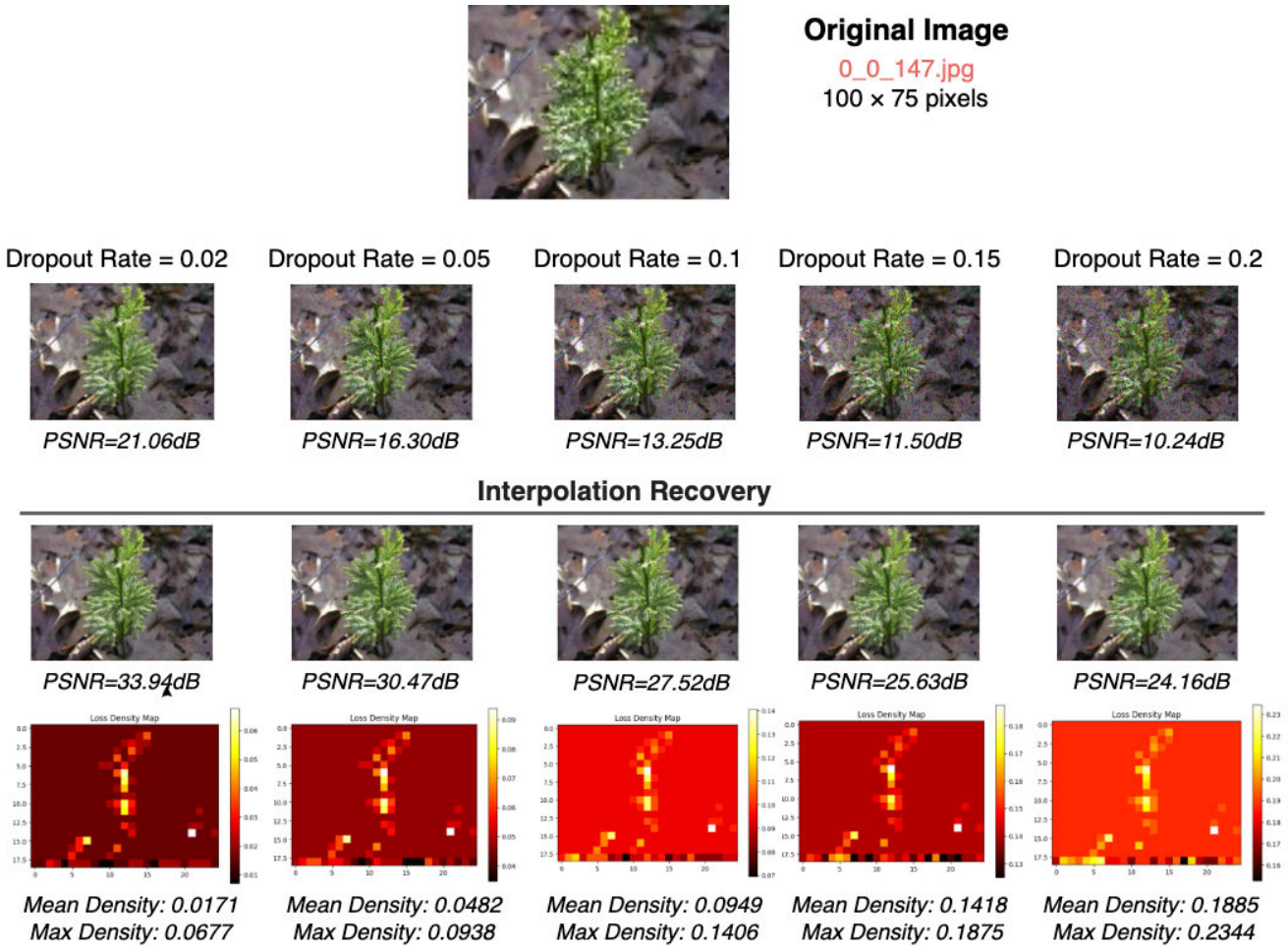


FIGURE 4. Impact of dropout rates on SDC-processed images and interpolation recovery. The first row shows SDC-encoded images under different dropout rates (0.02 to 0.2), while the second row presents their interpolation-based reconstruction. The third row visualizes the loss density maps, where SDC effectively disperses missing data, reducing localized dropout regions. Despite using only basic interpolation, the recovered images maintain reasonable visual fidelity, highlighting SDC's capability to mitigate structured data loss.

experiences extreme data loss, a crucial advantage of SDC over conventional encoding.

- **Basic Interpolation Recovery**: It is important to note that these results are achieved using only a basic interpolation method without advanced restoration techniques. Despite this, the SDC-processed images exhibit significantly improved recoverability. This strongly suggests that when combined with more sophisticated restoration networks (such as our proposed ResCNN), the potential recovery quality would be even higher.

These results confirm that SDC effectively transforms structured data loss into a more manageable stochastic noise pattern, improving resilience against DNA synthesis and sequencing dropout errors. Additionally, SDC provides adaptability by allowing fine-tuning of block size and dropout parameters based on the constraints of DNA synthesis and sequencing. This flexibility ensures that SDC can be optimized for various DNA storage scenarios, making it a practical and scalable solution.

c: SDC VALIDATION THROUGH NOISE PATTERN ANALYSIS

To validate the effectiveness of SDC in transforming structured data loss into distributed noise-like patterns, we perform a comparative analysis between SDC-processed images and well-understood noise distributions in both spatial and frequency domains (Fig. 5).

Spatial Domain Analysis The top row of Figure 5 presents the spatial domain comparison between an image processed with SDC and reference noise patterns.

- The PSNR similarity between SDC-processed images and Gaussian/uniform noise (46.38 dB and 48.14 dB, respectively) quantitatively confirms that SDC converts Type II error into more isotropic noise patterns.
- This transformation is crucial because it mitigates catastrophic block-wise information loss, allowing interpolation-based recovery methods to infer missing information more effectively.

Frequency Domain Validation The bottom row of Figure 5 shows the corresponding frequency spectra, further

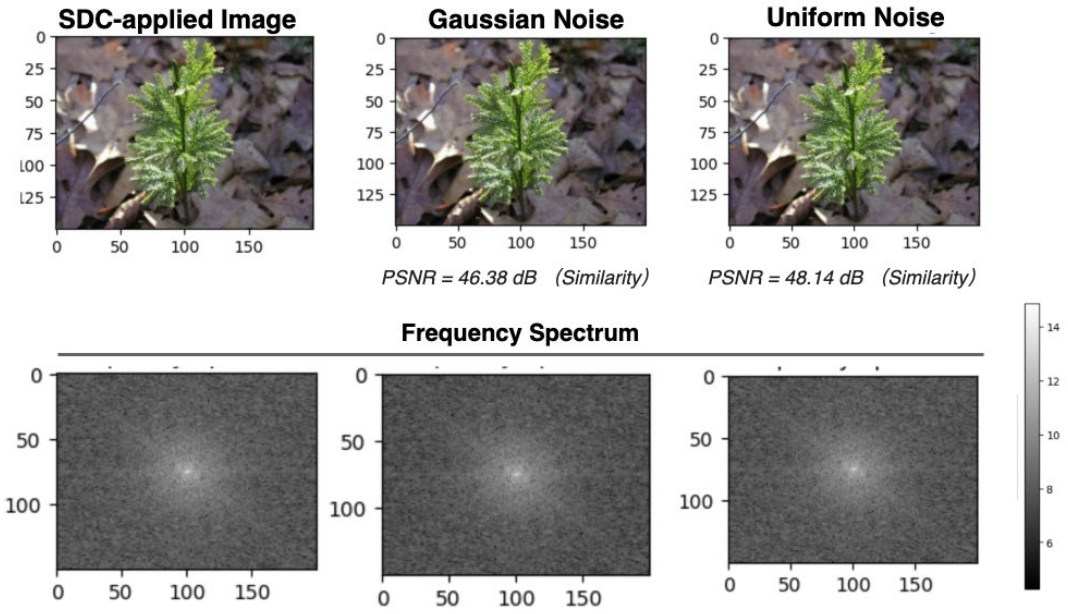


FIGURE 5. Comparison between SDC-processed images and reference noise patterns. Top row: SDC-processed image (left) exhibits high spatial similarity to Gaussian noise (middle, PSNR = 46.38 dB) and uniform noise (right, PSNR = 48.14 dB), suggesting that SDC redistributes structured Type II error into more isotropic noise patterns. Bottom row: Frequency spectrum analysis reveals that all three cases share similar radial energy distributions, validating SDC's ability to transform concentrated information loss into a well-distributed form that is statistically favorable for recovery.

validating the transformation of structured Type II error into noise-like characteristics.

- The frequency spectra of the SDC-processed images exhibit radial symmetry and energy distribution similar to Gaussian and uniform noise.
- This distribution suggests that SDC effectively scatters dropout pixels across different frequency bands, thereby preventing concentrated high-frequency loss, which is typically detrimental to image quality.
- A quantitative spectral correlation analysis shows high similarity between SDC-processed images and Gaussian/Uniform noise distributions (0.92 and 0.94 correlation coefficients, respectively), reinforcing the argument that SDC mitigates structured loss by transforming it into well-distributed noise patterns.

Discussion SDC provides a significant advantage in DNA-based image storage by redistributing Type II error into more recoverable noise-like structures. Instead of concentrated regions of missing data, the resulting dropout pattern is spread out in a pseudo-random manner, reducing local information loss and improving restoration feasibility.

Additionally, this transformation ensures compatibility with ResCNN-based recovery. Since many deep learning models are trained on natural images that inherently contain Gaussian and uniform noise components, SDC-preprocessed images are better aligned with conventional noise models, allowing for more effective reconstruction.

Notably, SDC does not eliminate the loss, but rather transforms it into a form that is easier to handle with standard restoration techniques. This transformation ensures that even at high loss rates, the loss patterns remain statistically favorable for interpolation and subsequent neural network ResCNN-based restoration.

B. EXPERIMENTS ON ResCNN

1) EXPERIMENTAL SETUP

In this section, we present the experimental setup and evaluation of our re-trained ResCNN model, incorporating a more comprehensive DNA storage error model as described in Section III. Although previous approaches have often relied on Gaussian noise to approximate DNA storage errors [13], this oversimplification does not fully capture the nature of DNA-specific distortions. Traditional communication noise models, such as additive Gaussian noise, primarily introduce continuous-valued perturbations, whereas DNA storage errors involve insertion, deletion, and substitution (IDS) errors, as well as sequence dropout. These errors fundamentally differ from signal noise in that they disrupt the structural integrity of the stored information, rather than merely altering pixel intensities.

To better model the real-world challenges of DNA data storage, we extend the ResCNN training pipeline by incorporating a broader range of error patterns that more accurately reflect the types of degradation observed in DNA synthesis, storage, and sequencing. This enhanced training process enables ResCNN to generalize more effectively to DNA storage scenarios, where errors are discrete, position dependent, and highly variable between different synthesis technologies.

FIGURE 6. Super-resolution performance at ×4 scaling factor across diverse test images. The network effectively reconstructs structural and textural details in different scenarios, including text (mailbox), complex textures (stuffed animal), motion (skateboard), geometric patterns (hourglass), low-light conditions (portrait), and fine natural details (flower). PSNR and SSIM results confirm consistent high-quality restoration.

Our experiments were carried out on a dataset of 8,600 images selected from multiple standard benchmarks, ensuring a diverse range of visual content. The dataset was randomly partitioned into 70% for training, 10% for validation, and 20% for testing. Each training batch consisted of 16 randomly extracted low-resolution RGB (LR), with random enhancement (flips and rotations) applied to enhance generalization.

To adapt ResCNN to a more realistic DNA storage environment, we introduced additional corruption patterns in the training process to model a broader range of DNA storage errors. These distortions account for different types of degradation commonly observed in DNA-based storage, allowing ResCNN to generalize more effectively under real-world conditions.

The model was implemented using PyTorch 1.9.0 and trained on an NVIDIA RTX 3090 GPU. We used Adam optimizer with an initial learning rate of 10^{-4} and a batch size of 32. Training was carried out for up to 100 epochs, with early stopping based on validation loss to prevent overfitting.

2) BASELINE SUPER-RESOLUTION FOR DNA STORAGE

Before evaluating the denoising capability of ResCNN, we first assess its super-resolution performance in a noise-free environment. ResCNN is designed to jointly enhance image resolution and correct DNA-storage-induced distortions, making it highly suitable for compressed DNA image storage.

Since storing high-resolution images in DNA is impractical due to synthesis and sequencing constraints, we propose to use ResCNN's super-resolution capability to store low-resolution images while reconstructing high-fidelity outputs. This approach balances storage efficiency and image quality, reducing the required synthesis cost while maintaining retrieval quality.

Fig. 6 presents super-resolution results at ×4 scaling across a range of image types. The dataset includes:

- Complex textures (stuffed animal)
- Dynamic motion (skateboard)
- Geometric structures (hourglass)
- Lowlight conditions (portrait)
- Natural textures (flower)

Across all cases, ResCNN reconstructs fine details with consistently high PSNR values (28.62-39.97 dB) and SSIM (0.8783-0.9783). These results validate that even without explicit noise correction, ResCNN can successfully restore high-quality images from compressed representations, making it ideal for DNA-based storage workflows.

3) NETWORK RETRAINING FOR DNA STORAGE NOISE

Although ResCNN has demonstrated strong super-resolution capabilities, DNA-based image storage introduces unique error patterns that differ significantly from conventional image transmission noise. Unlike Gaussian noise, which is commonly used in image restoration, DNA storage errors primarily involve insertion, deletion, and substitution (IDS) errors, resulting from synthesis, storage degradation, and sequencing processes.

To adapt ResCNN to DNA storage conditions, we re-trained it using synthetic datasets with multiple noise types

TABLE 1. Mapping of common noise types in image processing to DNA storage error patterns. This mapping serves as a reference for evaluating recovery algorithms under simulated conditions.

Noise Type	Corresponding DNA Storage Error	Scenario
Salt-and-Pepper	Bit Flips	Base information turns into extreme values
Gaussian	Continuous Errors	Continuous accumulation of random errors
Uniform	Random Base Errors	Randomly distributed base modifications
Periodic	Periodic Distortions	Systematic errors in DNA synthesis
Poisson	Discrete Random Errors	Signal-related errors during DNA synthesis

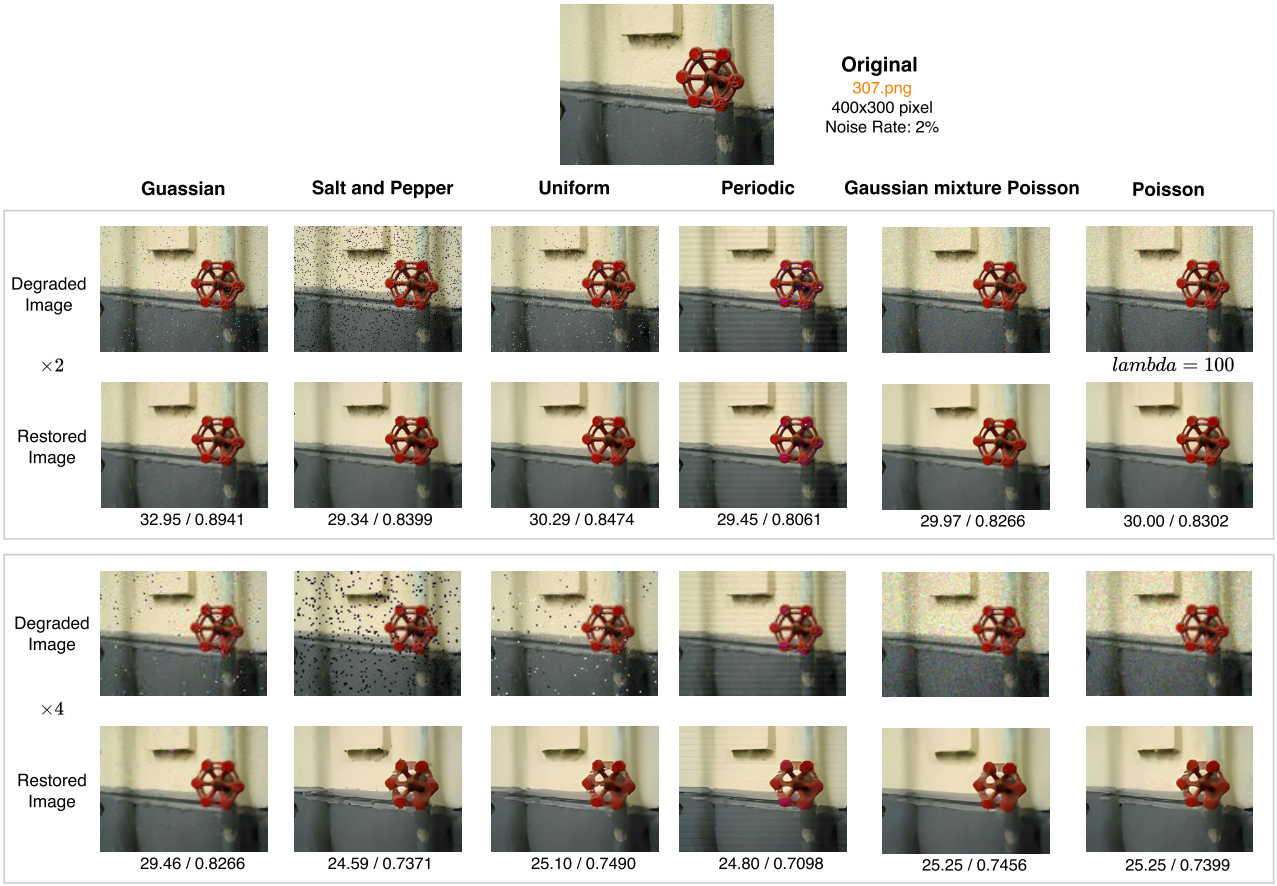


FIGURE 7. Comparative analysis of DR-ResCNN performance across various noise types simulating DNA storage errors. The original 400 × 300 pixel image (top) is degraded with 2% noise of different types (Gaussian, Salt and Pepper, Uniform, Periodic, Gaussian mixture Poisson, and Poisson). Restored images are shown for both ×2 and ×4 upscaling, with the corresponding PSNR/SSIM values. Note the consistent recovery quality across different noise types, particularly evident in the ×2 upscaling scenario.

that statistically resemble real DNA storage errors. As shown in Table 1, each noise type serves as a proxy for different error mechanisms encountered in DNA-based data storage:

- **Gaussian Noise**: Models continuous sequencing fluctuations.
- **Salt-and-Pepper Noise**: Simulates bit flips during DNA readout.
- **Uniform Noise**: Represents random base modifications.
- **Periodic Noise**: Mimics systematic synthesis distortions.
- **Poisson Noise**: Captures signal-related sequencing degradation.

These noise types were introduced at varying intensities (1% to 5%) to simulate different levels of DNA sequence corruption.

Training Procedure: To incorporate these noise models, we extended the original ResCNN training pipeline by augmenting the dataset with synthetic distortions at controlled levels. The network was optimized with Adam Optimizer with an initial learning rate of 10^{-4}, which was lowered by a factor of 0.5 every 2×10^5 iteration. Training was performed over 100 epochs, leveraging standard image augmentation techniques (flips, rotations) to improve generalization.

TABLE 2. Comparison of DNA storage methods from various literature sources, highlighting features such as redundancy level, error-handling capability, decoding complexity, cost implications, scalability, and suitability for image-specific data. ResCNN and ResCNN+SDC demonstrate advantages in terms of low redundancy, error tolerance, and scalability compared to traditional approaches like DNA-Fountain coding [23], ECC-based designs [9], and strand overlapping with duplication and voting [10], [11]. These results illustrate the unique optimization provided by the proposed ResCNN and SDC techniques for DNA image storage workflows.

Features/Methods	Our Work	DNA-Fountain Code [23]	ECC Design [9]	Strand Overlapping [10]	Duplicate & Voting [11]
Redundancy Level	Low	Medium	High	High	Low-Medium
Error Handling	Tolerates major errors	Error correction	Error correction	Error mitigation	Error consensus
Dropout Handling					
Decoding Complexity	Low (CNN-based)	Moderate	Potentially complex	Moderate	Moderate
Cost Implications	Low-Moderate	Moderate-High	Moderate-High	High	High
Scalability	Very High	Moderate	Moderate	Low	Moderate
Image-Specific Optimization	Yes	No	No	No	No

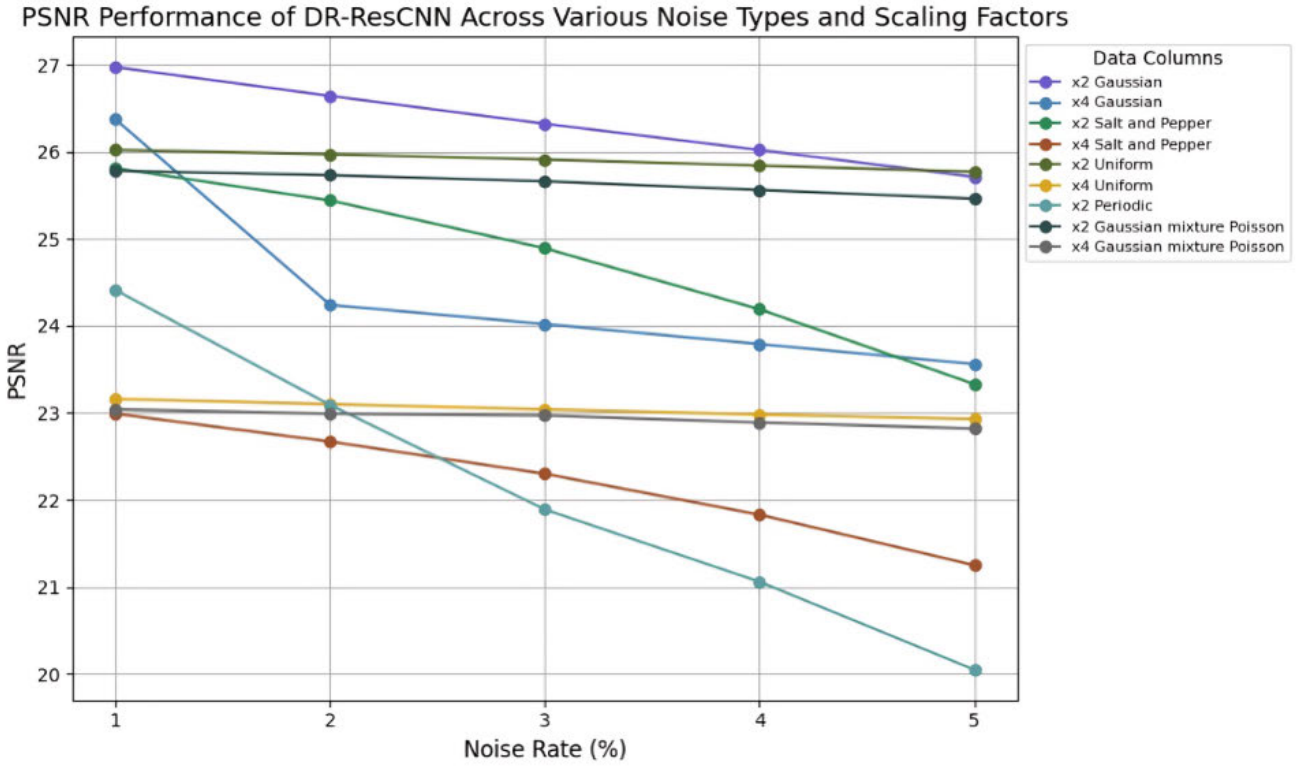


FIGURE 8. Average PSNR performance of ResCNN for various noise types and scaling factors (×2 and ×4) across different noise rates. Each data point represents the mean PSNR value calculated from all test images in our dataset. The graph illustrates the model's robustness in handling various noise conditions typical in DNA-based data storage. Gaussian noise demonstrates the best overall performance, while Salt and Pepper noise presents significant challenges at higher noise rates. The ×2 scaling consistently outperforms the ×4 scaling.

By integrating this broader noise model, our retrained ResCNN learns robust feature representations that generalize across diverse DNA storage error distributions while maintaining its super-resolution capability. This allows the model to effectively restore images from DNA storage with minimal visual degradation, even in the presence of severe IDS errors.

4) EXPERIMENTAL RESULTS AND VISUAL ANALYSIS

To validate the effectiveness of our proposed DSI-ResCNN in DNA storage environments, we conducted experimental evaluations to assess both its super-resolution capabilities and its robustness against DNA storage noise, using synthetic error models. DSI-ResCNN is designed not only to upscale low-resolution images but also to mitigate the impact of

various types of noise commonly encountered in DNA-based storage.

The results are visualized in Figure 7, which highlights the effects of different types of noise on image degradation and restoration quality. The evaluation covers two key aspects:

- Super-resolution performance at upscaling factors $\times 2$ and $\times 4$, demonstrating the model's ability to enhance image quality from compressed storage.
- Denoising robustness, assessing how well DSI-ResCNN restores images corrupted by synthetic DNA storage noise, including Gaussian, Salt-and-Pepper, Uniform, Periodic, and Poisson noise.

We observe that:

- Gaussian noise exhibits the best recoverability, with PSNR remaining above 25dB even at high noise levels.
- Salt-and-Pepper noise is the most challenging, as localized extreme errors lead to significant degradation.
- The $\times 2$ upscaling consistently outperforms $\times 4$, reflecting the impact of scale on restoration difficulty.

These observations are further quantified in Figure 8, which illustrates the degradation of PSNR as a function of the intensity of the noise. The key findings include the following.

- Gaussian noise shows the most graceful degradation, maintaining PSNR > 24 dB even at 5% noise.
- Salt-and-Pepper noise exhibits the sharpest decline, confirming its disruptive effect on structured data.
- Upscaling $\times 2$ provides a consistently higher PSNR than $\times 4$, validating its robustness under different error conditions.

These results demonstrate that DSI-ResCNN maintains stable restoration performance in multiple noise scenarios, making it a viable solution for DNA-based image storage recovery.

a: COMPARISON WITH CONVENTIONAL DNA STORAGE METHODS

To further assess the advantages of DSI-ResCNN, we compare its capabilities with conventional DNA storage error correction strategies, as summarized in Table 2. Unlike traditional approaches that rely on redundant DNA strands for error correction, DSI-ResCNN achieves high error tolerance with low redundancy, making it a more storage-efficient solution. Its CNN-based dropout handling reduces decoding complexity while effectively mitigating data loss, and its deep learning architecture ensures high scalability, allowing the model to generalize across various noise types encountered in DNA storage systems. This comparison highlights that DSI-ResCNN offers a unique trade-off, providing robust error recovery without the excessive redundancy required by conventional methods.

b: SUMMARY AND IMPLICATIONS FOR DNA IMAGE STORAGE

Our results demonstrate that DSI-ResCNN is a practical and efficient deep learning-based solution for error-tolerant DNA image storage. By incorporating super-resolution, our approach enables images to be stored at lower resolutions while achieving high-quality restoration, significantly reducing the cost of DNA synthesis and sequencing. Furthermore, DSI-ResCNN is resilient to various noise distortions, effectively reconstructing images despite synthesis, degradation, and sequencing errors. Unlike traditional DNA storage techniques that require high redundancy, our model provides strong recovery performance without additional redundant strands, ensuring both scalability and storage efficiency. These findings underscore the potential of deep learning to redefine error correction in DNA storage, paving the way for future optimizations that integrate both biological and computational constraints to further enhance real-world applications.

C. DISCUSSION

Our results demonstrate that DSI-ResCNN, the combination of ResCNN and SDC, effectively addresses the key challenges of DNA-based image storage by targeting two major types of errors: IDS (Type I) and sequence dropout (Type II).

SDC mitigates Type II errors by evenly distributing dropout losses across the image, preventing catastrophic data loss in concentrated regions. This transforms structured information loss into a more recoverable noise pattern, improving resilience to severe disruptions. ResCNN, on the other hand, corrects Type I errors by leveraging deep learning-based super-resolution and denoising, restoring fine image details even under significant degradation. Together, ResCNN + SDC creates a robust recovery pipeline that ensures both error-tolerant reconstruction and storage efficiency.

Unlike traditional DNA storage techniques that rely on high redundancy or complex symbol-level error correction, DSI-ResCNN introduces a scalable and low redundancy alternative. The network generalizes well across various noise types, handling both Gaussian-like distortions and extreme outlier errors (e.g., salt-and-pepper noise). This makes it highly adaptable to different DNA storage workflows, reducing the need for extensive physical redundancy while maintaining strong recovery quality.

Although our results validate the feasibility of deep learning-driven DNA storage, future optimizations include end-to-end training of ResCNN+SDC, better alignment with biological constraints, and exploring low-power implementations for practical deployment. However, DSI-ResCNN establishes a promising direction for hybrid DNA storage solutions, bridging biological encoding strategies with AI-based reconstruction, enabling more efficient and scalable DNA-based image preservation.

V. CONCLUSION AND FUTURE WORKS

DNA-based data storage holds immense potential for long-term, high-density information preservation, yet its practical deployment remains constrained by error-prone synthesis, degradation, and sequencing processes. Traditional redundancy-based error correction strategies significantly

increase storage overhead, limiting efficiency and scalability. In this work, we propose DSI-ResCNN, an error-tolerant deep learning-based information subsystem that shifts the focus from static redundancy mechanisms to an adaptive, learning-driven decoding approach. By integrating convolutional neural networks with a dropout-aware encoding strategy, our system effectively mitigates two major categories of DNA storage errors: local distortions induced by synthesis and sequencing noise and large-scale dropout due to sequence loss.

Our framework achieves this through two complementary components. ResCNN functions as both an image restoration and super-resolution model, enabling high-fidelity reconstruction from compressed DNA representations, thereby reducing synthesis and sequencing costs. This is complemented by SDC (Sequence Dropout Control), which strategically redistributes dropout-prone image information across sequence space, mitigating the effects of missing DNA fragments without introducing excessive redundancy. Together, these components establish a scalable and efficient DNA storage workflow, demonstrating that deep learning-based decoding can serve as an effective alternative to conventional error correction codes.

Looking ahead, the evolution of DNA storage must move towards end-to-end adaptive pipelines that seamlessly integrate biological and information subsystems. One promising direction is dynamic encoding strategies that adjust redundancy in real time based on sequencing conditions, rather than relying on rigid, one-size-fits-all error correction schemes. Feedback-driven adaptive error mitigation could optimize space efficiency by applying redundancy only where necessary, significantly reducing storage overhead while maintaining retrieval reliability.

Furthermore, bridging computational advances with biological constraints remains a critical challenge. Although our work focuses on computational robustness, incorporating real DNA synthesis and sequencing data into the training pipeline would refine the model performance and improve its practical applicability. Emerging technologies such as microfluidic DNA synthesis and high-throughput sequencing may enable real-time learning-based error handling, improving both synthesis efficiency and readout accuracy.

Beyond error correction, deep learning is poised to play a more expansive role in DNA storage systems. Future applications may leverage neural networks for sequence optimization, real-time quality control, and adaptive error detection during synthesis and sequencing. The ability of deep learning models to dynamically learn from sequencing trends could further reduce error rates, improve storage density, and improve system reliability.

While our in vitro approach with short DNA sequences demonstrates advantages in stability, future work must address data security concerns for practical DNA storage deployment. Implementation of molecular-level security mechanisms represents a critical research direction. Possibilities include embedding cryptographic signatures within DNA sequences through selective nucleotide modifications that preserve information content while enabling authentication. Access control could be implemented through specialized sequencing conditions or primer-dependent readouts, effectively creating biological encryption keys. These approaches, when integrated with our error-tolerance framework, would address the dual challenges of data integrity and security that currently limit DNA storage adoption in sensitive applications.

The robustness and efficiency of DSI-ResCNN position it as a promising candidate for applications where long-term, error-resilient data preservation is critical. In digital forensics, DNA-based storage could revolutionize the archiving of immutable evidence. In biomedical imaging, secure ultra-long-term storage of patient data could be realized, ensuring fidelity over decades without degradation. These domains will benefit from the dual capabilities of DSI-ResCNN for image restoration and noise mitigation, offering a compelling alternative to conventional storage mediums.

The principles underlying DSI-ResCNN also extend beyond image storage. As DNA storage technologies mature, the integration of intelligent biological and computational codesigns will be key to unlocking the full potential of molecular information storage. Although cost and speed remain major hurdles, continued advancements in both synthetic biology and machine learning may progressively establish DNA storage as a viable complement to conventional digital storage solutions. By pioneering a deep learning-based decoding paradigm, our work lays the groundwork for the next generation of intelligent, error-tolerant, and secure molecular data storage architectures.

ACKNOWLEDGMENT

The authors are grateful for the help of Dr. Chenchen Zhu and Dr. Bingqing Zhao from Stanford University School of Medicine. Special thanks should be given to Dr. Meng Wang from MM Computational Medicine and Bioinformatics, University of Michigan. They express their sincere gratitude to the anonymous reviewers whose insightful comments and suggestions substantially improved the quality and clarity of this manuscript. Their feedback helped them better articulate their technical contributions and strengthen the experimental validation. They utilized large-language models from OpenAI and Anthropic to assist in refining the language and clarity of the manuscript.

They would also like to express gratitude to the memory of Dr. Jishou Ruan, whose guidance and encouragement were instrumental in shaping the research direction and inspiring this journey.

REFERENCES

[1] D. Reinsel, J. Gantz, and J. Rydning, "The digitization of the world-from edge to core," Int. Data Corp. (IDC), White Paper US44413318, Nov. 2018. [Online]. Available: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

[2] A. Extance, "How DNA could store all the world's data," *Nature*, vol. 537, no. 7618, pp. 1–12, 2016.
[3] N. Tang, S. Ma, and J. Tian, "New tools for cost-effective DNA synthesis," in *Synthetic Biology*. Amsterdam, The Netherlands: Elsevier, 2013, pp. 3–21.
[4] J.-Y. Lee, "The principles and applications of high-throughput sequencing technologies," *Develop. Reproduction*, vol. 27, no. 1, pp. 9–24, Mar. 2023.
[5] S. K. Tabatabaei, B. Wang, N. B. M. Athreya, B. Enghiad, A. G. Hernandez, C. J. Fields, J.-P. Leburton, D. Soloveichik, H. Zhao, and O. Milenkovic, "DNA punch cards for storing data on native DNA sequences via enzymatic nicking," *Nature Commun.*, vol. 11, no. 1, p. 1742, Apr. 2020.
[6] L. Organick et al., "Random access in large-scale DNA data storage," *Nature Biotechnol.*, vol. 36, no. 3, pp. 242–248, Feb. 2018.
[7] D. Panda, K. A. Molla, M. J. Baig, A. Swain, D. Behera, and M. Dash, "DNA as a digital information storage device: Hope or hype?" *3 Biotech*, vol. 8, no. 5, pp. 1–9, May 2018.
[8] J. Bornholt, R. Lopez, D. M. Carmean, L. Ceze, G. Seelig, and K. Strauss, "A DNA-based archival storage system," *ACM SIGPLAN Notices*, vol. 51, no. 4, pp. 637–649, Jun. 2016.
[9] W. H. Press, J. A. Hawkins, S. K. Jones, J. M. Schaub, and I. J. Finkelstein, "HEDGES error-correcting code for DNA storage corrects indels and allows sequence constraints," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 31, pp. 18489–18496, Aug. 2020.
[10] Z. Ping, D.-Z. Ma, X. Huang, S. Chen, L. Liu, F. Guo, S. Zhu, and Y. Shen, "Carbon-based archiving: Current progress and future prospects of DNA-based data storage," *GigaScience*, vol. 8, no. 6, p. 075, Jun. 2019.
[11] C. Xu, B. Ma, Z. Gao, X. Dong, C. Zhao, and H. Liu, "Electrochemical DNA synthesis and sequencing on a single electrode with scalability for integrated data storage," *Sci. Adv.*, vol. 7, no. 46, Nov. 2021, Art. no. eabk0100.
[12] C. Pan, S. M. Hossein Tabatabaei Yazdi, S. Kasra Tabatabaei, A. G. Hernandez, C. Schroeder, and O. Milenkovic, "Image processing in DNA," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 8831–8835.
[13] C. Ruan, L. Yang, R. Han, S. Gao, H. Wu, and N. Ling, "Robust DNA image storage decoding with residual CNN," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2024, pp. 1–5.
[14] K. Reinert, B. Langmead, D. Weese, and D. J. Evers, "Alignment of next-generation sequencing reads," *Annu. Rev. Genomics Hum. Genet.*, vol. 16, no. 1, pp. 133–151, May 2015.
[15] P. McInerney, P. Adams, and M. Z. Hadi, "Error rate comparison during polymerase chain reaction by DNA polymerase," *Mol. Biol. Int.*, vol. 2014, pp. 1–8, Aug. 2014.
[16] B. H. Nguyen, C. N. Takahashi, G. Gupta, J. A. Smith, R. Rouse, P. Berndt, S. Yekhanin, D. P. Ward, S. D. Ang, P. Garvan, H.-Y. Parker, R. Carlson, D. Carmean, L. Ceze, and K. Strauss, "Scaling DNA data storage with nanoscale electrode wells," *Sci. Adv.*, vol. 7, no. 48, p. 6714, Nov. 2021.
[17] T.-H. Khuat and S. Kim, "A quaternary code correcting a burst of at most two deletion or insertion errors in DNA storage," *Entropy*, vol. 23, no. 12, p. 1592, Nov. 2021.
[18] Y. Pi and Z. Zhang, "Two-insertion/deletion/substitution correcting codes," 2024, *arXiv:2401.11231*.
[19] D. Brinza and F. Hyland, "Workshop: Error correction methods in next generation sequencing," in *Proc. IEEE 1st Int. Conf. Comput. Adv. Bio Med. Sci. (ICCABS)*, Feb. 2011, p. 268.
[20] I. F. Bronner, M. A. Quail, D. J. Turner, and H. Swerdlow, "Improved protocols for illumina sequencing," *Current Protocols Hum. Genet.*, vol. 79, no. 1, pp. 1–46, Oct. 2013.
[21] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.
[22] C. Ruan, R. Han, Y. Li, S. Gao, H. Wu, and N. Ling, "Efficient DNA-based image coding and storage," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2023, pp. 1–5.
[23] Y. Erlich and D. Zielinski, "DNA fountain enables a robust and efficient storage architecture," *Science*, vol. 355, no. 6328, pp. 950–954, Mar. 2017.

CIHAN RUAN (Graduate Student Member, IEEE) received the B.S. degree in software engineering from Nankai University, Tianjin, China, in 2014, and the M.S. degree in computer science and engineering from Santa Clara University, Santa Clara, CA, USA, in 2016, where she is currently pursuing the Ph.D. degree in computer science and engineering, working with Dr. Nam Ling on image and video processing, compression, and DNA-based data storage.

Her research focuses on neural network-based image, video, and haptic data compression. She also works on deep learning for DNA-based data storage, investigating error-tolerant decoding, and adaptive encoding strategies. Her research interests include advancing generative model-driven multimodal data compression, enhancing DNA-based archival systems, and exploring AI-driven approaches for secure, long-term digital preservation.

LIANG YANG received the B.S. and M.S. degrees in software engineering from Nankai University, Tianjin, China, in 2014 and 2017, respectively, where he is currently pursuing the Ph.D. degree with the College of Software, under the supervision of Dr. Tiegang Gao. His current research interests include information security, reversible data hiding, and multimedia security.

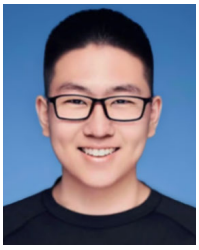
RONGDUO HAN (Graduate Student Member, IEEE) received the B.S. degree in mathematics and the M.S. degree in bioinformatics from the School of Mathematical Sciences, Nankai University, Tianjin, China, in 2021 and 2024, respectively. He is currently pursuing the Ph.D. degree in software engineering with the College of Software, Nankai University, under the supervision of Prof. Haining Zhang.

In 2023, he was a Visiting Scholar with Santa Clara University, Santa Clara, CA, USA. His research interests include deep learning, computer vision, and biomedical imaging, with a focus on AI-driven medical diagnostics, 3D medical image reconstruction, and multimodal haptic-video processing. He is also involved in DNA data storage research and video processing techniques for film production. His interdisciplinary work integrates computational intelligence with healthcare, digital archiving, and creative media applications.

SHAN GAO received the B.S. degree in electronic engineering from the National University of Defense Technology, in 2000, the M.S. degree in biochemistry and molecular biology from Shanghai University, in 2007, and the Ph.D. degree in bioinformatics from Nankai University, Tianjin, China, in 2010. He then completed two postdoctoral fellowships: one at the University of Kansas, from 2010 to 2011, and another at Cornell University, from 2011 to 2014. Since 2014, he has been an Associate Professor with Nankai University. He is currently an Associate Professor with the College of Life Sciences, Nankai University. His research interests include high-throughput omics data analysis, long noncoding RNAs, single-cell transcriptome sequencing data analysis, and bioinformatics applications in cancer, viruses, and other related fields.

HAOYU WU received the dual bachelor's degrees in computer science and public relations from Boston University and the master's degree in software engineering from Carnegie Mellon University, in 2021. He is currently a Software Engineer with Roku, Inc., where he has been contributing, since January 2022, with a focus on developing cutting-edge software solutions. His expertise includes large-scale software engineering projects and a deep interest in multimedia systems.

QIMING YUAN (Student Member, IEEE) received the B.S. degree in computer science from the University of Nebraska–Lincoln, Lincoln, NE, USA, in 2019, and the M.S. degree in computer science and engineering from Santa Clara University, Santa Clara, CA, USA, in 2022, where she is currently pursuing the Ph.D. degree in computer science and engineering. She is working with her advisor, Dr. Nam Ling, on deep learning-based 3D image and point cloud generation and compression. Her research interests include generative models for 3D vision, neural representations for point cloud compression, and efficient coding techniques for immersive media applications.



YANTING GUO is currently pursuing the Bachelor of Science degree in mathematics and statistics with Yunnan University, China. He is actively involved in bioinformatics research under the supervision of Associate Prof. Shan Gao from Nankai University. His research interest includes the analysis of high-dimensional data.

NAM LING (Life Fellow, IEEE) received the B.Eng. degree from the National University of Singapore, Singapore, in 1981, and the M.S. and Ph.D. degrees from the University of Louisiana at Lafayette, Lafayette, LA, USA, in 1985 and 1989, respectively. He was the Department Chair, from 2010 to 2023, an Associate Dean for Graduate Studies/Research/Faculty Development, from 2002 to 2010, and the Sanfilippo Family Chair Professor, from 2010 to 2020. He is/was also a Chair/Distinguished/Guest/Consulting Professor with several universities internationally. He is currently the Wilmot J. Nicholson Family Chair Professor with the Department of Computer Science and Engineering and the Associate Dean for Research with the School of Engineering, Santa Clara University, USA. He has authored or co-authored over 350 publications and seven adopted standard contributions. He has been granted more than 20 U.S./European/PCT patents. He has delivered more than 120 invited colloquia worldwide. He is an IEEE Fellow due to his contributions to video coding algorithms and architectures. He is also an IET Fellow and an AAIA Fellow. He was a recipient of the IEEE ICCE Best Paper Award (First Place) and the Umedia Best/Excellent Paper Award (thrice). He received six awards from Santa Clara University, four at the university level and two at the school/college level. He has served as the General Chair/Co-Chair for IEEE Hot Chips, VCVP (twice), IEEE ICME, IEEE VCIP, IEEE SiPS, SocialSec, and Umedia (six times). He has also served as the Technical Program (TPC) Co-Chair for IEEE ISCAS (twice), APSIPA ASC, IEEE APCCAS, IEEE SiPS (twice), DCV, and IEEE VCIP. He is the TPC Chair of IEEE ICME 2026. He was the Technical Committee Chair of IEEE CASCOM TC and IEEE TCMM and the Chair of the APSIPA U.S. Chapter. He served as the Guest Editor or an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS, IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, the *Journal of Signal Processing Systems* (Springer), *Multidimensional Systems and Signal Processing* (Springer), and other journals. He was a Keynote Speaker for IEEE APCCAS, VCVP (twice), JCPC, IEEE ICAST, IEEE ICIEA, IET FC & Umedia, IEEE Umedia, IEEE ICCIT, ICNLP/SSPS/CVPS, and Workshop at XUPT (twice). He was named as an IEEE Distinguished Lecturer twice. He was also an APSIPA Distinguished Lecturer.