A Survey of Retentive Network

Anonymous ACL submission

Abstract

Retentive Network (RetNet) represents a significant advancement in neural network architecture, offering an efficient alternative to the Transformer. While Transformers rely on self-attention to model dependencies, they suffer from high memory costs and limited scalability when handling long sequences due to their quadratic complexity. To mitigate these limitations, RetNet introduces a retention mechanism that unifies the inductive bias of recurrence with the global dependency modeling of attention. This mechanism enables linear-time inference, facilitates efficient modeling of extended contexts, and remains compatible with fully parallelizable training pipelines. RetNet has garnered significant research interest due to its consistently demonstrated cross-domain effectiveness, achieving robust performance across machine learning paradigms including natural language processing, speech recognition, and time-series analysis. However, a comprehensive review of RetNet is still missing from the current literature. This paper aims to fill that gap by offering the first detailed survey of the RetNet architecture, its key innovations, and its diverse applications. We also explore the main challenges associated with RetNet and propose future research directions to support its continued advancement in both academic research and practical deployment.

1 Introduction

Vaswani et al. (2017) proposed the Transformer architecture, which relies solely on the self-attention mechanisms. Owing to its ability to model long-range dependencies and its high degree of parallelism, the Transformer has emerged as the dominant paradigm in natural language processing (NLP). Beyond NLP, the Transformer has been successfully applied to a wide range of domains such as computer vision (CV), speech, and scientific areas like chemistry and bioinformatics, reflecting its versatility in modeling complex, long-range dependencies across modalities. Despite its strengths, the Transformer architecture faces notable limitations. During training, its quadratic time complexity makes modeling long sequences computationally costly. In the inference phase, linear memory complexity arises from storing KV cache for each token, resulting in significant memory overhead. Although various approaches have been explored to mitigate the complexity of the Transformer, achieving substantial reductions in computational overhead remains challenging(Choromanski et al., 2020; Katharopoulos et al., 2020; Wang et al., 2020).

To address the computational limitations of traditional Transformers, many research advances have emerged. Gated linear recurrent neural networks (Qin et al., 2023; De et al.) incorporated gating mechanisms to reduce the quadratic time complexity typically associated with Transformer training. State Space Models compressed sequence data into fixed-size representations, effectively mitigating the scaling issues inherent in Transformers (Gu et al., 2021; Gu and Dao, 2023). Linear Transformers (Katharopoulos et al., 2020) further alleviated memory and computational overhead by employing linear attention mechanisms, allowing both time and memory complexity to scale linearly with sequence length. The Receptance Weighted Key Value (RWKV) leverages linear attention to reduce computational complexity and memory usage during inference (Peng et al., 2023; Li et al., 2024). Among these, RetNet (Sun et al., 2023) stands out as a compelling solution, it integrated a multi-scale retention mechanism which employs three computational paradigms namely parallel, recurrent, and chunkwise recurrent representations. By leveraging these paradigms, RetNet achieves performance comparable to Transformers while enabling constant-time O(1) inference, reduced memory overhead, and efficient long-sequence model-

*Corresponding authors

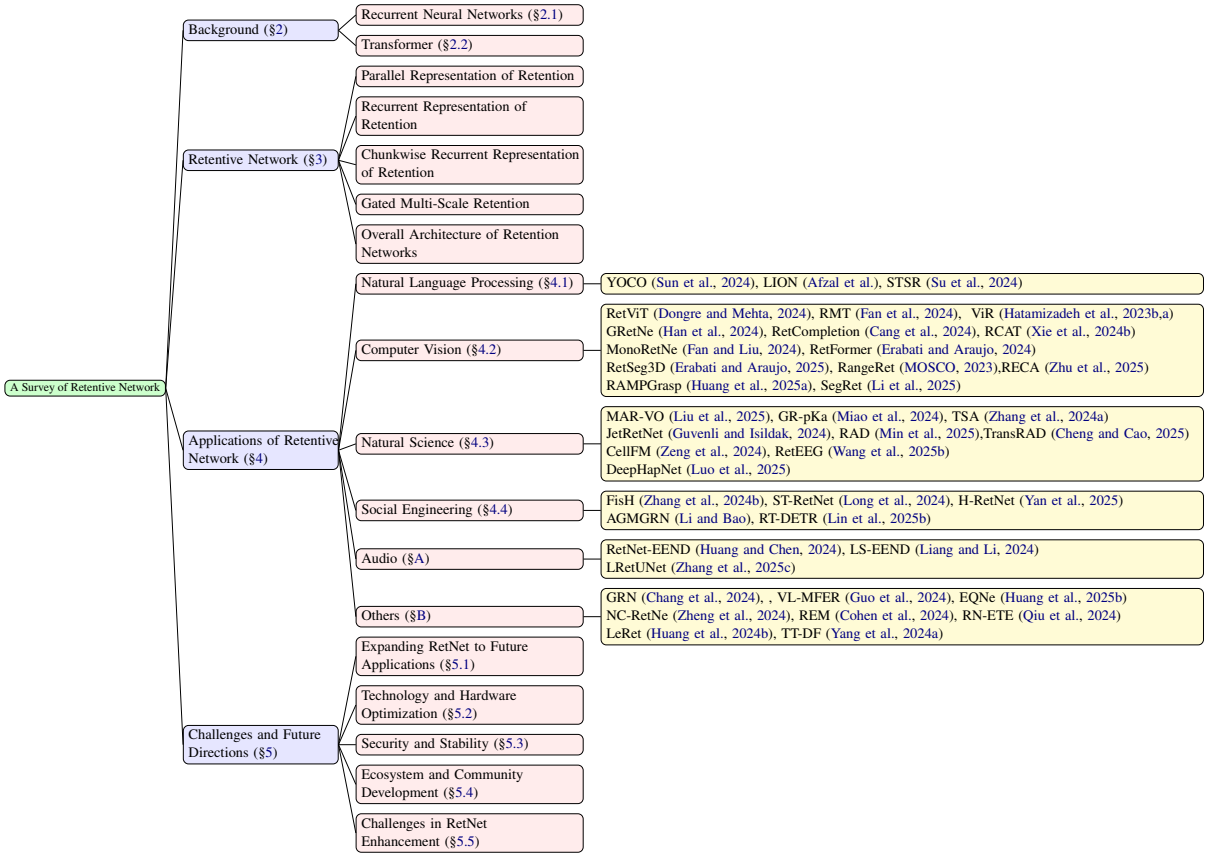


Figure 1: Structure of this paper.

ing.

By employing the retention mechanism, the decay mask makes RetNet very versatile for a wide range of applications, from NLP (Cheng et al., 2024), CV (Fan et al., 2024), natural science (Luo et al., 2025) to social engineering (Yan et al., 2025). With the rapid expansion of research and applications of RetNet, this survey aims to shed light on current progress in this field. As depicted in Figure 1, the remainder of this paper is organized as follows: Section 2 provides a systematic review of basic concepts, including RNN and Transformer architectures, Section 3 delves into the principle and mechanism of RetNet, and Section 4 explores the extensive applications of RetNet in diverse domains, including NLP, CV, natural sciences, social engineering, and audio processing. Section 5 examines the primary challenges confronting RetNet and outlines prospective directions for future research.

2 Background

2.1 Recurrent Neural Networks

Recurrent neural networks (RNNs) are capable of learning features and long-term dependencies from sequential and time-series data (Salehinejad et al., 2017). Specifically, RNN introduced a recurrent architecture that maintains a hidden state, enabling the modeling of sequential data with variable lengths through shared weights across time steps (Hochreiter and Schmidhuber, 1997). The operational mechanism of RNN is captured by the following mathematical formulation:

$$h_t = f_H(W_{hh} \cdot h_{t-1} + W_{hx} \cdot x_t + b_h) \quad (1)$$

$$y_t = f_O(W_{ho} \cdot h_t + b_o) \quad (2)$$

where h_t denotes the hidden state at time step t, and x_t is the input at time t. The function $f_H(\cdot)$ is the hidden layer activation function, and $f_O(\cdot)$ is the output activation function. y_t denotes the output at time t. W_{hh}, W_{hx}, and W_{ho} are the weight matrices connecting the hidden-to-hidden, input-to-hidden, and hidden-to-output layers, respectively. b_h and b_o are the bias vectors for the hidden and output layers.

Despite the RNN's strength in modeling temporal sequences, a major limitation is the vanishing gradient problem, which causes gradients to decay exponentially over time steps. This significantly impairs the network's ability to retain and utilize information from distant past inputs (Bengio et al., 1994).

To mitigate the challenges of vanishing or exploding gradients encountered by RNN when pro-

cessing extended sequences, researchers have developed several advanced variants. Long short-term memory network (LSTM) (Hochreiter and Schmidhuber, 1997) incorporates gating mechanisms to regulate information flow, enabling effective capture of long-term dependencies. Gated recurrent unit (GRU) (Cho et al., 2014) offers a simplified architecture compared to LSTM while delivering comparable performance. Bidirectional recurrent neural network (Bi-RNN) (Schuster and Paliwal, 1997) processes sequences in both forward and reverse directions simultaneously, providing a more comprehensive understanding of sequential patterns.

2.2 Transformer

The Transformer architecture dispenses with recurrence entirely, relying instead on self-attention mechanisms to model global dependencies between inputs and outputs (Vaswani et al., 2017). This fundamental shift enables the model to more effectively capture long-range relationships and supports stable, efficient training without the gradient propagation issues commonly associated with recurrent structures.

The attention mechanism enables models to capture dependencies among different positions within an input sequence, which is essential for learning contextual relationships. In Self-Attention, the input sequence is represented as a matrix $X \in \mathbb{R}^{n \times d}$, where n is the number of tokens and d is the dimensionality of each token embedding. To generate the necessary attention components, the Transformer applies three independent trainable linear transformations to the input: the matrix X is projected into the query, key, and value spaces using the weight matrices $W^Q \in \mathbb{R}^{d \times d_q}$, $W^K \in \mathbb{R}^{d \times d_k}$, and $W^V \in \mathbb{R}^{d \times d_v}$. As a result, we obtain $Q = XW^Q$, $K = XW^K$, and $V = XW^V$.

Each row in Q, K, and V corresponds to a token in the sequence, where Q represents the queries used to attend to other tokens, K represents the keys that determine relevance, and V carries the actual content of the tokens. The Self-Attention output is computed using the scaled dot-product attention mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \quad (3)$$

where d_k denotes the dimensionality of the key vectors. The scaling factor $\sqrt{d_k}$ is used to mitigate the impact of large dot-product values, ensuring stable gradients and more effective learning.

The self-attention mechanism in the Transformer is extended to multiple attention heads, each capable of learning distinct attention weights to effectively capture diverse relational patterns. Multi-head attention enables the model to process different informational subspaces in parallel, enhancing its representational capacity.

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (4)$$

$$\begin{aligned} &MultiHead(Q, K, V) \\ &= Concat(head_1, \ldots, head_h)W^O \end{aligned} \quad (5)$$

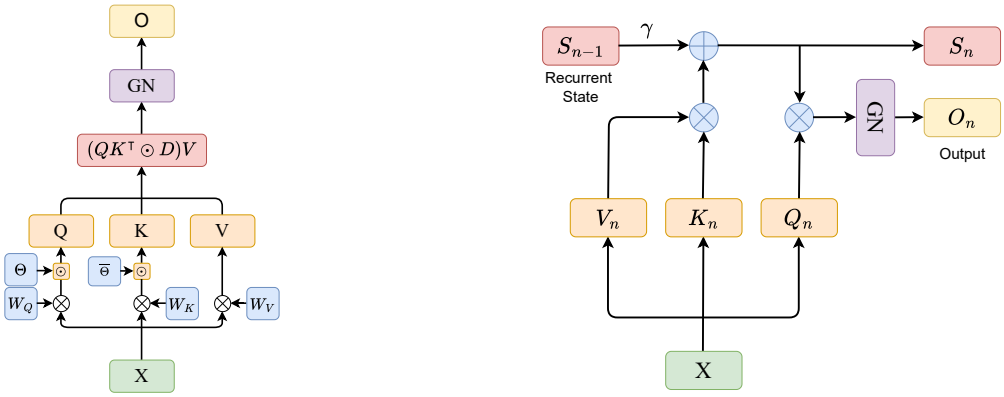
where h denotes the number of attention heads, $Concat$ represents the concatenation operation, and W^O is the trainable projection matrix, W_i^Q, W_i^K, and W_i^V are the parameter matrice. Multi-head attention significantly enhances the Transformer's capability to address NLP tasks and other forms of sequential data processing with improved efficiency and expressiveness.

3 Retentive Network

RNNs have difficulty capturing long-range dependencies due to the vanishing gradient problem and their inherently sequential structure, which also limits parallelism (Yu et al., 2019). Transformer, while effective at capturing long-range dependencies, face high computational complexity and inefficiency in processing long sequences (Lin et al., 2022). RetNet(Sun et al., 2023) theoretically derived the connection between recurrence and attention and proposed retention mechanism for sequence modeling. RetNet has been shown to achieve low-cost inference, efficient long-sequence modelling, Transformer-comparable performance, and parallel model training simultaneously.

RetNet is constructed as a stack of L identical blocks, each comprising two core components: a Multi-Scale Retention (MSR) module and a Feed-Forward Network (FFN) module. For a given sequence of input $x = x_1 \cdots x_{|j|}$, where $|j|$ represents the length of the sequence, RetNet utilizes an autoregressive encoding method to process the sequence. The input is packed into $X^0 = [\mathbf{x}_1, \cdots, \mathbf{x}_{|j|}] \in \mathbb{R}^{|j| \times d_{\text{model}}}$, where d_{model} is the dimension of the hidden layer. Then compute the contextualized vector representations as follows:

$$X^l = \text{RetNet}_l(X^{l-1}), l \in [1, L]. \quad (6)$$



(a) Parallel representation.

(b) Recurrent representation.

Figure 2: Dual form of RetNet. "GN" denotes GroupNorm.

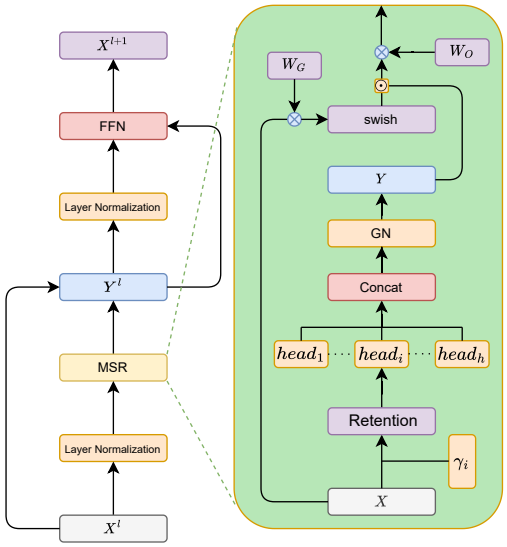


Figure 3: Overall architecture of RetNet.

Retention mechanism with a dual form of recursion and parallelism is the key to the success of RetNet. Project the input $X \in \mathbb{R}^{|j| \times d_{\text{model}}}$ to $v_n = X_n \cdot w_v$, where w_v is the trainable matrix that maps inputs to value vectors. Then make the projection Q, K:

$$Q = XW^Q, \quad K = XW^K, \tag{7}$$

where $W^Q, W^K \in \mathbb{R}^{d \times d}$ are learnable matrices.

Consider a sequence modeling problem, through the state $\mathbf{s}_n \in \mathbb{R}^{d \times d}$ mapping v_n to a vector of o_n.

$$\begin{aligned} \mathbf{s}_n &= A\mathbf{s}_{n-1} + K_n^\top v_n \\ o_n &= Q_n \mathbf{s}_n = \sum_{m=1}^{n} Q_n A^{n-m} K_m^\top v_m \end{aligned} \tag{8}$$

where K_n, Q_n is the projection of the time step n.

Further, diagonalize $A = \Lambda(\gamma e^{i\theta})\Lambda^{-1}$, where Λ is the reversible matrix, γ is the decay mask, according to Euler's formula $e^{i\theta} = [\cos\theta_1, \sin\theta_2, \cdots, \cos\theta_{d-1}, \sin\theta_d]$, then $A^{n-m} = \Lambda(\gamma e^{i\theta})^{n-m}\Lambda^{-1}$, n, m is the time step. Equation 8 becomes:

$$\begin{aligned} o_n &= \sum_{m=1}^{n} (Q_n(\gamma e^{i\theta})^n)(K_m(\gamma e^{i\theta})^{-m})^\top v_m \\ &= \sum_{m=1}^{n} \gamma^{n-m}(Q_n e^{in\theta})(K_m e^{im\theta})^\dagger v_m \end{aligned} \tag{9}$$

where $Q_n(\gamma e^{i\theta})^n$, $K_m(\gamma e^{i\theta})^{-m}$ is the xPos (Sun et al., 2022), † is the conjugate transpose. $e^{in\theta}$ and $e^{im\theta}$ serve as rotational factors that encode positional information using complex exponential forms, where θ denotes the learnable parameters employed to model relative phase differences for the purpose of capturing sequential dependencies.

Parallel Representation of Retention As shown in Figure 2a, the retention layer is defined as:

$$\begin{aligned} &Q = (XW^Q) \odot \Theta, \quad K = (XW^K) \odot \overline{\Theta}, \\ &V = XW^V, \\ &D_{nm} = \begin{cases} \gamma^{n-m}, & n \geq m \\ 0, & n < m \end{cases}, \\ &\text{Retention}(X) = (QK^\top \odot D)V \end{aligned} \tag{10}$$

where $\odot$ is the Hadamard product, Θ is the position-dependent modulation term, and $\overline{\Theta}$ denotes its complex conjugate, and $D \in \mathbb{R}^{|j| \times |j|}$ constitutes a unified matrix that jointly encodes causal masking and exponential decay as a function of relative positional distance.

Recurrent Representation of Retention As shown in Figure 2b, at the n-th timestep, the output is recurrently obtained as follows:

$$\begin{aligned} S_n &= \gamma S_{n-1} + K_n^\top V_n \\ \text{Retention}(X_n) &= Q_n S_n, n = 1, \cdots, |j| \end{aligned} \tag{11}$$

Chunkwise Recurrent Representation of Retention The input sequences are segmented into chunks. Within each chunk, the computation is carried out using the parallel representation Equation 10. In contrast, information across chunks is propagated using the recurrent representation Equation 11. Specifically, let B denote the chunk length. The retention output of the i-th chunk is computed as follows:

$$\begin{aligned} Q_{[i]} &= Q_{Bi:B(i+1)}, \\ K_{[i]} &= K_{Bi:B(i+1)}, \\ V_{[i]} &= V_{Bi:B(i+1)}, \\ R_i &= K_{[i]}^\top (V_{[i]} \odot \zeta) + \gamma^B R_{i-1}, \\ \text{Retention}(X_{[i]}) &= \underbrace{(Q_{[i]} K_{[i]}^\intercal \odot D) V_{[i]}}_{\text{Inner-Chunk}} \\ &\quad + \underbrace{(Q_{[i]} R_{i-1}) \odot \xi}_{\text{Cross-Chunk}} \\ \xi_{ij} &= \gamma^{i+1}, \quad \zeta_{ij} = \gamma^{B-i-1} \end{aligned} \tag{12}$$

where $[i]$ indicates the i-th chunk, i.e., $x_{[i]} = [x_{(i-1)B+1}, \cdots, x_{iB}]$. ζ and ξ are exponential decay factors that modulate the influence of intra-chunk and inter-chunk information.

Gated Multi-Scale Retention In each layer, the number of retention heads is defined as $h = d_{\text{model}}/d$, where d denotes the head dimension. Each head is associated with distinct parameter matrices $W^Q, W^K, W^V \in \mathbb{R}^{d \times d}$. MSR mechanism assigns a unique decay factor γ to each head. For simplicity, identical γ values are used across different layers and kept fixed. To enhance the non-linearity of the retention layers, a swish gate (Hendrycks and Gimpel, 2016; Ramachandran et al., 2017) is introduced. Given the input X, the computation of the layer is defined as follows:

$$\begin{aligned} \gamma &= 1 - 2^{-5-\text{arange}(0,h)} \in \mathbb{R}^h \\ \text{head}_i &= \text{Retention}(X, \gamma_i) \\ Y &= \text{GN}_h(\text{Concat}(\text{head}_1, \cdots, \text{head}_h)) \\ \text{MSR}(X) &= (\text{swish}(XW^G) \odot Y) W^O \end{aligned} \tag{13}$$

where $W^G, W^O \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ are learnable parameter matrices. $\text{arange}(0, h)$ denotes a vector of integers from 0 to $h-1$, used to assign distinct decay scales across h attention heads. GN denotes Group Normalization (Wu and He, 2018), applied to each head output following the SubLN strategy in (Shoeybi et al., 2019). Since each head employs a distinct γ scale, their output variances differ, which necessitates separate normalization.

Overall Architecture of Retention Networks As illustrated in Figure 3, an L-layer retention network is constructed by stacking MSR and FFN modules. The input sequence $\{x_i\}_{i=1}^{|j|}$ is first mapped to vector representations via a word embedding layer. The resulting embeddings, denoted as $X_0 = [x_1, \cdots, x_{|j|}] \in \mathbb{R}^{|j| \times d_{\text{model}}}$, serve as the initial input to the model. The final output is represented as X^L.

$$\begin{aligned} Y^l &= \text{MSR}(\text{LN}(X^l)) + X^l \\ X^{l+1} &= \text{FFN}(\text{LN}(Y^l)) + Y^l \end{aligned} \tag{14}$$

where $\text{LN}(\cdot)$ denotes the Layer Normalization function (Ba et al., 2016). The feed-forward network (FFN) is defined as

$$\text{FFN}(X) = \text{gelu}(XW_1)W_2,$$

where W_1 and W_2 are learnable parameter matrices, and $\text{gelu}(\cdot)$ is the Gaussian Error Linear Unit activation function.

4 Applications of Retentive Network

4.1 Natural Language Processing

RetNet has proven to be highly effective in a variety of NLP tasks due to its efficient retention mechanism. In language modeling, the decoder-decoder architecture with gated retention mechanism was introduced by Sun et al. (2024) to improve contextual understanding. For knowledge graph reasoning, Cheng et al. (2024) utilized RetNet as an encoder. In multi-hop reasoning tasks, the STSR model presented by Su et al. (2024) employed RetNet's parallel retention module to speed up training

and enhance performance in sequence-to-sequence reasoning tasks. The LION framework, developed by Afzal et al., adapted RetNet for bidirectional language tasks, incorporating fixed decay masks to efficiently capture long-range dependencies and reduce computational costs. Afzal et al. (2025) further refined this idea in LION-D, a bidirectional variant of RetNet that supports linear-time inference while preserving the efficiency of parallel training. He et al. (2024) introduced DenseRetNet, which improves feature extraction by integrating dense hidden connections.

4.2 Computer Vision

RetNet and their variants have demonstrated broad applicability across various CV domains. The fixed decay mask enables RetNet to efficiently capture long-range spatial or temporal dependencies in images or videos.

Image Tasks. RetViT replaces standard attention with parallelizable retention blocks to accelerate training while maintaining representational capacity (Dongre and Mehta, 2024). Fan et al. (2024) extend RetNet's one-dimensional unidirectional decay matrix to a two-dimensional bidirectional decay matrix, thereby designing Manhattan Self-Attention (MaSA). ViR explores efficient vision backbones by redesigning the retention mechanism to support both parallel training and recurrent inference (Hatamizadeh et al., 2023b). Another ViR model leverages RetNet's block structure and multi-scale design to recursively capture contextual dependencies across spatial scales (Hatamizadeh et al., 2023a). Hu et al. (2024a) proposed the SwiFTeR architecture, which employs the Retention mechanism in the fusion model's decoder. SegRet applies multi-scale retention modules to strengthen hierarchical feature aggregation, boosting semantic segmentation accuracy (Li et al., 2025). Retention mechanism helps hyperspectral models reduce memory cost while preserving spectral discriminability (Arya et al., 2025; Paheding et al., 2024). GRetNet enhances spatial feature modeling via Gaussian-decayed retention based on Manhattan distance (Han et al., 2024). Incorporating MaSA into LoFTR-like frameworks improves coarse feature matching in challenging correspondence tasks (Sui et al., 2024). Multi-focus image fusion benefits from bidirectional 2D retention that captures local spatial consistency (Huang et al., 2024a). RetCompletion applies a fast parallelized retentive decoder for real-time image inpainting (Cang et al., 2024). The Cross-Axis Transformer integrates RetNet's recurrent retention mechanism to process visual attention across chunked image regions (Erickson, 2023). The RetNet-based retention module is cleverly applied to rotating target detection (Liu et al., 2024b).

Video Tasks. RCAT combines RetNet with CLIP adapters, yielding strong results in video recognition across different datasets (Xie et al., 2024b). Maskable RetNet introduces learnable masking strategies, improving temporal localization in moment retrieval (Hu et al., 2024b). MonoRetNet proposed a half-duplex bidirectional retention design for monocular depth prediction from sequential frames (Fan and Liu, 2024).

3D Data Modeling. RetFormer incorporates spatial retention module tailored to 3D Transformer backbones (Erabati and Araujo, 2024). LION models point cloud sequences with linear-time complexity by applying groupwise retention in RNN-style architectures (Liu et al., 2024c). RetSeg3D extends the retention concept from one-dimensional sequences to 3D voxel grids for improved semantic parsing (Erabati and Araujo, 2025). RangeRet introduces a Manhattan distance-based spatial decay, enhancing context aggregation in LiDAR segmentation tasks (MOSCO, 2023). Octree-Retention Fusion exploits parallel retention with exponential decay masks to improve hierarchical context modeling in point cloud compression (Zhang et al., 2024c).

Cross-modal Tasks. RECA refines multi-hop reasoning in VQA tasks by integrating decay-aware attention across modalities (Zhu et al., 2025). In UAV geolocation, RMT's spatially constrained retention enables robust matching between aerial and satellite views (Lin et al., 2024). For image fusion, RetNet facilitates cross-modal shared feature extraction, enabling more coherent integration of text and visual signals (Wang et al., 2025c).

Robotic Perception. RAMPGrasp deploys multiscale retention to improve robustness against occlusions and cluttered scenes (Huang et al., 2025a). VVNet fuses RetNet modules with ViT backbones to address noise and visibility challenges in underwater imagery (Liu et al., 2024a). HFA-Net employs RMT's MaSA to embed fine-grained spatial priors for subtle facial movement detection (Zhang et al., 2025b). The RetNet-based RMT block is

integrated into the YOLOv9s backbone network, enhancing local and global feature extraction capabilities (Xu et al., 2024).

4.3 Natural Science

RetNet has demonstrated remarkable versatility across diverse domains in the natural sciences, attributed primarily to its efficient retention mechanism, scalable attention modeling, and capacity to encode long-range dependencies.

Chemistry. Miao et al. (2024) augmented molecular feature learning by embedding the retention structure during information propagation. Knitter et al. (2024); Knitter (2024) applied Retenet to Neural-Network Quantum States (NQS). RetNet's utility extends to lithium-ion battery state-of-health (SoH) estimation, where its retention mechanism excels in capturing temporal degradation patterns (Chen et al., 2024).

Physics. JetRetNet exploits retention mechanism to encode multiscale dependencies among tracking and vertex features (Guvenli and Isildak, 2024). Radio-frequency signal classification integrates bidirectional retention and cross-block state fusion to accommodate the causal structure of modulation tasks (Han et al., 2025). RAD addresses anomaly detection in cyber-physical systems by employing multi-scale retention and rotational positional encodings to model long-term dependencies efficiently (Min et al., 2025). Cheng and Cao (2025) exploited RMT enables fine-grained target modeling in radar perception by distributing attention according to spatial proximity. RetNet has been adapted to assess transient stability through a time-adaptive framework (Zhang et al., 2024a). In planetary environments, RetNet's retention mechanism has been employed for unmanned aerial vehicle (UAV) monocular visual odometry (Liu et al., 2025).

Biology and Medicine. In biomedical sequence analysis, RetNet have been explored for haplotype assembly (Luo et al., 2025). Liu et al. (2024d) incorporated RetNet to capture spike protein features. Two RETNets are used to extract drug features and protein features, respectively (Peng et al., 2024). In transcriptomic analysis, RetNet variant processes large-scale cell data (Zeng et al., 2024).

For EEG decoding, RetNet captures bidirectional temporal dependencies (Wang et al., 2025b). RetNet denoises EEG signals (Wang et al., 2024a). RetNet decodes temporal patterns for emotion recognition, fused with spatial features (Xu et al., 2025).

In medical image processing, ELKarazle et al. (2023) enabled real-time polyp segmentation with bidirectional retention. Chu et al. (2024) enhanced polyp segmentation with spatial distance-based retention. RetNet models spatial correlations for echocardiography segmentation (Lin et al., 2025a). Zhou et al. (2024) improved CT denoising via co-retention mechanism. RetNet captures rotation-invariant features for medical image classification (Li and Huang, 2025).

4.4 Social Engineering

RetNet has been extensively adopted in various societal engineering tasks due to its powerful capability to capture long-range dependencies and retain critical information throughout sequential modeling.

In building change detection, RetNet extracts and preserves spatial features from remote sensing images (Lin and Piao, 2024). In fire detection, RetNet's Local Attention (LA) enhances global feature extraction in YOLO-based models (Kim et al., 2024). For earthquake early warning, RetNet encodes nonlinear couplings in seismic wave embeddings (Zhang et al., 2024b). In photovoltaic forecasting, RetNet extracts high-order features from hazy weather data (Yang et al., 2024b). In bridge damage assessment, RetNet captures critical features under varying conditions, enhancing anomaly detection (Wang et al., 2025a). For track circuit entity recognition, multi-scale retention (MSR) optimizes long-distance dependency modeling (Chen et al., 2025). In human-robot collaboration, 3DMaSA extends RetNet to predict force and velocity from video sequences (Domínguez-Vidal and Sanfeliu, 2024). In coal gangue identification, RetNet's retention mechanism optimizes model size and inference speed (Zhang et al., 2025d). The application in tea disease detection, using RetNet for spatial modeling (Lin et al., 2025b).

For urban traffic flow prediction, The Temporal Self-Retention (TSR) block to decode time-dependent features extracted by the Temporal Self-Attention module (Li and Bao). Spatial RetNet and temporal RetNet both utilizing multi-scale retention mechanism to effectively capture spatial and temporal dependencies (Zhu et al., 2024). Another study by employing RetNet's causal decay mechanism in the temporal branch (Long

et al., 2024). Meanwhile, H-RetNet supports heterogeneous inputs through parallel branches and modality-specific retention strategies (Yan et al., 2025). For more applications of RetNet, the reader is referred to A, B.

5 Challenges and Future Directions

5.1 Expanding RetNet to Future Applications

In multimodal sentiment analysis, RetNet integrates textual, visual, and auditory signals for accurate emotion recognition. In autonomous systems, it processes real-time LIDAR, camera, and audio streams to enhance decision-making. Future work should focus on improving cross-modal alignment and robustness to noise and resolution variability.

In healthcare, RetNet's capacity to model high-dimensional longitudinal data enables predictive modeling for personalized medicine. By leveraging data from electronic health records, wearables, and medical imaging, RetNet can forecast disease trajectories such as chronic or neurodegenerative conditions.

5.2 Technology and Hardware Optimization

As RetNet's applications scale to large-scale deployments, energy efficiency becomes a critical consideration, particularly for edge and mobile environments. The retention mechanism's reduced computational complexity offers inherent energy savings compared to Transformers, but further optimizations are necessary to meet the demands of sustainable computing.

RetNet's distinct computational patterns, including parallel, recurrent, and chunkwise recurrent blending, require customized hardware solutions to fully exploit its efficiency. Developing RetNet-specific accelerators, such as retention-aware compute units, can significantly enhance performance.

5.3 Security and Stability

Similar to Transformer-based models, RetNet exhibits vulnerabilities to adversarial attacks. Its retention mechanism, while effective for capturing long-term dependencies, may amplify sensitivity to adversarial perturbations, potentially compromising reliability in safety-critical applications such as secure communications or surveillance systems.

RetNet models, trained on large-scale datasets, often inherit societal biases, such as those related to gender, race, or socioeconomic status, embedded in the training data. These biases can subtly skew predictions, leading to unintended and inequitable outcomes.

5.4 Ecosystem and Community Development

The widespread adoption of RetNet, in its capacity as an emerging neural architecture, is contingent on the development of a robust open source ecosystem. The success of architectures such as Transformer, BERT, and GPT, whose proliferation has been significantly supported by accessible and well-maintained open source libraries, has provided a foundation for RetNet. The requirement for an easy-to-use, feature-complete, and extensible software framework to accelerate research and deployment is equally important.

In addition, the establishment of standardized, challenging benchmark datasets and evaluation protocols is critical for objectively assessing RetNet's performance across a range of tasks and domains. Such benchmarks will not only facilitate fair comparisons with existing models, but also ensure the reliability, robustness, and generalizability of RetNet in real-world applications.

5.5 Challenges in RetNet Enhancement

While the explicit decay mechanism in RetNet enables efficient modeling of long-range dependencies by attenuating past information over time, it also introduces limitations in adaptively preserving salient contextual signals. The current decay formulation, typically based on fixed or predefined functions, may not fully capture the dynamic nature of temporal importance across different tasks or modalities. Future research should explore adaptive or learnable decay strategies that allow the model to modulate memory retention based on input characteristics or task demands. For instance, incorporating gating mechanisms or attention-informed decay functions could enable RetNet to selectively preserve or forget past information more effectively.

6 Conclusion

This paper presents a thorough survey of the current literature on RetNet, offering analytical insights, exploring practical applications, and outlining key challenges and future research directions. As the first dedicated review of RetNet to our knowledge, this work seeks to capture the evolving landscape of RetNet-related studies and provide valuable perspectives to support continued innovation and exploration in this emerging field.

Limitations

This paper provides a comprehensive review of Retentive Network (RetNet) and its theoretical foundations, empirical performance, and practical applications. Nevertheless, due to the rapidly evolving nature of this research area, especially in terms of variant designs and scalability optimization, certain notable works may have been inadvertently omitted. Additionally, a number of studies discussed in this survey rely on earlier benchmarks or smaller-scale evaluations, which may not fully reflect the performance characteristics of RetNet models in large-scale or real-world scenarios. We encourage future research to incorporate state-of-the-art implementations and diverse application contexts to offer more comprehensive and practical guidance.

References

Arshia Afzal, Elias Abad Rocamora, Leyla Naz Candogan, Pol Puigdemont, Francesco Tonin, Yongtao Wu, Mahsa Shoaran, and Volkan Cevher. Lion: A bidirectional framework that trains like a transformer and infers like an rnn.

Arshia Afzal, Elias Abad Rocamora, Leyla Naz Candogan, Pol Puigdemont, Francesco Tonin, Yongtao Wu, Mahsa Shoaran, and Volkan Cevher. 2025. Linear attention for efficient bidirectional sequence modeling. *arXiv preprint arXiv:2502.16249*.

Rajat Kumar Arya, Subhojit Paul, and Rajeev Srivastava. 2025. An efficient hyperspectral image classification method using retentive network. *Advances in Space Research*, 75(2):1701–1718.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Yueyang Cang, Pingge Hu, Xiaoteng Zhang, Xingtong Wang, Yuhang Liu, and Li Shi. 2024. Retcompletion: High-speed inference image completion with retentive network. *arXiv preprint arXiv:2410.04056*.

Qian Chang, Xia Li, and Xiufeng Cheng. 2024. Graph retention networks for dynamic graphs. *arXiv preprint arXiv:2411.11259*.

Guanxu Chen, Fangfang Yang, Weiwen Peng, Yuqian Fan, and Ximin Lyu. 2024. State-of-health estimation for lithium-ion batteries based on kullback–leibler divergence and a retentive network. *Applied Energy*, 376:124266.

Yanrui Chen, Guangwu Chen, and Peng Li. 2025. Named entity recognition in track circuits based on multi-granularity fusion and multi-scale retention mechanism. *Electronics*, 14(5):828.

Jun Cheng, Tao Meng, Xiao Ao, and Xiaohua Wu. 2024. Pre-training retnet of simulating entities and relations as sentences for knowledge graph reasoning. In *2024 4th Asia Conference on Information Engineering (ACIE)*, pages 6–10. IEEE.

Lei Cheng and Siyang Cao. 2025. Transrad: Retentive vision transformer for enhanced radar object detection. *IEEE Transactions on Radar Systems*.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, and 1 others. 2020. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*.

Jinghui Chu, Wangtao Liu, Qi Tian, and Wei Lu. 2024. Pfprnet: A phase-wise feature pyramid with retention network for polyp segmentation. *IEEE Journal of Biomedical and Health Informatics*.

Lior Cohen, Kaixin Wang, Bingyi Kang, and Shie Mannor. 2024. Improving token-based world models with parallel observation prediction. *arXiv preprint arXiv:2402.05643*.

Soham De, Samuel L Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, and 1 others. Griffin: Mixing gated linear recurrences with local attention for efficient language models, 2024. *URL https://arxiv.org/abs/2402.19427*, page 50.

JE Domínguez-Vidal and Alberto Sanfeliu. 2024. Force and velocity prediction in human-robot collaborative transportation tasks through video retentive networks. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9307–9313. IEEE.

Shreyas Dongre and Shrushti Mehta. 2024. Retvit: Retentive vision transformers. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–8. IEEE.

Khaled ELKarazle, Valliappan Raman, Caslon Chua, and Patrick Then. 2023. Retseg: Retention-based colorectal polyps segmentation network. *arXiv preprint arXiv:2310.05446*.

Gopi Krishna Erabati and Helder Araujo. 2024. Retformer: Embracing point cloud transformer with retentive network. *IEEE Transactions on Intelligent Vehicles*.

Gopi Krishna Erabati and Helder Araujo. 2025. Retseg3d: Retention-based 3d semantic segmentation for autonomous driving. *Computer Vision and Image Understanding*, 250:104231.

Lily Erickson. 2023. Cross-axis transformer with 3d rotary positional embeddings. *Preprint*, arXiv:2311.07184.

Dengxin Fan and Songyan Liu. 2024. Monoretnet: A self-supervised model for monocular depth estimation with bidirectional half-duplex retention. In *International Conference on Intelligent Computing*, pages 361–372. Springer.

Qihang Fan, Huaibo Huang, Mingrui Chen, Hongmin Liu, and Ran He. 2024. Rmt: Retentive networks meet vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5641–5651.

Meng Feiyu. 2024. Cfpsg: Collaborative filtering poi similarity graph enhanced retentive network for next poi recommendation. In *2024 21st International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 1–4. IEEE.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Albert Gu, Karan Goel, and Christopher Ré. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Chen Guo, Xinran Li, Jiaman Ma, Yimeng Li, Yuefan Liu, Haiying Qi, Li Zhang, and Yuhan Jin. 2024. Vl-mfer: A vision-language multimodal pretrained model with multiway-fuzzy-experts bidirectional retention network. *IEEE Transactions on Fuzzy Systems*.

Ayse Asu Guvenli and Bora Isildak. 2024. B-jet tagging with retentive networks: A novel approach and comparative study. *arXiv preprint arXiv:2412.08134*.

Jia Han, Zhiyong Yu, and Jian Yang. 2025. Radio frequency-retentive network for automatic modulation classification. *Electronics Letters*, 61(1):e70203.

Zhu Han, Shuyi Xu, Lianru Gao, Zhi Li, and Bing Zhang. 2024. Gretnet: Gaussian retentive network for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*.

Ali Hatamizadeh, Michael Ranzinger, and Jan Kautz. 2023a. Vir: Vision retention networks. *arXiv. org*.

Ali Hatamizadeh, Michael Ranzinger, Shiyi Lan, Jose M Alvarez, Sanja Fidler, and Jan Kautz. 2023b. Vir: Towards efficient vision retention backbones. *arXiv preprint arXiv:2310.19731*.

Wei He, Kai Han, Yehui Tang, Chengcheng Wang, Yujie Yang, Tianyu Guo, and Yunhe Wang. 2024. Densemamba: State space models with dense hidden connection for efficient large language models. *arXiv preprint arXiv:2403.00818*.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jia Cheng Hu, Roberto Cavicchioli, and Alessandro Capotondi. 2024a. Shifted window fourier transform and retention for image captioning. *arXiv preprint arXiv:2408.13963*.

Jingjing Hu, Dan Guo, Kun Li, Zhan Si, Xun Yang, and Meng Wang. 2024b. Maskable retentive network for video moment retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1476–1485.

Jianan Huang, Xuebing Liu, Qing Zhu, Yaonan Wang, Mingtao Feng, Jiaming Zhou, Zhen Zhou, Lin Chen, and Danwei Wang. 2025a. Rampgrasp: Retentive attention-based multiscale perception grasp detection network. *IEEE Transactions on Circuits and Systems for Video Technology*.

Jingjia Huang, Jingyan Tu, Ge Meng, Yingying Wang, Yuhang Dong, Xiaotong Tu, Xinghao Ding, and Yue Huang. 2024a. Efficient perceiving local details via adaptive spatial-frequency information integration for multi-focus image fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9350–9359.

Kai-Wei Huang and Chia-Ping Chen. 2024. Long audio file speaker diarization with feasible end-to-end models. In *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1–6. IEEE.

Qihe Huang, Zhengyang Zhou, Kuo Yang, Gengyu Lin, Zhongchao Yi, and Yang Wang. 2024b. Leret: Language-empowered retentive network for time series forecasting. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*.

Yuyao Huang, Kai Chen, Wei Tian, and Lu Xiong. 2025b. Boost query-centric network efficiency for multi-agent motion forecasting. *IEEE Robotics and Automation Letters*.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR.

Sangwon Kim, In-su Jang, and Byoung Chul Ko. 2024. Domain-free fire detection using the spatial–temporal attention transform of the yolo backbone. *Pattern Analysis and Applications*, 27(2):45.

Oliver Knitter. 2024. *Exploration of Quantum-Inspired Deep Learning Architectures for Fundamental Applications in Scientific Computing*. Ph.D. thesis.

Oliver Knitter, Dan Zhao, James Stokes, Martin Ganahl, Stefan Leichenauer, and Shravan Veerapaneni. 2024. Retentive neural quantum states: Efficient ansätze for ab initio quantum chemistry. *Machine Learning: Science and Technology*.

Bin Lei, Caiwen Ding, and 1 others. 2023. Flashvideo: A framework for swift inference in text-to-video generation. *arXiv preprint arXiv:2401.00869*.

Sihan Li and Juhua Huang. 2025. Resgdanet: An efficient residual group attention neural network for medical image classification. *Applied Sciences*, 15(5):2693.

Xing Li and Yuequan Bao. Adaptive gated meta graph retention network: A model for urban traffic flow prediction. *Available at SSRN 5170149*.

Zhiyuan Li, Yi Chang, and Yuan Wu. 2025. Segret: An efficient design for semantic segmentation with retentive network. *arXiv preprint arXiv:2502.14014*.

Zhiyuan Li, Tingyu Xia, Yi Chang, and Yuan Wu. 2024. A survey of rwkv. *arXiv preprint arXiv:2412.14847*.

Di Liang and Xiaofei Li. 2024. Ls-eend: Long-form streaming end-to-end neural diarization with online attractor extraction. *arXiv preprint arXiv:2410.06670*.

Huangbin Lin, Qing Zhu, Zhen Zhou, Yaonan Wang, Yongjie Sui, Yijiang Li, Tianming Li, and Zichen Chen. 2024. Single-stage uav geolocation enhancement with masa and self-homologous loss. In *2024 China Automation Congress (CAC)*, pages 5439–5444. IEEE.

Ruixing Lin and Shunmei Piao. 2024. Change detection of building remote sensing images based on rmt-bit. In *2024 4th International Conference on Electronic Information Engineering and Computer Science (EIECS)*, pages 428–432. IEEE.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of transformers. *AI open*, 3:111–132.

Zhicheng Lin, Rongpu Cui, Limiao Ning, and Jian Peng. 2025a. Temporal features-fused vision retentive network for echocardiography image segmentation. *Sensors*, 25(6):1909.

Zhijie Lin, Zilong Zhu, Lingling Guo, Jingjing Chen, and Jiyi Wu. 2025b. Disease detection algorithm for tea health protection based on improved real-time detection transformer. *Applied Sciences (2076-3417)*, 15(4).

Zhou Linhao, Zhong Shenghua, and Xiao Zhijiao. 2024. Discovering multi-relational integration for knowledge tracing with retentive networks. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 960–968.

Jiaji Liu, ZhiTao Liu, YePeng Wang, and Fang Li. 2024a. Vvnet: Underwater object detection network based on vision transformer and vision retnet for underwater robot picking. In *2024 International Symposium on Digital Home (ISDH)*, pages 19–24. IEEE.

Jiayuan Liu, Bo Zhou, Xue Wan, Yan Pan, Zicong Li, and Yuanbin Shao. 2025. Mar-vo: A match-and-refine framework for uav's monocular visual odometry in planetary environments. *IEEE Transactions on Geoscience and Remote Sensing*.

Jing Liu, Donglin Jing, Yanyan Cao, Ying Wang, Chaoping Guo, Peijun Shi, and Haijing Zhang. 2024b. Lightweight progressive fusion calibration network for rotated object detection in remote sensing images. *Electronics*, 13(16):3172.

Zhe Liu, Jinghua Hou, Xinyu Wang, Xiaoqing Ye, Jingdong Wang, Hengshuang Zhao, and Xiang Bai. 2024c. Lion: Linear group rnn for 3d object detection in point clouds. *Advances in Neural Information Processing Systems*, 37:13601–13626.

Ziyu Liu, Yi Shen, Yunliang Jiang, Hancan Zhu, Hailong Hu, Yanlei Kang, Ming Chen, and Zhong Li. 2024d. Variation and evolution analysis of sars-cov-2 using self-game sequence optimization. *Frontiers in Microbiology*, 15:1485748.

Baichao Long, Wang Zhu, and Jianli Xiao. 2024. St-retnet: A long-term spatial-temporal traffic flow prediction method. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 3–16. Springer.

Junwei Luo, Jiaojiao Wang, Jingjing Wei, Chaokun Yan, and Huimin Luo. 2025. Deephapnet: a haplotype assembly method based on retnet and deep spectral clustering. *Briefings in Bioinformatics*, 26(1):bbae656.

Omayma Mahjoub, Sasha Abramowitz, Ruan de Kock, Wiem Khlifi, Simon du Toit, Jemma Daniel, Louay Ben Nessir, Louise Beyers, Claude Formanek, Liam Clark, and Arnu Pretorius. 2025. Sable: a performant, efficient and scalable sequence model for marl. *Preprint*, arXiv:2410.01706.

Runyu Miao, Danlin Liu, Liyun Mao, Xingyu Chen, Leihao Zhang, Zhen Yuan, Shanshan Shi, Honglin Li, and Shiliang Li. 2024. Gr-p k a: a message-passing neural network with retention mechanism for p k a prediction. *Briefings in Bioinformatics*, 25(5):bbae408.

Zhaoyi Min, Qianqian Xiao, Muhammad Abbas, and Duanjin Zhang. 2025. Retentive network-based time series anomaly detection in cyber-physical systems. *Engineering Applications of Artificial Intelligence*, 145:110215.

SIMONE MOSCO. 2023. Exploiting retentive networks in 3d lidar semantic segmentation.

Cheng Nian, Weiyi Zhang, Fasih Ud Din Farrukh, Liting Niu, Dapeng Jiang, Fei Chen, and Chun Zhang. 2024. A 77.79 gops/w retentive network fpga inference accelerator with optimized workload. In *IECON 2024-50th Annual Conference of the IEEE Industrial Electronics Society*, pages 1–7. IEEE.

Masashi Okada, Mayumi Komatsu, and Tadahiro Taniguchi. 2024. A contact model based on denoising diffusion to learn variable impedance control for contact-rich manipulation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7286–7293. IEEE.

Sidike Paheding, Nusrat Zahan, Abel A Reyes, Ernesto Martinez Jr, and Eung-Joo Lee. 2024. Hyperspectral image classification with retentive network. In *Pattern Recognition and Tracking XXXV*, volume 13040, pages 15–21. SPIE.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, and 1 others. 2023. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*.

Lihong Peng, Xin Liu, Min Chen, Wen Liao, Jiale Mao, and Liqian Zhou. 2024. Mgndti: A drug-target interaction prediction framework based on multimodal representation learning and the gating mechanism. *Journal of Chemical Information and Modeling*, 64(16):6684–6698.

Zhen Qin, Songlin Yang, and Yiran Zhong. 2023. Hierarchically gated recurrent neural network for sequence modeling. *Advances in Neural Information Processing Systems*, 36:33202–33221.

Yufan Qiu, Yaping Liu, and Shuo Zhang. 2024. Rnete: A retentive network-based encryption traffic encoder. In *Proceedings of the 2024 3rd International Conference on Cryptography, Network Security and Communication Technology*, pages 214–219.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. 2017. Swish: a self-gated activation function. *arXiv preprint arXiv:1710.05941*, 7(1):5.

Hojjat Salehinejad, Sharan Sankar, Joseph Barfett, Errol Colak, and Shahrokh Valaee. 2017. Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Ruilin Su, Wanshan Zhang, and Dunhui Yu. 2024. Sequence-to-sequence multi-hop knowledge reasoning based on retentive network. In *2024 7th International Conference on Computer Information Science and Application Technology (CISAT)*, pages 360–366. IEEE.

Yongjie Sui, Yan Zheng, Qing Zhu, Zhen Zhou, Lin Chen, Jianqiao Luo, Yaonan Wang, and Zihao Yang. 2024. An efficient transformer incorporating a spatial decay matrix and attention decomposition for image matching. In *2024 China Automation Congress (CAC)*, pages 5216–5221. IEEE.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. 2023. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. 2022. A length-extrapolatable transformer. *arXiv preprint arXiv:2212.10554*.

Yutao Sun, Li Dong, Yi Zhu, Shaohan Huang, Wenhui Wang, Shuming Ma, Quanlu Zhang, Jianyong Wang, and Furu Wei. 2024. You only cache once: Decoder-decoder architectures for language models. *Advances in Neural Information Processing Systems*, 37:7339–7361.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Baoquan Wang, Yan Zeng, and Dongming Feng. 2025a. Deep learning-based damage assessment of hinge joints for multi-girder bridges utilizing vehicle-induced bridge responses. *Engineering Structures*, 333:120148.

Bin Wang, Fei Deng, and Peifan Jiang. 2024a. Eegdir: Electroencephalogram denoising network for temporal information storage and global modeling through retentive network. *Computers in Biology and Medicine*, 177:108626.

Junliang Wang, Wenlong Hang, Shuang Liang, Qiong Wang, Badong Chen, and Jing Qin. 2025b. Convolutional retentive network for eeg decoding. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Siyu Wang, Xiaocong Chen, and Lina Yao. 2024b. Retentive decision transformer with adaptive masking for reinforcement learning based recommendation systems. *arXiv preprint arXiv:2403.17634*.

Zeyu Wang, Libo Zhao, Jizheng Zhang, Rui Song, Haiyu Song, Jiana Meng, and Shidong Wang. 2025c. Multi-text guidance is important: Multi-modality image fusion via large generative vision-language model. *International Journal of Computer Vision*, pages 1–23.

Tengqing Wu. 2024. A diffusion data enhancement retentive model for sequential recommendation. In *2024 7th International Conference on Computer Information Science and Application Technology (CISAT)*, pages 114–118. IEEE.

Yuxin Wu and Kaiming He. 2018. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.

Yunyang Xie, Kai Chen, Shenghui Li, Bingqian Li, and Ning Zhang. 2024a. Uarc: Unsupervised anomalous traffic detection with improved u-shaped autoencoder and retnet based multi-clustering. In *International Conference on Information and Communications Security*, pages 187–207. Springer.

Zexun Xie, Min Xu, Shudong Zhang, and Lijuan Zhou. 2024b. Rcat: Retentive clip adapter tuning for improved video recognition. *Electronics*, 13(5):965.

Keyu Xu, Chengtian Song, Yue Xie, Lizhi Pan, Xiaozheng Gan, and Gao Huang. 2024. Rmt-yolov9s: An infrared small target detection method based on uav remote sensing images. *IEEE Geoscience and Remote Sensing Letters*.

Zhangyong Xu, Ning Chen, Guangqiang Li, Jing Li, Hongqing Zhu, and Zhiying Zhu. 2025. The mitigation of heterogeneity in temporal scale among different cortical regions for eeg emotion recognition. *Knowledge-Based Systems*, 309:112826.

Yimo Yan, Songyi Cui, Jiahui Liu, Yaping Zhao, Bodong Zhou, and Yong-Hong Kuo. 2025. Multimodal fusion for large-scale traffic prediction with heterogeneous retentive networks. *Information Fusion*, 114:102695.

Wenkui Yang, Zhida Zhang, Xiaoqiang Zhou, Junxian Duan, and Jie Cao. 2024a. Tt-df: A large-scale diffusion-based dataset and benchmark for human body forgery detection. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 429–443. Springer.

Xuan Yang, Yunxuan Dong, Lina Yang, and Thomas Wu. 2024b. Short-term photovoltaic forecasting model for qualifying uncertainty during hazy weather. *arXiv preprint arXiv:2407.19663*.

Xuan Yang, Tao Peng, Haijia Bi, and Jiayu Han. 2024c. Span-level bidirectional retention scheme for aspect sentiment triplet extraction. *Information Processing & Management*, 61(5):103823.

Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270.

Yuansong Zeng, Jiancong Xie, Zhuoyi Wei, Yun Su, Ningyuan Shangguan, Shuangyu Yang, Chengyang Zhang, Wenbing Li, Jinbo Zhang, Nan Fang, and 1 others. 2024. Cellfm: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. *bioRxiv*, pages 2024–06.

Fanlong Zhang, Huanming Chen, Quan Chen, and Jianqi Liu. 2025a. Cloud software code generation via knowledge graphs and multi-modal learning.

Lingzhe Zhang, Zewen Xiao, and Huaiyuan Wang. 2024a. Transient stability assessment of power system based on time-adapative retnet. In *2024 3rd Asia Power and Electrical Technology Conference (APET)*, pages 391–396. IEEE.

Meng Zhang, Wenzhong Yang, Liejun Wang, Zhonghua Wu, and Danny Chen. 2025b. Hfa-net: hierarchical feature aggregation network for micro-expression recognition. *Complex & Intelligent Systems*, 11(3):1–20.

Tianning Zhang, Feng Liu, Yuming Yuan, Rui Su, Wanli Ouyang, and Lei Bai. 2024b. Fast information streaming handler (fish): A unified seismic neural network for single station real-time earthquake early warning. *arXiv preprint arXiv:2408.06629*.

Yuxuan Zhang, Zipeng Zhang, Weiwei Guo, Wei Chen, Zhaohai Liu, and Houguang Liu. 2025c. Lretunet: A u-net-based retentive network for single-channel speech enhancement. *Computer Speech & Language*, page 101798.

Zhikang Zhang, Zhongjie Zhu, Yongqiang Bai, Ming Wang, and Zhijing Yu. 2024c. Octree-retention fusion: A high-performance context model for point cloud geometry compression. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 1150–1154.

Zipeng Zhang, Zhencai Zhu, Bin Meng, Zheng Yang, Mingke Wu, Xinyu Cheng, Binhong Li, and Houguang Liu. 2025d. Intelligent coal gangue identification: A novel amplitude frequency sensitive neural network. *Expert Systems with Applications*, 274:126880.

Kaili Zheng, Feixiang Lu, Yihao Lv, Liangjun Zhang, Chenyi Guo, and Ji Wu. 2024. 3d human pose estimation via non-causal retentive networks. In *European Conference on Computer Vision*, pages 111–128. Springer.

Li Zhou, Dayang Wang, Yongshun Xu, Shuo Han, Bahareh Morovati, Shuyi Fan, and Hengyong Yu. 2024. Gradient guided co-retention feature pyramid network for ldct image denoising. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 153–163. Springer.

Wang Zhu, Baichao Long, and Jianli Xiao. 2024. Spatial-temporal retentive heterogeneous graph convolutional network for traffic flow prediction. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Zijie Zhu, Feng Ding, Chenglong Chu, and Fangming Zhong. 2025. Retention enhanced cross-modal attention for multi-hop vqa. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

A Audio

In recent years, the RetNet has demonstrated significant promise across various audio-related tasks, particularly due to its capacity to efficiently model long-range dependencies while maintaining linear computational complexity.

Huang and Chen (2024) advanced long-form speaker diarization by developing the RetNet-EEND framework, which replaces the Transformer encoder in the EEND-EDA model with a RetNet-based architecture. Liang and Li (2024) proposed the LS-EEND model, which replaces the conventional masked self-attention mechanism in the encoder with Retention, enabling the model to achieve linear time complexity and improved efficiency. RetNet has also demonstrated strong potential in the domain of speech enhancement. In the LRetUNet architecture, Zhang et al. (2025c) integrated RetNet with LSTM units to construct a time-frequency representation module specifically designed for single-channel speech enhancement.

B Others

RetNet has found versatile applications in multiple domains, leveraging its capability to efficiently model long-term dependencies and handle complex sequential data.

In multimodal representation learning, VL-MFER introduces a bidirectional RetNet (Bi-RetNet) that exploits both parallel and recursive forms of multiscale retention to fuse visual and language modalities (Guo et al., 2024).

In the educational domain, a customized Retentive Module extends the original RetNet with a multiscale retention layer, capturing not only the temporal dependencies between student interactions but also their forgetting patterns, thereby enhancing the precision of learning state modeling (Linhao et al., 2024).

RetNet has also proven effective for time-series forecasting. LeRet leverages a causal retention encoder alongside multiscale retention modules to enhance nonlinear feature extraction in repaired sequences (Huang et al., 2024b). In reinforcement learning-based recommender systems, RetNet substitutes traditional masked attention with a segmented multiscale retention scheme, significantly improving efficiency and robustness in long-range modeling (Wang et al., 2024b). For sequential recommendation, dual RetNet modules encode both item and user history, enriching the personalized representation space (Wu, 2024). CFPSG further incorporates RetNet into a unified framework to support next-POI prediction by capturing fine-grained temporal correlations (Feiyu, 2024).

RetNet's architectural flexibility makes it well-suited for modeling motion and control dynamics. In EQNet, RetNet serves as the core of a structured state-space model, improving encoding efficiency in multi-agent motion forecasting (Huang et al., 2025b). Similarly, the NC-RetNet introduces a non-causal retention mask to access both past and future frames within blocks, improving 3D human pose estimation while maintaining low latency (Zheng et al., 2024). In TT-DF, RetNet powers the motion-guided branch, balancing long-range dependency modeling and computational cost (Yang et al., 2024a). For robotic manipulation, it is employed to process time-series inputs from learned impedance control dynamics (Okada et al., 2024).

Language and reasoning tasks also benefit from RetNet's capabilities. In ASTE, a novel bidirectional retention scheme inspired by RetNet bridges sequential and syntactic modeling gaps, boosting sentiment triplet extraction performance (Yang et al., 2024c). For world modeling, RetNet supports observation, reward, and termination prediction within REM, and is extended via the POP mechanism to generate observation sequences in parallel during imagination (Cohen et al., 2024).

System-level advancements further highlight RetNet's efficiency. A high-throughput FPGA inference accelerator incorporates RetNet to maximize hardware utility via dual-mode structure and linear computation (Nian et al., 2024). In FlashVideo, RetNet functions as the decoder, with parallel retention for training and autoregressive decoding for inference (Lei et al., 2023). Sable introduces an encoder-decoder RetNet for MARL with cross-retention and dynamic state resetting to better capture long-term dependencies in online settings (Mahjoub et al., 2025). In software engineering, RetNet is adopted as the encoder for cloud software code generation from multimodal knowledge (Zhang et al., 2025a).

In the domain of network analysis, RetNet proves invaluable in both encrypted and anomalous traffic scenarios. RN-ETE extends RetNet

by incorporating a multi-resolution self-attention mechanism, enabling bidirectional retention within encrypted traffic encoding for more effective network traffic encryption and analysis (Qiu et al., 2024). On the other hand, UARC applies RetNet's retention-based reconstruction module to model long-term temporal patterns in network traffic, addressing the challenges of anomaly detection in traffic streams (Xie et al., 2024a).

In dynamic graph deep learning, Chang et al. (2024) proposed the Graph Retention Network (GRN) as a unified architecture for deep learning on dynamic graphs.