Neural Networks 192 (2025) 107863

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Full Length Article

Decoding natural visual scenes via learnable representations of neural spiking sequences

Jing Peng [a,b,1], Shanshan Jia [c,1], Jiyuan Zhang [c], Yongxing Wang [a], Zhaofei Yu [c,*], Jian K. Liu [a,b,*]

[a] School of Computer Science, University of Leeds, Leeds, UK
[b] School of Computer Science, Centre for Human Brain Health, University of Birmingham, Birmingham, UK
[c] Institute for Artificial Intelligence, School of Computer Science, Peking University, Beijing, China

ARTICLE INFO

Keywords:
Vision
Video
Neural spike
Wavelet
Neural network
Deep learning

ABSTRACT

Visual input underpins cognitive function by providing the brain with essential environmental information. Neural decoding of visual scenes seeks to reconstruct pixel-level images from neural activity, a vital capability for vision restoration via brain-computer interfaces. However, extracting visual content from time-resolved spiking activity remains a significant challenge. Here, we introduce the Wavelet-Informed Spike Augmentation (WISA) model, which applies multilevel wavelet transforms to spike trains to learn compact representations that can be directly fed into deep reconstruction networks. When tested on recorded retinal spike data responding to natural video stimuli, WISA substantially improves reconstruction accuracy, especially in recovering fine-grained details. These results emphasize the value of temporal spike patterns for high-fidelity visual decoding and demonstrate WISA as a promising model for visual decoding.

1. Introduction

In daily life, the brain continuously receives and processes a vast array of sensory information from the external environment. This sensory input is transmitted through a complex neural system, ultimately driving behavioral responses. Neurons play a central role in this process, serving as critical elements for information transmission and efficient computation. Individual neurons respond to input stimuli by altering their membrane potential, generating discrete electrical events known as neural spikes (Rieke, 1997). These spikes are widely recognized as the fundamental units of neural computation, encoding and representing sensory stimuli, including visual information (Chichilnisky, 2001; Gollisch and Meister, 2008), where the earlier visual system carries out a significant part of visual information (Chen et al., 2024; Gollisch and Meister, 2010; Karamanlis et al., 2022).

Neural coding seeks to understand how neurons encode the relationship between sensory stimuli and neural spikes (Rieke, 1997). This field encompasses two primary components: neural encoding and neural decoding. Neural encoding investigates how individual neurons or neural populations process environmental stimuli (Liu and Gollisch, 2015; Liu et al., 2017; Olshausen et al., 1996; Onken et al., 2016; Simoncelli and Olshausen, 2001; Wang et al., 2020), while neural decoding focuses on extracting and interpreting information embedded in neural signals (Quiroga et al., 2009; Wu et al., 2006). In the context of the visual system, this involves exploring how visual scenes are represented by spiking signals and developing methodologies to decode these signals to reconstruct the input stimuli (Shah and Chichilnisky, 2020; Yu et al., 2020; Zhang et al., 2022).

To uncover the principles of neural encoding, researchers have developed a variety of models based on the intrinsic properties of neurons and neural circuits (Meyer et al., 2017; Pillow et al., 2008; Wu et al., 2025; Yan et al., 2022; Yu et al., 2020). Early spike-feature transformation methods exploited the wavelet transform to perform unsupervised detection and sorting of action potentials, demonstrating superior discrimination of spike classes without manual intervention (Quiroga et al., 2004). Subsequent work has addressed the computational bottlenecks of large-scale recordings by developing scalable, automated spike-sorting pipelines that maintain high reliability across thousands of channels (Carlson and Carin, 2019). These studies utilize the dual importance of time-frequency feature representation and algorithmic efficiency in spike-based decoding. Building on these insights, our current work focuses on integrating learnable multiscale wavelet representations with temporal convolutional feature learning to directly enhance downstream image reconstruction. During the neural transmission of

* Corresponding authors.
E-mail addresses: yuzf12@pku.edu.cn (Z. Yu), j.liu.22@bham.ac.uk (J.K. Liu).
[1] Equal contribution.

https://doi.org/10.1016/j.neunet.2025.107863
Received 14 January 2025; Received in revised form 1 July 2025; Accepted 10 July 2025
Available online 16 July 2025

visual information, neural spikes are essential units to represent visual stimuli. Consequently, constructing models capable of directly reconstructing visual scenes from spiking signals is critical to advance our understanding of neural representation (Botella-Soler et al., 2018; Gollisch and Meister, 2008). An effective neural decoder should reliably extract and reconstruct stimulus information encoded in neural spikes (Yu et al., 2020).

Traditional neural decoding approaches often involve classifying neural signal patterns to infer the corresponding stimulus type (Ragni et al., 2021; Shen et al., 2021; Wen et al., 2018). More recently, stimulus reconstruction techniques have emerged that allow the recovery of pixel-level details of visual scenes from neural signals (Nishimoto et al., 2011; Zhang et al., 2020, 2022), including functional magnetic resonance imaging (fMRI) activity (Naselaris et al., 2009; Nishimoto et al., 2011; Qiao et al., 2018; Thirion et al., 2006; Wen et al., 2018), spiking neural signals (Botella-Soler et al., 2018; Gollisch and Meister, 2008; Marre et al., 2015; Parthasarathy et al., 2017), and calcium imaging signals (Garasto et al., 2018; Yoshida and Ohki, 2020). Despite these advancements, current methodologies face limitations in decoding high-resolution dynamic natural scenes (Shah and Chichilnisky, 2020). Addressing these challenges is critical for realizing the full potential of neural decoding in visual neuroscience, brain-machine interface, and visual neuroprosthesis technologies (Yang et al., 2023; Yu et al., 2020).

In this study, we introduce the Wavelet-Informed Spike Augmentation (WISA) model, based on previous deep learning decoding models (Yu et al., 2024; Zhang et al., 2020, 2022). WISA first applies a discrete wavelet transform (DWT) to spike trains to obtain multilevel coefficients that capture essential time-frequency structure (Strang and Nguyen, 1996). A compact temporal convolutional network then adaptively refines these coefficients, and an inverse DWT reconstructs augmented spiking neural signals for downstream decoding. By integrating learnable, multiscale preprocessing directly into the decoding pipeline, WISA enables end-to-end optimization of time-frequency features, substantially improving reconstruction fidelity of video stimuli. To evaluate the effectiveness of WISA, the augmented spikes were integrated into a downstream decoding network to align the spikes with corresponding visual stimuli. The entire framework, comprising WISA and the decoding network, operates as an end-to-end deep learning system, enabling the reconstruction of dynamic natural visual scenes directly from spiking signals. Using experimentally recorded neural spiking signals in response to video stimuli, we demonstrated that WISA significantly improves the quality of visual scene reconstruction. These results highlight the robustness and practical potential of the WISA approach for learning representations of spiking signals and providing an advanced modeling framework for neural decoding of dynamic visual scenes using neural spike sequences.

2. Results

2.1. Decoding visual scenes with WISA

The proposed decoding framework comprises two stages (Fig. 1). In the first stage, the WISA module is utilized to learn representations of temporal sequences of spiking signals (see Methods). The second stage involves a decoding network that converts the augmented spiking signals into visual images. To highlight the benefits of WISA, we use previously developed CNN models as decoders (Yu et al., 2024; Zhang et al., 2020, 2022). Although alternative decoding networks can be incorporated, we use the CNN decoder throughout the study as a baseline model for comparison with WISA.

The proposed decoding framework was evaluated using experimental data recorded from retinal ganglion cells to reconstruct two natural video stimuli: salamander and trigger videos, which show varying levels of spatial and temporal complexity (Zheng et al., 2021). The results demonstrated that WISA significantly enhances the decoding performance (Fig. 2). Both the CNN decoder without WISA and the

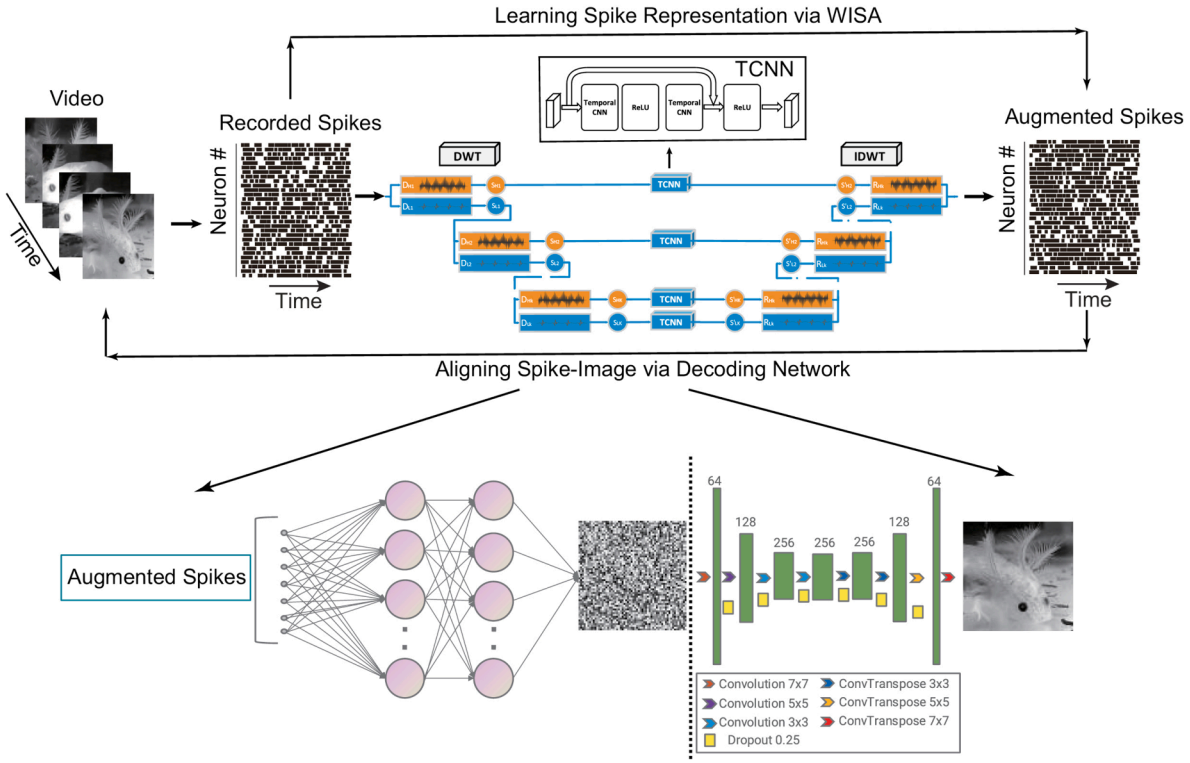

Fig. 1. The decoding framework with WISA. (Top) The WISA module has three sequential steps: 1) DWT is applied to spiking signals to obtain multilevel wavelet coefficients; 2) Temporal convolutional neural network (TCNN) module is employed to learn and extract high- and low-frequency features from the wavelet coefficients; 3) The wavelet coefficients of the extracted features are reconstructed through inverse DWT, generating the augmented spikes. (Bottom) The CNN decoding network uses augmented spikes to obtain reconstructed visual images.

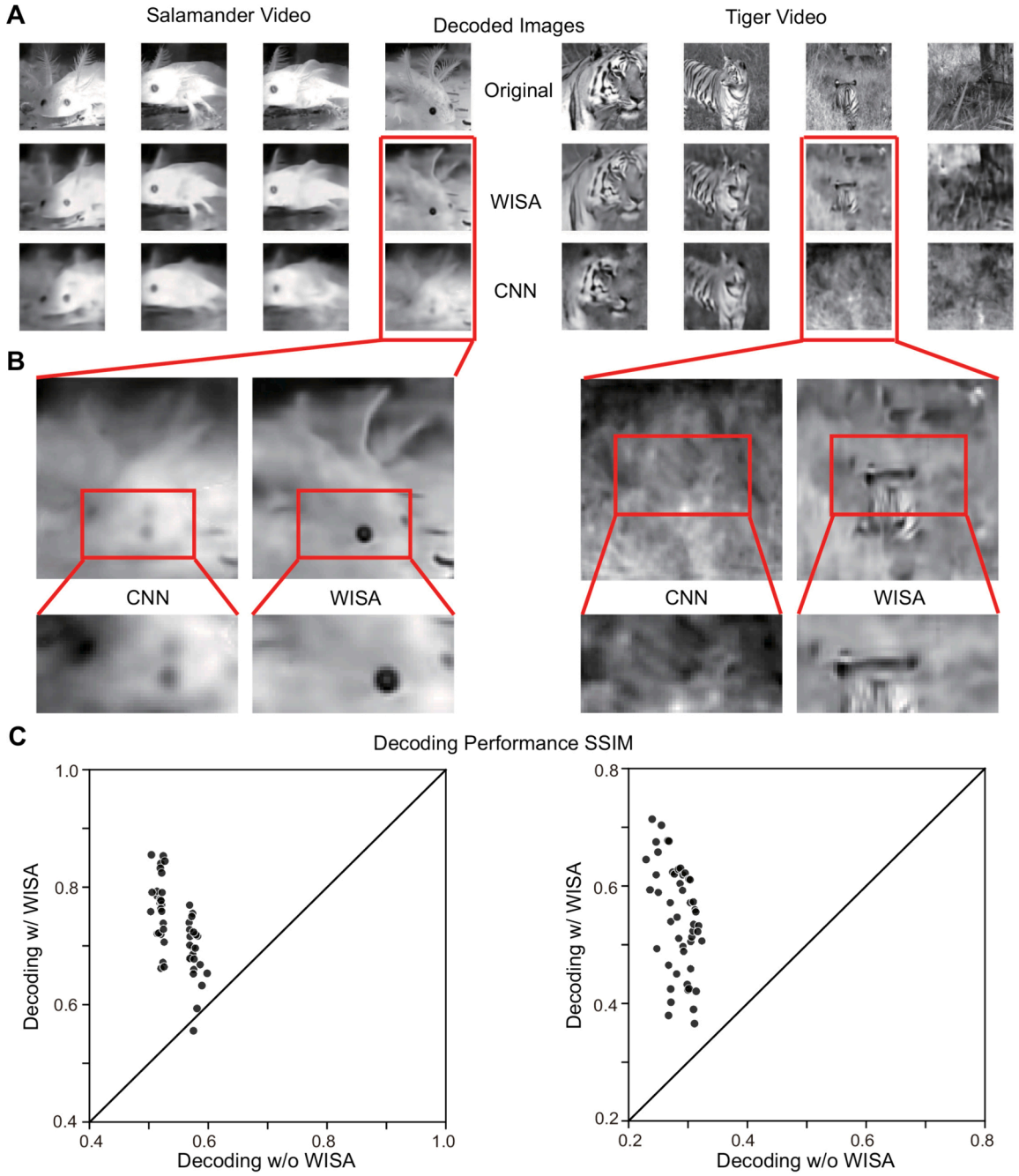

Fig. 2. WISA improves the decoding results. (A) Decoded images from the Salamander (left) and Tiger (right) video by the decoder network with WISA and without WISA. (B) Highlighted image details were better retrieved by WISA. (C) The decoding performance quantified by SSIM with and without WISA. Each data point is the SSIM of a test image.

WISA-enhanced model successfully reconstructed the global content of each video frame (Fig. 2(A)). However, the WISA-enhanced model exhibited superior ability to capture fine-grained details within the images (Fig. 2(B)). The decoding performance was quantitatively evaluated using the structural similarity index measure (SSIM) between the original and decoded images, revealing a marked improvement with WISA compared to the baseline CNN decoder (Fig. 2(C)).

2.2. Correlation analysis of decoded images

To further assess the reconstruction performance of the WISA model across different datasets, we randomly sampled 10 % of all frames from both videos, 180 frames from the salamander video and 160 frames from the tiger video, and calculated the Pearson correlation coefficients (CC) for each pixel in the reconstructed images (Fig. 3). The WISA model consistently exhibited higher accuracy, with average CCs concentrated above 0.9 for the salamander video and above 0.8 for the tiger video. In contrast, the CNN decoder without WISA displayed more dispersed and generally lower CCs. The significantly higher CCs achieved by WISA indicate a stronger relationship between the reconstructed and original images, showing its superior performance (Fig. 3(A)).

To further quantify pixel-level reconstruction fidelity, we computed the Pearson correlation coefficient between flattened original and reconstructed images for both the baseline CNN decoder (without WISA) and the WISA-enhanced model. The WISA-enhanced scatter plot aligns tightly with the identity line, yielding a Pearson correlation of 0.97 and

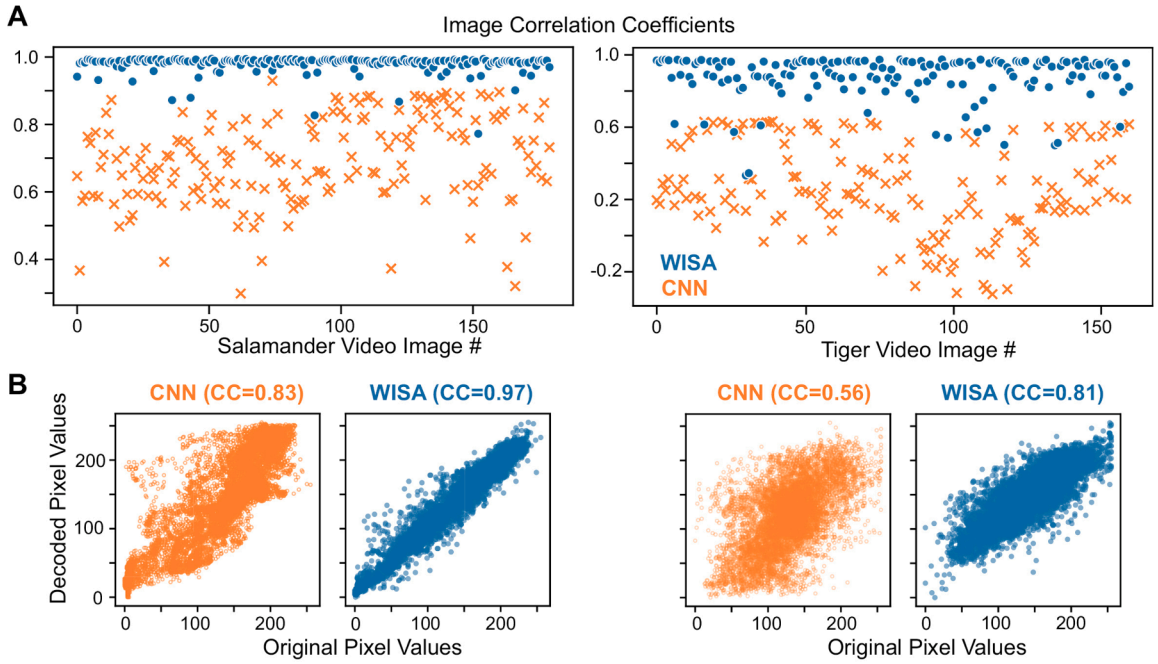

Fig. 3. The comparison of individual pixel correlation for the decoded images. WISA shows higher correlations than CNN. (A) Pearson correlation coefficients averaged over each image by WISA and CNN. Each point indicates the correlation coefficient averaged over all the pixels of each image. (B) The scatter plots of individual pixel values of all the original and decoded images by both models. Each point represents a pixel value from the original image and its corresponding reconstructed image.

indicating accurate recovery of most pixel intensities (Fig. 3(B)). In contrast, the baseline CNN exhibits a more dispersed scatter and a lower correlation coefficient of 0.83, reflecting larger reconstruction errors. These results demonstrate that WISA not only achieves superior overall correlation but also substantially reduces pixel-wise errors, thereby more effectively preserving fine-grained image details.

These results confirm that the WISA model is capable of preserving the details and pixel values of the original images, resulting in higher reconstruction accuracy and consistency. Conversely, the plain CNN model without WISA exhibited larger errors during reconstruction and performed worse overall, particularly in complex regions with high pixel values.

2.3. Low-dimensional embedding of original and reconstructed images

To address the challenge of visually distinguishing reconstructed images from the originals, we employed Principal Component Analysis (PCA) to compare their low-dimensional embedded representations. Specifically, 10 % of all frames from both videos were randomly sampled and divided into 32×32 patches. The original images and their corresponding reconstructed images were pooled and analyzed using PCA for dimensionality reduction and visualization (Fig. 4).

The PCA results reveal that the low-dimensional embeddings of images reconstructed by the WISA model closely align with those of the original images. This alignment suggests that the WISA-reconstructed images successfully preserve the fine-grained visual patterns and structural characteristics of their corresponding original frames. These findings further validate the effectiveness of the WISA model in accurately reconstructing visual scenes.

2.4. Comparative evaluation of WISA against state-of-the-art spike decoding models

To further validate the superiority of the WISA framework, we conducted quantitative comparison experiments on two video datasets against recently proposed spike-based decoding models, Spk2ImgNet (Zhao et al., 2021) and S2INet (Li et al., 2023). Spk2ImgNet applies parallel, multi-branch learnable filters over spike time windows of varying lengths to estimate instantaneous luminance, extracts features via shared-weight multi-layer residual blocks, and fuses these features across time using reliability-weighted deformable convolutions, producing the reconstructed image end-to-end. S2INet employs an end-to-end autoencoder architecture whose encoder comprises two fully connected layers with Gabor-oriented convolutional blocks (to mimic V1 spatial-frequency and orientation selectivity), and whose decoder uses symmetric deconvolutional layers to restore the image. The experiments with Spk2ImgNet and S2INet were carried out by optimizing the model parameters to suit our current dataset.

To further quantify the model performance, we used another similarity metric, peak signal-to-noise ratio (PSNR), calculated between the stimulus and the decoded images. WISA substantially outperforms all competing methods on both datasets in terms of PSNR and SSIM (Table 1). Compared to the baseline CNN, WISA achieves better PSNR and SSIM gains. Spk2ImgNet and S2INet achieved lower reconstruction fidelity under our experimental conditions. To evaluate the robustness of these improvements, Welch's t-tests was conducted on the SSIM and PSNR values comparing WISA with Spk2ImgNet, S2INet and CNN, where both metrics exhibit highly significant improvement in WISA ($p < 0.001$). These results demonstrate that WISA utilizes learnable wavelet representation to take into account temporal information in spiking sequences, improving the ability of spike-to-scene reconstruction.

2.5. Optimizing wavelet transformation on model performance

To examine the robustness of wavelet hyperparameters in WISA, we conducted systematic ablation experiments along two dimensions: wavelet basis function and decomposition level. First, holding the Daubechies wavelet basis (dB) constant, we varied the decomposition level from 1 to 4 and compared each to our chosen level of 5

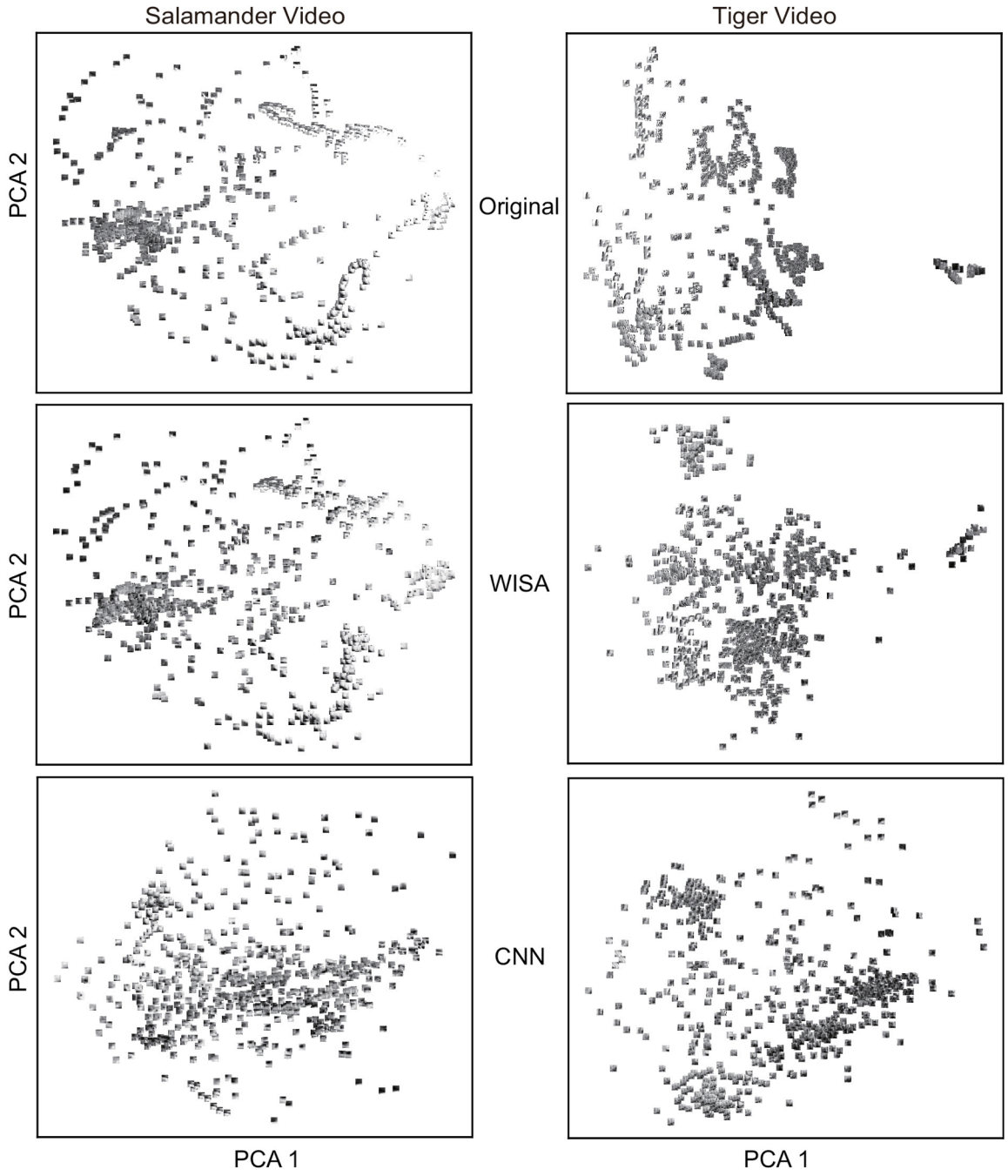

Fig. 4. Low-dimensional embedding distribution of decoding results preserved by WISA. WISA better preserves the embedding distribution of the original images, compared to CNN. Each point indicates a 32x32 patch image.

Table 1
A quantitative comparison of PSNR and SSIM for each evaluated model on the salamander and tiger datasets. The values (Mean ± Std) are obtained from different model training settings.

	Salamander		Tiger	
Model	PSNR	SSIM	PSNR	SSIM
Spk2ImgNet	12.1762±0.2098	0.3289±0.0142	11.1660±0.1262	0.1732±0.0072
S2INet	14.5524±0.1731	0.2352±0.0171	12.7392±0.0971	0.1650±0.0036
CNN	19.0123±0.7959	0.5319±0.0186	15.0715±0.9295	0.2324±0.0139
WISA	**23.9370±0.0217**	**0.7290±0.0651**	**20.7847±0.1458**	**0.5459±0.0378**

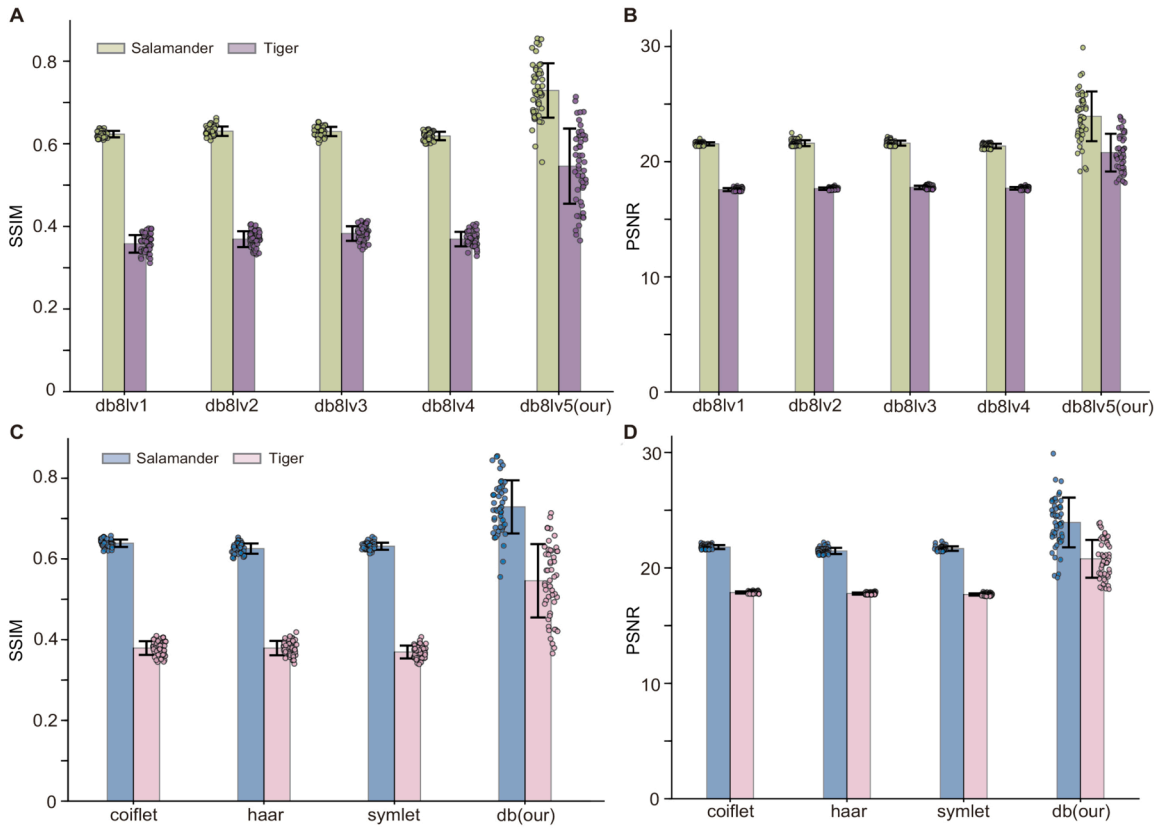

Fig. 5. Effect of individual wavelet components on decoding performance. SSIM (A, C) and PSNR (B, D) metrics are compared across different decomposition levels (A, B) and wavelet bases (C, D) on two datasets.

(Fig. 5(a–b)). The results indicate that lower decomposition levels (Level ≤ 4) or shorter filters (e.g., dB8 at low levels) can extract only partial time-frequency features, leading to substantially reduced PSNR and SSIM. In contrast, at Level 5, the model both fully captures multi-scale time-frequency information and avoids performance saturation, since it does not exceed the signal's maximum permissible decomposition level $\lfloor \log_2 N \rfloor$ (where N is the signal length)-thereby achieving peak PSNR and SSIM on both Salamander and Tiger datasets. Second, with the decomposition level fixed at 5, we compared three alternative wavelet bases-Coiflet, Haar, and Symlet-against our final choice of Daubechies-8 (Fig. 5(c–d)). Coiflet, Haar, and Symlet all underperform relative to dB8-Level 5 in terms of PSNR and SSIM, owing to their shorter filter lengths or differing design priorities that prevent balanced feature extraction across time and frequency domains. These findings demonstrate that the dB8 wavelet combined with a five-level decomposition optimally balances computational efficiency and maximal multi-scale time-frequency feature extraction, thereby enhancing image reconstruction quality and validating the robustness of this hyperparameter configuration.

2.6. Contribution of different model components in WISA

As shown in Table 2, we investigated the effect of different model components on the decoding performance. By removing various components, the performance drops compared to the full WISA model. Compared to the baseline CNN, TCNN increases performance, demonstrating that the TCNN module alone can effectively enhance the learning and integration of time-frequency features, thus strengthening image reconstruction capability. Omitting the IDWT step leads to a pronounced drop in reconstruction quality compared to the complete WISA model, showing the necessity of inverse wavelet reconstruction for preserving multi-scale temporal dynamics. To evaluate the significance of these improvements, Welch's t-tests were applied to the SSIM and PSNR values of the full WISA in contrast to three ablated models, showing both metrics exhibit highly significant increases ($p < 0.001$). Therefore, the full WISA model achieves the highest SSIM and PSNR among all configurations, confirming that the synergy of multilevel wavelet decomposition, TCNN enhancement, and inverse transform optimally boosts decoding performance.

Table 2
Ablation analysis of the WISA model components for the decoding performance. The full WISA is represented as a combination DWT + TCNN + IDWT + CNN. (−) Component removed; (+) Component added. The values (Mean±Std) are obtained from different model training settings.

	Salamander		Tiger	
WISA model components	PSNR	SSIM	PSNR	SSIM
−DWT −TCNN −IDWT +CNN	19.0123±0.7959	0.5319±0.0186	15.0715±0.9295	0.2324±0.0139
−DWT +TCNN −IDWT +CNN	20.6344±0.5667	0.5959±0.0141	16.6029±0.7111	0.3334±0.0161
+DWT +TCNN −IDWT +CNN	18.7159±0.4980	0.5087±0.0134	13.9396±0.7515	0.1919±0.0612
+DWT +TCNN +IDWT +CNN	**23.9370±0.0217**	**0.7290±0.0651**	**20.7847±0.1458**	**0.5459±0.0378**

3. Conclusions and discussions

In this study, we proposed a novel WISA module that operates on spike sequences by performing wavelet transformation, learning to enhance its wavelet coefficients, and outputting new representations. The augmented representation of spikes retains the temporally ordered effective information embedded in spiking signals. Compared to other deep learning models without wavelet, the present study shows the advantage of exploring spikes in the time-frequency domain. We demonstrated the effectiveness of WISA in reconstructing visual stimulus images, enabling future studies on the utilization of time-resolved neural activity for visual decoding.

3.1. Visual decoding

The goal of neural decoding in vision is to predict visual stimuli based on neural activity. Traditionally, linear methods have been employed as neural decoders due to their simplicity and computational efficiency (Schoenmakers et al., 2013). However, the limited representational capacity of linear approaches results in suboptimal reconstruction performance, particularly for complex visual tasks such as decoding natural scenes. While linear methods provide high interpretability, their inability to capture intricate features of visual stimuli restricts their applicability in high-fidelity reconstruction.

In recent years, deep neural networks (DNNs) have become increasingly prevalent in neural decoding tasks, offering enhanced representational capabilities. CNNs, in particular, are widely adopted in various DNN architectures for visual decoding. CNNs progressively extract hierarchical visual features from input images through convolutional layers, utilizing kernels of varying sizes that mimic receptive fields in neuroscience (Ciresan et al., 2011; Goodfellow et al., 2016; LeCun et al., 2002; Richards et al., 2019).

Advances in CNN-based algorithms have demonstrated exceptional performance in image decoding using neural signals (Güçlü et al., 2015; Iqbal et al., 2019; Kim et al., 2021; Li et al., 2022; Zhang et al., 2020, 2022). Compared to linear methods, DNN-based approaches significantly enhance the accuracy of reconstructing natural scenes. Despite their dependency on large-scale neural data, deep learning technologies remain among the most promising methodologies for advancing visual neural decoding.

3.2. Enhancing spiking signals

Action potentials, or spikes, represent a specialized form of electrical signaling in neural cells, serving as a fundamental mechanism for neuronal information transmission and communication. When the membrane potential of a neuron surpasses a threshold during depolarization, an action potential is generated (Kleinfeld et al., 2019). Spikes are typically recorded using microelectrodes, an invasive method where electrodes are inserted into or near neurons to capture neural activity (Orsborn et al., 2014; Shanechi et al., 2017; Steinmetz et al., 2021). The quality and complexity of the recorded neural signals depend heavily on the underlying electrophysiological techniques employed (Nurmikko, 2018; Steinmetz et al., 2018). Often, spike events are inferred indirectly through spike sorting, a computational process applied to multi-unit recordings. However, this approach is prone to imprecise estimations, noise, and the mixing of signals from multiple neuronal sources (Carlson and Carin, 2019; Quiroga et al., 2004; Rey et al., 2015).

Analyzing neural spiking signals to extract meaningful features relevant to behavior has been a central focus in neuroscience (Quiroga et al., 2009). Computational methods have been used to represent spike features through wavelet transforms (Jia et al., 2022; Quiroga et al., 2004) or latent embeddings learned via deep neural networks (Schneider et al., 2023; Shen et al., 2021). In this study, we provide an approach to process spiking signals to reconstruct augmented spiking sequences, forming a novel representation. The quality of these augmented spikes was evaluated using a decoding framework, demonstrating improved performance in reconstructing visual stimulus images. Our findings align with recent advancements in spike generation and enhancement using deep learning generative models (Kapoor et al., 2024; Shen et al., 2025), offering a promising avenue for representing spiking signals for a wide range of downstream applications.

3.3. Wavelet analysis enhances deep learning models for decoding of temporal neural signals

The wavelet transform is conceptualized as a sequence of multi-scale high-pass and low-pass filters that allow signals within specific frequency ranges to pass while attenuating others. This method first identifies frequency components in the original signal and then localizes these components by sliding the filter across the signal. Consequently, wavelet transform is widely utilized in temporal signal processing (Chen et al., 2018; Liu et al., 2020; Pan et al., 2022; Sakar et al., 2019; Suzuki, 2020). By decomposing a signal through a combination of high-pass and low-pass filters, the wavelet transform generates wavelet coefficients containing high-frequency and low-frequency information. Manipulating these coefficients enables the removal of redundant information, while the original signal can be reconstructed via the inverse wavelet transform. Our work aligns with recent research utilizing wavelet transformations to analyze temporal information embedded in neural signals (Duraivel et al., 2020; Jia et al., 2022; Lopes-dos-Santos et al., 2018).

In recent years, the integration of deep learning models with wavelet transforms has gained significant attention, proving beneficial for various downstream tasks. Its potential has been demonstrated to enhance computational efficiency in image reconstruction (Liu et al., 2018) and improve image denoising capabilities (Yang et al., 2020) in the general field of computer vision. However, there are few studies on its application to neural spiking signals and visual coding. Our recent work shows that wavelet analysis can improve the decoding of static images using temporal spiking signals, compared to CNN models without using wavelets (Jia et al., 2022). The current work extends this paradigm by integrating wavelets with CNNs to process temporal spike stream signals in an end-to-end fashion for decoding dynamic videos. Within this framework, CNNs function as feature selectors and extractors of wavelet coefficients, optimizing downstream processing by reconstructing enhanced wavelet coefficients for improved model performance.

In future work, our framework is potentially extended for analysis of visual coding in other types of temporal neural signals, including invasive recordings of intracranial EEG (iEEG) or electrocorticography (ECoG), and spiking signals in humans. The challenges of these experimental modalities are the limited recording scale and time (Quiroga, 2019), as well as limited access to video stimulus (Cao et al., 2025, 2022). Although noninvasive recordings of scalp electroencephalogram (EEG) (Grootswagers et al., 2022), fMRI and magnetoencephalography (MEG) (Hebart et al., 2023), can record large scale neural populations and long-term visual stimulation, the limited spatial and/or temporal resolution and signal-to-noise ratio pose new challenges for analysis, particularly for dynamic scenes beyond static images (Allen et al., 2022). Our proposed wavelet-informed deep learning approach may provide new insights into these challenges. Further extension of our model needs to address the application ability of temporal resolutions of different types of signals to be suitable for various types of neural signals.

4. Methods

4.1. The wavelet-informed spike augmentation (WISA) module

The entire decoding network consists of two major modules, WISA for processing spikes and CNN for decoding images.

In the WISA module, we first use the discrete wavelet transform (DWT) in the time domain to decompose the spike stream signals into

multiple high-frequency components and one low-frequency component. Secondly, we construct a compact, small yet efficient module, the Temporal CNN (TCNN), which processes the signal and outputs a new signal through a learning process. Finally, we restore a new signal S' with the same size as $H \times W \times T$ through inverse discrete wavelet transform (IDWT), generating a new representation of the spikes.

4.1.1. Application of DWT on spiking signals

The DWT performs multi-scale analysis on the original signal through scaling and translation operations. In the process of wavelet decomposition, using low-pass filters (D_L) and high-pass filters (D_H) allows us to capture different frequency components of the original signal and locate their specific positions in the time domain. Typically, one-dimensional DWT includes a set of decomposition filters (D_L and D_H) and reconstruction filters (R_L and R_H) to complete the signal's decomposition and reconstruction.

When processing spike stream signals, the record of spike counts over a certain period is denoted as S. The wavelet decomposition process first uses D_H and D_L to convolve and downsample S, obtaining high-frequency wavelet coefficients S_{H1} and low-frequency wavelet coefficients S_{L1}. The reconstruction process then involves the inverse operation on S_{H1} and S_{L1}, reconstructing the signal S' as follows:

$$S_{L1} = (2\downarrow)\big(D_L * S\big),$$
$$S_{H1} = (2\downarrow)\big(D_H * S\big),$$
$$S' = R_L * (2\uparrow)S_{L1} + R_H * (2\uparrow)S_{H1}.$$

Here, $(2\downarrow)$ and $(2\uparrow)$ represent downsampling and upsampling operations, respectively, and $*$ denotes convolution. Through these steps, we obtain the low-frequency wavelet coefficient matrix F_{L1} and the high-frequency wavelet coefficient matrix F_{H1} as the results of the first layer of wavelet transform. In multi-level wavelet transformations, the low-frequency wavelet coefficients S_{L1} are further decomposed using D_H and D_L. After five levels of decomposition, the resulting wavelet coefficient set $F = F_{H1}, F_{H2}, F_{H3}, F_{H4}, F_{H5}, F_{L5}$ includes five high-frequency wavelet coefficients and one low-frequency wavelet coefficient.

4.1.2. The TCNN module

After decomposing spike stream signals along the time axis using DWT, we obtain the original signal's low-frequency and high-frequency wavelet coefficients. Our goal is to learn and extract low-frequency and high-frequency features of the signal from these wavelet coefficients. Utilizing the wavelet coefficient matrix F obtained from filter decomposition, we process F with a CNN-based feature extraction module, named the Temporal CNN (TCNN) module.

This CNN-based feature extraction module includes several convolutional layers with skip connections and activation layers. Each coefficient F_i (where F_i is a wavelet coefficient in matrix F) is input into two consecutive temporal CNN layers, with a ReLU layer in between, to obtain residual features. Adding the original F_i generates an intermediate output F_i^{mid}, which is then input into a temporal CNN layer with ReLU, ultimately producing output F_i'. In this module, all CNN layers are one-dimensional convolutional layers, mainly processing the temporal channel T in F_i. The computational process is as follows:

$$F_i^{\text{mid}} = \text{ReLU}\left(F_i * W_{\text{conv1}}\right) * W_{\text{conv2}} + F_i,$$
$$F_i' = \text{ReLU}\left(F_i^{\text{mid}} * W_{\text{conv3}}\right),$$

where, $*$ denotes the convolution operation, and W_{conv1}, W_{conv2}, and W_{conv3} are the weight matrices of the three convolutional layers.

4.1.3. Generating new spiking signals by IDWT

After the TCNN processing, we obtain a new set of wavelet coefficients $F' = F_{H1'}, F_{H2'}, F_{H3'}, F_{H4'}, F_{L5'}$, and generate a new spike signal S' of the same size as S using IDWT.

In the WISA module, the spike signal S undergoes wavelet transformation and feature learning and extraction via the TCNN module, followed by reconstruction into S' using IDWT. This process can be expressed as $S' = f_{\text{WISA}}(S)$. Since S and S' are of the same size, for the network architecture of the downstream task, we only need to introduce the WISA module at the network's head. We refer to this downstream network as 'task', and its prediction is denoted as Y_{task}. The entire computational process can be described as: $Y_{\text{task}} = f_{\text{task}}\left(f_{\text{WISA}}(S)\right)$. The outcome of this new representation of spikes is a new temporal sequence of spiking signals, which can be used for different downstream tasks.

4.2. The CNN decoding network

After processing spikes with the WISA module, a decoding network model was used to align spikes with images. We aim to evaluate the augmented spikes to validate the effectiveness of WISA in image reconstruction, so we used our previously developed decoding network model based on the CNN architecture (Yu et al., 2024; Zhang et al., 2020, 2022). Briefly, the CNN model is divided into two stages. The first stage consists of a three-layer fully connected network. The first layer receives the spike signals processed by the wavelet transform as input, with the number of units corresponding to the number of neurons in the spike signals. The second layer is a hidden layer containing 512 neurons, using the ReLU activation function and including a dropout layer to prevent overfitting. The output layer has a number of neurons equal to the number of pixels in the target image. Considering the target image size of 90x90 pixels, the output layer has 8100 neurons, using the sigmoid activation function. In this way, we obtain an intermediate image.

At the second stage, the CNN first uses convolution operations and downsampling to process and reduce the size of the intermediate image. This includes four convolutional layers with kernel sizes of (64,7), (128,5), (256,3), and (256,3), padding of 0, and a stride of (1,1), aiming to denoise the intermediate image while preserving important parts. Then, deconvolution and upsampling are used to restore the image size and texture. This stage includes four convolutional layers with kernel sizes of (256,3), (128,3), (64,5), and (1,7), padding and stride of 0 and (1,1) respectively, and the output is a grayscale image of size $H \times W \times C$. The second stage uses the ReLU activation function, and a dropout layer is included after each convolutional layer to prevent overfitting.

4.3. End-to-end training of the WISA-CNN model

The entire decoding network needs to be trained by an end-to-end fashion, with the spike trains at the input end and the visual images at the other end. The model was implemented using PyTorch. During training, we used the Adam optimizer and set the initial learning rate to 0.0001. We trained all networks for 600 epochs and adjusted the learning rate to 0.0002 after reaching 400 epochs. The models were trained on an NVIDIA-A100 (40GB) GPU with a batch size set to 16.

The Mean Squared Error (MSE) was used as the loss function that measures the pixel-level difference between the reconstructed image and the ground-truth image, making it suitable for error quantification in image restoration. By minimizing this MSE loss, the network learns to adjust its parameters to generate a reconstructed image that more closely resembles the actual image, thereby reducing reconstruction error at the pixel level.

4.4. Experimental data

The spiking signals, along with the corresponding video stimulus datasets, are taken from the previous studies (Onken et al., 2016), and detailed experimental methods and data collection processes can be found (Liu et al., 2017; Onken et al., 2016). Briefly, the decoder was applied to reconstruct dynamic videos from spikes recorded simultaneously by a population of retinal ganglion cells (RGCs) in an isolated sala-

mander retina. The retina was placed on a multielectrode array within a recording chamber. Visual stimuli were presented on an OLED display and projected onto the photoreceptor layer through a lens positioned above the retina. The dataset has two videos with various scenes of salamanders and tigers, consisting of 1800 and 1600 frames, respectively, representing different levels of scene complexity (Zheng et al., 2021). The images are 90x90 pixels, and the responses are from 90 retinal ganglion cells. The entire dataset was randomly split into 90 % as training set and 10 % test set.

4.5. Image reconstruction evaluation metrics

We employed two popular metrics to compare the similarity between the reconstructed images and the original stimulus images.

1) Peak Signal-to-Noise Ratio (PSNR) is a metric for assessing image quality, comparing the error between the original image and the processed image to the maximum possible error. The formula is as follows:

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}}\right),$$

where MAX_I is the maximum pixel value of the image (usually 255), and MSE is the mean squared error. PSNR is measured in decibels (dB), with higher values indicating better image quality.

2) Structural Similarity Index Metric (SSIM) (Wang et al., 2004) is a comprehensive image quality assessment metric that considers luminance, contrast, and structural errors. It is calculated for multiple windows of an image and is based on luminance (l), contrast (C), and structure (S). The SSIM index ranges from 0 to 1, with values closer to 1 indicating greater similarity between the reconstructed image and the original stimulus image. The formula is as follows:

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = [l(\mathbf{x},\mathbf{y})]^{\alpha} \cdot [c(\mathbf{x},\mathbf{y})]^{\beta} \cdot [s(\mathbf{x},\mathbf{y})]^{\gamma}$$

The values of α, β, and γ are usually set to 1. Then

$$\begin{cases} l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \\ c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \\ s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \end{cases},$$

where μ_x is the average of x, μ_y is the average of y, σ_x is the variance of x, σ_y is the variance of y, $c_1 = (k_1 l)^2$ and $c_2 = (k_2 l)^2$ are constants, $c_3 = c_2/2$. l is the dynamic range of the pixel values (from 0 to 255), $k_1 = 0.01$, and $k_2 = 0.03$.

CRediT authorship contribution statement

Jing Peng: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Investigation, Formal analysis, Data curation; **Shanshan Jia:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation; **Jiyuan Zhang:** Software, Resources, Methodology; **Yongxing Wang:** Supervision, Resources; **Zhaofei Yu:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization; **Jian K. Liu:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Data availability

The code is available at https://github.com/jiankliu/WISA.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grant numbers 62422601, 62176003, 62088102, Beijing Nova Program under Grant numbers 20230484362, 20240484703, and the UK Royal Society Newton Advanced Fellowship under Grant number NAF-R1-191082.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., & Kay, K., (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116–126.

Botella-Soler, V., Deny, S., Martius, G., Marre, O., Tkačik, G. (2018). Nonlinear decoding of a complex movie from the mammalian retina. *PLoS Computational Biology*, *14*(5), e1006057.

Cao, R., Brunner, P., Brandmeir, N. J., Willie, J. T., & Wang, S., (2025). A human single-neuron dataset for object recognition. *Scientific Data*, *12*(1), 79.

Cao, R., Lin, C., Brandmeir, N. J., & Wang, S. (2022). A human single-neuron dataset for face perception. *Scientific Data*, *9*(1), 365.

Carlson, D., & Carin, L. (2019). Continuing progress of spike sorting in the era of big data. *Current Opinion in Neurobiology*, *55*, 90–96.

Chen, Y., Beech, P., Yin, Z., Jia, S., Zhang, J., Yu, Z., & Liu, J. K. (2024). Decoding dynamic visual scenes across the brain hierarchy. *PLoS Computational Biology*, *20*(8), e1012297.

Chen, T., Lin, L., Zuo, W., Luo, X., Zhang, L. (2018). Learning a wavelet-like auto-encoder to accelerate deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*.

Chichilnisky, E. J. (2001). A simple white noise analysis of neuronal light responses. *Network: Computation in Neural Systems*, *12*(2), 199.

Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M., and Schmidhuber, J. (2011). Flexible, high performance convolutional neural networks for image classification. In *Proceedings of the twenty-second international joint conference on artificial intelligence* (pp. 1237–1242).

Duraivel, S., Rao, A. T., Lu, C. W., Bentley, J. N.„ Stacey, W. C., Chestek, C. A., Patil, P. G. (2020). Comparison of signal decomposition techniques for analysis of human cortical signals. *Journal of Neural Engineering*, *17*(5), 056014.

Garasto, S., Bharath, A. A., Schultz, S. R. (2018). Visual reconstruction from 2-photon calcium imaging suggests linear readout properties of neurons in mouse primary visual cortex. bioRxiv preprint 300392.

Gollisch, T. and Meister, M. (2008). Rapid neural coding in the retina with relative spike latencies. *Science*, *319*(5866), 1108–1111.

Gollisch, T. and Meister, M. (2010). Eye smarter than scientists believed: Neural computations in circuits of the retina. *Neuron*, *65*(2), 150–164.

Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep learning. MIT Press.

Grootswagers, T., Zhou, I., Robinson, A. K., Hebart, M. N., & Carlson, T. A. (2022). Human EEG recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, *9*(1), 3.

Güçlü, U., van, G., Marcel, A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., & Baker, C. I. (2023). THINGS-Data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, *12*, e82580.

Iqbal, A., Dong, P., Kim, C. M., Jang, H. (2019). Decoding neural responses in mouse visual cortex through a deep neural network. In *2019 International joint conference on neural networks (IJCNN)* (pp. 1–7).

Jia, S., Li, X., Huang, T., Liu, J. K., Yu, Z. (2022). Representing the dynamics of high-dimensional data with non-redundant wavelets. *Patterns*, *3*(3), 100424.

Kapoor, J., Schulz, A., Vetter, J., Pei, F. and Gao, R., Macke, J. H. (2024). Latent diffusion for neural spiking data. arXiv preprint arXiv:2407.08751

Karamanlis, D., Schreyer, H. M., Gollisch, T. (2022). Retinal encoding of natural scenes. *Annual Review of Vision Science*, *8*(1), 171–193.

Kim, Y. J., Brackbill, N., Batty, E., Lee, J., Mitelut, C., Tong, W., Chichilnisky, E. J., Paninski, L. (2021). Nonlinear decoding of natural images from large-scale primate retinal ganglion recordings. *Neural Computation*, *33*(7), 1719–1750.

Kleinfeld, D., Luan, L., Mitra, P. P., Robinson, J. T., Sarpeshkar, R., Shepard, K., Xie, C., Harris, T. D. (2019). Can one concurrently record electrical spikes from every neuron in a mammalian brain? *Neuron*, *103*(6), 1005–1015.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., (2002). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Li, W., Joseph, R., Alex, N., Tjahjadi, T., Zhuang, Z. (2022). Fusion of ANNs as decoder of retinal spike trains for scene reconstruction. *Applied Intelligence*, *52*(13), 15164–15176.

Li, W., Zheng, S., Liao, Y., Hong, R., He, C. and Chen, W., Deng, C., Li, X. (2023). The brain-inspired decoder for natural visual image reconstruction. *Frontiers in Neuroscience*, *17*, 1130606.

Liu, J. K., Gollisch, T. (2015). Spike-triggered covariance analysis reveals phenomenological diversity of contrast adaptation in the retina. *PLoS Computational Biology*, *11*(7), e1004425.

Liu, J. K., Schreyer, H. M., Onken, A., Rozenblit, F., Khani, M. H., Krishnamoorthy, V., Panzeri, S., Gollisch, T. (2017). Inference of neuronal functional circuitry with spike-triggered non-negative matrix factorization. *Nature Communications*, *8*(1), 149.

Liu, P., Zhang, H., Zhang, K., Lin, L., Zuo, W. (2018). Multi-level wavelet-CNN for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 773–782).

Liu, Q., Yang, S., Liu, J., Xiong, P., Zhou, M. (2020). A discrete wavelet transform and singular value decomposition-based digital video watermark method. *Applied Mathematical Modelling*, *85*, 273–293.

Lopes-dos-Santos, V., Rey, H. G., Navajas, J., Quiroga, R. Q. (2018). Extracting information from the shape and spatial distribution of evoked potentials. *Journal of Neuroscience Methods*, *296*, 12–22.

Marre, O. and Botella-Soler, V., Simmons, K. D., and Mora, T., Tkačik, G., Berry II, M. J. (2015). High accuracy decoding of dynamical motion from a large retinal population. *PLoS Computational Biology*, *11*(7), e1004304.

Meyer, A. F., Williamson, R. S., Linden, J. F., Sahani, M. (2017). Models of neuronal stimulus-response functions: elaboration, estimation, and evaluation. *Frontiers in Systems Neuroscience*, *10*.

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, *63*(6), 902–915.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, *21*(19), 1641–1646.

Nurmikko, A. V. (2018). Approaches to large scale neural recording by chronic implants for mobile BCIs. In *2018 6th international conference on brain-computer interface (BCI)* (pp. 1–2).

Olshausen, B. A., Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609.

Onken, A., Liu, J. K., Karunasekara, P. P., Chamanthi, R., Delis, I., Gollisch, T., Panzeri, S. (2016). Using matrix and tensor factorizations for the single-trial analysis of population spike trains. *PLoS Computational Biology*, *12*(11), e1005189.

Orsborn, A. L., Moorman, H. G., Overduin, S. A., Shanechi, M. M., Dimitrov, D. F. Carmena, J. M. (2014). Closed-loop decoder adaptation shapes neural plasticity for skillful neuroprosthetic control. *Neuron*, *82*(6), 1380–1393.

Pan, W., Shi, H., Zhao, Z., Zhu, J., He, X., Pan, Z., Gao, L., Yu, J., Wu, F., Tian, Q. (2022). Wnet: Audio-guided video object segmentation via wavelet-based cross-modal denoising networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1320–1331).

Parthasarathy, N., Batty, E., Falcon, W., Rutten, T. and Rajpal, M., Chichilnisky, E. J., and Paninski, L. (2017). Neural networks for efficient bayesian decoding of natural images from retinal neurons. In *Advances in neural information processing systems*.

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J., Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, *454*(7207), 995–999.

Qiao, K., Zhang, C., Wang, L., Chen, J. and Zeng, L., Tong, L., Yan, B. (2018). Accurate reconstruction of image stimuli from human functional magnetic resonance imaging based on the decoding model with capsule network architecture. *Frontiers in Neuroinformatics*, *12*, 62.

Quiroga, Q. R. (2019). Plugging in to human memory: advantages, challenges, and insights from human single-neuron recordings. *Cell*, *179*(5), 1015–1032.

Quiroga, R. Q., Nadasdy, Z., Ben-Shaul, Y. (2004). Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Computation*, *16*(8), 1661–1687.

Quiroga, R. Q., Panzeri, S. (2009). Extracting information from neuronal populations: information theory and decoding approaches. *Nature Reviews Neuroscience*, *10*(3), 173–185.

Ragni, F., Lingnau, A., Turella, L. (2021). Decoding category and familiarity information during visual imagery. *NeuroImage*, *241*, 118428.

Rey, H. G., Pedreira, C., Quian, Q. R. (2015). Past, present and future of spike sorting techniques. *Brain Research Bulletin*, *119*, 106–117.

Richards, B. A. et al. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770.

Rieke, F. (1997). Spikes: Exploring the neural code. MIT Press.

Sakar, C. O., Serbes, G., Gunduz, A., Tunc, H. C., Nizam, H., Sakar, B. E., Tutuncu, M., Aydin, T., Isenkul, M. E., Apaydin, H. (2019). A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-Factor wavelet transform. *Applied Soft Computing*, *74*, 255–263.

Schneider, S., Lee, J. H. & Mathis, M. W. (2023). Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, *617*(7960), 360–368.

Schoenmakers, S., Barth, M., Heskes, T. & van Gerven, M. (2013). Linear reconstruction of perceived images from human brain activity. *NeuroImage*, *83*, 951–961.

Shah, N. P., Chichilnisky, E. J. (2020). Computational challenges and opportunities for a bi-directional artificial retina. *Journal of Neural Engineering*, *17*(5), 055002.

Shanechi, M. M., Orsborn, A. L., Moorman, H. G. and Gowda, S., Dangi, S., Carmena, J. M. (2017). Rapid control and feedback rates enhance neuroprosthetic control. *Nature Communications*, *8*(1), 13825.

Shen, J., Liu, J. K., Wang, Y. (2021). Dynamic spatiotemporal pattern recognition with recurrent spiking neural network. *Neural Computation*, *33*(11), 2971–2995.

Shen, J., Wang, K., Gao, W., Liu, J. K., Xu, Q., Pan, G., Chen, X., Tang, H. (2025). Temporal spiking generative adversarial networks for heading direction decoding. *Neural Networks*, *184*, 106975.

Simoncelli, E. P., Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*(1), 1193–1216.

Steinmetz, N. A. et al. (2021). Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, *372*(6539), eabf4588.

Steinmetz, N. A., Koch, C., Harris, K. D., Carandini, M. (2018). Challenges and opportunities for large-scale electrophysiology with neuropixels probes. *Current Opinion in Neurobiology*, *50*, 92–100.

Strang, G., Nguyen, T. (1996). Wavelets and filter banks. SIAM.

Suzuki, T. (2020). Wavelet-based spectral-spatial transforms for CFA-sampled raw camera image compression. *IEEE Transactions on Image Processing*, *29*, 433–444.

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.-B., Lebihan, D., Dehaene, S. (2006). Inverse retinotopy: Inferring the visual content of images from brain activation patterns. *NeuroImage*, *33*(4), 1104–1116.

Wang, M. et al. (2020). Single-neuron representation of learned complex sounds in the auditory cortex. *Nature Communications*, *11*(1), 4361.

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., Liu, Z. (2018). Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, *28*(12), 4136–4160.

Wu, M. C.-K., David, S. V., Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience*, *29*(1), 477–505.

Wu, R., Zhou, F., Yin, Z., Liu, K. J. (2025). Aligning neuronal coding of dynamic visual scenes with foundation vision models. In *Computer vision - ECCV 2024* (pp. 238–254).

Yan, Q., Zheng, Y., Jia, S., Zhang, Y., Yu, Z., Chen, F., Tian, Y., Huang, T., Liu, J. K. (2022). Revealing fine structures of the retinal receptive field by deep-Learning networks. *IEEE Transactions on Cybernetics*, *52*(1), 39–50.

Yang, R., Zhao, P., Wang, L., Feng, C., Peng, C., Wang, Z., Zhang, Y., Shen, M., Shi, K., Weng, S., Dong, C., Zeng, F., Zhang, T., Chen, X., Wang, S., Wang, Y., Luo, Y., Chen, Q., Chen, Y., Jiang, C., Jia, S., Yu, Z., Liu, J., Wang, F., Jiang, S., Xu, W., Li, L., Wang, G., Mo, X., Zheng, G., Chen, A., Zhou, X., Jiang, C., Yuan, Y., Yan, B., & Zhang, J. (2023). Assessment of visual function in blind mice and monkeys with subretinally implanted nanowire arrays as artificial photoreceptors. *Nature Biomedical Engineering*, *8*(8), 1018–1039.

Yang, H.-H., Yang, C.-H. H., Wang, Y.-C. F. (2020). Wavelet channel attention module with a fusion network for single image deraining. In *2020 IEEE international conference on image processing (ICIP)* (pp. 883–887).

Yoshida, T., Ohki, K. (2020). Natural images are reliably represented by sparse and variable populations of neurons in visual cortex. *Nature Communications*, *11*(1), 1–19.

Yu, Z., Bu, T., Zhang, Y., Jia, S., Huang, T., Liu, J. K. (2024). Robust decoding of rich dynamical visual scenes with retinal spikes. *IEEE Transactions on Neural Networks and Learning Systems*, (pp. 1–14).

Yu, Z., Liu, J. K., Jia, S., Zhang, Y., Zheng, Y., Tian, Y., Huang, T. (2020). Toward the next generation of retinal neuroprosthesis: Visual computation with spikes. *Engineering*, *6*(4), 449–461.

Zhang, Y.-J., Yu, Z.-F., Liu, J. K., Huang, T.-J., Huang, T.-J. (2022). Neural decoding of visual information across different neural recording modalities and approaches. *Machine Intelligence Research*, *19*(5), 350–365.

Zhang, Y., Jia, S., Zheng, Y., Yu, Z., Tian, Y., Ma, S., Huang, T., Liu, J. K. (2020). Reconstruction of natural visual scenes from neural spikes with deep neural networks. *Neural Networks*, *125*, 19–30.

Zhang, Y., Bu, T., Zhang, J., Tang, S., Yu, Z., Liu, J. K., Huang, T. (2022). Decoding pixel-level image features from two-photon calcium signals of macaque visual cortex. *Neural Computation*, *34*(6), 1369–1397.

Zhao, J., Xiong, R., Liu, H., Zhang, J., Huang, T. (2021). Spk2imgnet: Learning to reconstruct dynamic scene from continuous spike stream. In *Proceedings of the 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 11991–12000).

Zheng, Y., Jia, S., Yu, Z., Liu, J. K., Huang, T. (2021). Unraveling neural coding of dynamic natural visual scenes via convolutional recurrent neural networks. *Patterns*, *2*(10), 100350.