Expert Systems With Applications 263 (2025) 125658

Contents lists available at ScienceDirect

Expert Systems With Applications

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

LRCN: Layer-residual Co-Attention Networks for visual question answering

Dezhi Han [a,1], Jingya Shi [a,1], Jiahao Zhao [a,1], Huafeng Wu [b,*,2], Yachao Zhou [c], Ling-Huey Li [d], Muhammad Khurram Khan [e], Kuan-Ching Li [d,*,2]

[a] College of Information Engineering, Shanghai Maritime University, Shanghai, 201306, China
[b] Merchant Marine College, Shanghai Maritime University, Shanghai, 201306, China
[c] Shanghai Anheng Times Information Technology Co., Ltd., Shanghai, China
[d] Department of Computer Science and Information Engineering, Providence University, Taichung, Taiwan
[e] Center of Excellence in Information Assurance (CoEIA), King Saud University, Riyadh, Kingdom of Saudi Arabia

ARTICLE INFO

Dataset link: https://visualqa.org/terms.html

Keywords:
Visual question answering
Multimodal feature
Layer-residual mechanism
Transformer
Computer vision

ABSTRACT

Visual Question Answering (VQA) is a multimodal task requiring a collaborative understanding of fine-grained visual concepts and language semantics. The key to VQA research is constructing a framework for modeling inter and intramodal information interactions between intricate modalities. Due to the superior global view and the ability to capture the relationships within multimodal information, the Transformer and its variants have become the prime choice for VQA tasks. Despite this layer-by-layer architecture enabling answer reasoning by optimizing multimodal feature information, the information may still be lost when transferred from lower to higher layers. To solve such an issue, we propose a Layer-Residual Mechanism (LRM), a plug-and-play generic approach that can reduce the computation and memory overhead to almost negligible. By adding a residual straight-through line between adjacent layers that cascades the attention block in-depth, LRM mitigates the vanishing during information transfer, stabilizes training, and endeavors to overcome performance decline in VQA models at deeper layers. To verify the effectiveness and generality of the proposed LRM, we apply it to the Encoder–Decoder structure, the Pure-Stacking structure, and a specifically designed Co-Stacking structure that can simultaneously understand textual features and visual features called Layer-residual Co-Attention Networks (LRCN). Extensive ablation studies and comparative experiments on the benchmark VQA v2 and CLEVR datasets show that LRCN significantly outperforms the original architectures, demonstrating the superior effectiveness and compatibility of LRCN.

1. Introduction

In recent years, deep learning has been widely used in various fields (Han et al., 2023, 2021; Li et al., 2024, 2025) and has also enabled machines to effectively process Natural Language Processing (NLP) and Computer Vision (CV) tasks (Liang et al., 2023), also an increased interest in multimodal studies, such as image captioning (Cornia et al., 2020; Wang et al., 2022), visual grounding (Chen et al., 2023; Deng et al., 2021; Wu et al., 2024), and visual question answering (Nguyen et al., 2022; Sood et al., 2023). Transformer (Vaswani et al., 2017) and its variants (X-formers) (Li et al., 2024a; Lin et al., 2022) have achieved impressive performance in NLP and CV (Huang et al., 2022), where Carion et al. (2020) constructed an end-to-end detection framework based on CNN + Transformer in the CV field from the post-processing steps that rely on manual priors such as NMS and anchor generators. In the field of NLP, Devlin et al. (2019) proposed a BERT model based on Deep Bidirectional Transformers that shows new SOTA performance records in eleven different tasks. Due to the powerful cross-modal compatibility and effectiveness of Transformer and X-formers, more researchers introduce them to Vision-and-Language (VL) tasks. Proposed by Yu et al. (2019b), MCAN introduced the Transformer-base into the VQA task for the first time, learning deeper cross-modal interaction information through Modular Co-attention layers, and achieved the 2019 VQA challenge championship.

Multimodal learning bridges the gap between language and vision. VQA requires understanding the semantic information in text and images to obtain fine-grained correlations between questions and images

* Corresponding authors.
E-mail addresses: dzhan@shmtu.edu.cn (D. Han), 202230310118@stu.shmtu.edu.cn (J. Shi), 202130310194@stu.shmtu.edu.cn (J. Zhao), hfwu@shmtu.edu.cn (H. Wu), anna.zhou@dbappsecurity.com.cn (Y. Zhou), s1091858@gm.pu.edu.tw (L.-H. Li), mkhurram@ksu.edu.sa (M.K. Khan), kuancli@pu.edu.tw (K.-C. Li).
[1] Equally Contributed to this work.
[2] Equally Contributed to this work.

https://doi.org/10.1016/j.eswa.2024.125658
Received 23 July 2024; Received in revised form 28 October 2024; Accepted 28 October 2024
Available online 12 November 2024

for reasoning. Currently, there are two basic approaches to training VQA models: the pre-training approach and the end-to-end method.

The pre-training is a very extensive process, given that obtaining rich and fine-grained image-text pairs and then passing the constructed textual and visual features into the Transformer-based model by forced alignment. Alternatively, as proposed in Li et al. (2020), objects under the same semantics are used as Anchor Points for image and language alignment to significantly ease the learning of alignments. The rich fine-grained information obtained by the pre-training approaches allows the model parameters to no longer be initialized randomly. Subsequently, after initial parameter learning in a large corpus, the model is used for downstream tasks through fine-tuning, which is the advantage of this method. However, the pre-training approaches are costly regarding computer resources and time. A large amount of data is needed when designing a new model, but the cost limits their deployment ability and the speed of model update iterations. For this reason, researchers adhere to an alternative direction, namely the end-to-end training approach, by extracting visual and textual features using different deep network structures and then passing the features into the designed model for training. The end-to-end methods can achieve competitive results compared to pre-training ones and consume very few resources (Chen et al., 2024; Li et al., 2024). Due to the excellent performance of the Transformer in VL tasks, the majority of VQA tasks have gradually tended to use the X-former architecture.

Designing the attention mechanism based on the Transformer to obtain fine-grained text and image information is the current dominant end-to-end approach in VQA. Chen et al. (2024) proposed a Context-based Compact Transformer, which improved the reasoning ability of the model by using optimized text global context and spatial position context. Rahman et al. (2021) proposed an AOA module to replace Self-Attention (SA) and Guided-Attention (GA), which generates an information vector and an attention gate using the attention results and the current context. However, Yu et al. (2022) suggested that the success of the Transformer architecture stems not only from the computational power of the attention mechanism but also from the influential design architecture of the Transformer. Although attention mechanisms have been valued and applied in various fields, a large number of studies have shown that even if simple spatial pooling is used instead of attention models, Transformer-based models can still achieve good results (Li et al., 2023; Liang et al., 2022; Long et al., 2023). Unlike the above approaches, to build a model that performs better in VQA tasks, we optimize the multimodal features of information transfer paths between adjacent layers and redesign the pure-stacking structure in this work.

In the research that investigates the effectiveness of X-formers, Chefer et al. (2021) concluded that there are three main types of Transformer architectures in multimodal tasks: (i) pure self-attention, (ii) self-attention combined with co-attention, and (iii) Encoder–Decoder attention. Most current VQA tasks are in the form (iii) or variants, as they propose a more efficient attention mechanism that optimizes the inference process for multimodal feature information We cannot ignore that textual information in the model can only be inferred from multiple consecutive self-attention, inevitably causing the network to focus on words irrelevant to the localized image region; as a result, leading to the generation of irrelevant word attention weights. The form (ii) (Cho et al., 2020; Tan & Bansal, 2019) notes the need for textual information and corrects for keywords by acquiring fine-grained image information; however, the effect of joint attention is diminished by multiple successive self-attention guidance that may result in both modalities noticing the wrong content from the start. Yu et al. (2019b) proposed a stacking structure in which language and image information are updated simultaneously at each layer. The downside of the method is that the model's inference of question keywords only refers to textual features, and the performance is far from the Encoder–Decoder structure. Based on the above-mentioned issues, we propose a Layer-residual Co-Stacking (LRCS) layer to synchronize the inference of textual and visual features promptly and to use the image information to guide the inference of question keywords, and the main contributions are as follows:

- To propose a novel Layer-Residual Mechanism (LRM) that does not increase the additional parameter size and computation cost that can optimize the multimodal information transfer route between the layers of the Encoder–Decoder structure and establish efficient identical mapping relationships. Also, it gives the model a robust inference learning capability and fine-grained feature information by extracting the underlying semantic information to the high layers in forward propagation and building straight backpropagation from the high layers to the lower levels. The Pure-Stacking structure has an attention dispersion problem due to the early SA unit feeding into the GA unit prematurely, which may compromise the stability of the model. The LRM is universal in addressing this issue, and it also allows the stacking structure to retain the advantage of simultaneous updating of textual and visual information and the stability of the encoder–decoder model.
- To propose a novel Layer-residual Co-Attention Network (LRCN) that applies LRM to Encoder–Decoder and Stacking structures, including Pure-Stacking and Co-Stacking. The proposed LRCS structure consists of LRCS layers cascaded in depth. By applying the Stacking structure, the model can send the multimodal information produced by SA to co-attention on time and later use co-attention to perform dynamic inference on both modal contents simultaneously. Compared to the Pure-Stacking structure, the Co-Stacking structure focuses on the need for text information to be guided by visual information, allowing the model to acquire question keywords corresponding to the actual visual content.
- Our evaluation of the proposed LRCN model on the VQA v2 and CLEVR dataset reveals a significant performance improvement compared to other state-of-the-art approaches. With its unique features, this novel model fully demonstrates its validity and rationality, sparking intrigue in its potential.

The remainder of this work is organized as follows: Section 2 reviews and presents existing achievements in VQA, while Section 3 depicts the methodology and details of the proposed scheme. Section 4 analyzes and evaluates the experimental results achieved, and finally, concluding remarks and future directions of this work are given in Section 5.

2. Related work

This section provides a comprehensive review of the existing work related to VQA and introduces the various residual approaches and normalization in the Transformer architecture.

2.1. Visual question answering

VQA task aims to answer a human question in natural language for a given image. To continue the development in sync with the VQA task, learning fine-grained multimodal representations is crucial (Chen et al., 2024; Han et al., 2020; Li et al., 2024b). The standardized VQA model is divided into three main modules:

1. Multimodal feature representation: The extraction of image features has evolved from the early VGG networks to the powerful ResNet family and further to the Bottom-Up and Top-Down Attention (BUTD) (Anderson et al., 2018) network based on Faster R-CNN. BUTD leverages a Faster R-CNN pre-trained on the Visual Genome dataset to extract key object region features, significantly enhancing performance in image captioning tasks. Compared to traditional global feature extraction methods, BUTD better captures fine-grained visual information, especially excelling in aligning image and text semantics. Jiang

et al. (2020) proposed image grid features, which contain valuable background information for downstream in addition to the salient targets. As this method can achieve better results, more researchers have adopted it. The LSTM network (Hochreiter & Schmidhuber, 1997) is usually used for text feature representations, the pre-trained Glove network (Pennington et al., 2014) or the Bert network (Devlin et al., 2019) used in question word embedding for better semantic extraction performance.

2. Multimodal feature filtering through the attention mechanism: The attention mechanism allows the model to dynamically focus on or assign different weights to different parts of the image when processing images and questions. Researchers have recently proposed improvements based on the attention mechanism (Chen et al., 2022, 2024; Guo et al., 2021; Rahman et al., 2021; Yang et al., 2022; Zhou et al., 2021). For example, Chen et al. (2022) introduced contextual information to VQA for the first time, proposing contextual attention to address potential comprehension biases. Moreover, our proposed model is based on the TRAR base (Zhou et al., 2021).
3. Multi-modality feature fusion: The model requires joint representation to uncover potential relationships between images and question modalities. Earlier, researchers introduced multimodal factorized high-order pooling (Yu et al. 2018) into VQA, enabling high-order interaction between modalities. Researchers have made many attempts (Li et al., 2022; Manmadhan & Kovoor, 2023; Nguyen et al., 2023; Rahman et al., 2021; Zhang et al., 2022). For example, Li et al. (2022) proposed a high-performance method for two-stage text-based multimodal fusion. Zhang et al. (2022) proposed multimodal fusion networks that took full advantage of multi-head attention and cross-attention and used MLP to generate multi-head attention distributions, which led to a further improvement in performance. Rahman et al. (2021) proposed two new fusion methods, Multimodal Attention Fusion, and Multimodal Mutan Fusion, to dynamically decide how to weigh each modality to generate the final features.

2.2. Residual connection and normalization in Transformer

Layer Normalization (LN) (Ba et al., 2016) and Residual Connection (RC) (He et al., 2016) are essential components of the Transformer with the ability to make deep network training stable and alleviate ill-posed gradients and model degeneration problems (Qiu et al., 2023; Xie et al., 2023). Ioffe and Szegedy (2015) were the first to propose Batch Normalization (BN) to unify batches of data under a consistent spatial data distribution. By discarding unimportant and complex information, BN reduces the complexity of training the data and the risk of overfitting. However, BN has the problem of losing differential information between samples due to forced normalization in the case of little correlation of information between batch samples or varying lengths of input sentences.

Later, Ba et al. (2016) introduced LN to perform normalization operations on the entire layer data to better constrain the vector's size to solve this problem. The LN pulls the data distribution into the non-saturated region of the activation function, which also has the feature of weight/data scaling invariance. The Residual Connection was introduced to prevent the problem of gradient disappearance due to the original information being forgotten during transfer. However, the LN operation makes the residual connection effect less valid. This architecture of adding the LN after the RC makes the training initial phase suffer from a slight gradient in the front layers and a large gradient in the later layers. Hence, to reduce the training difficulty, it is often necessary to use the warm-up mechanism to alleviate the gradient imbalance problem.

As opposed to this PostLN architecture, PreLN is the architecture in which the LN is placed before the RC. A comparison concluded by Xiong et al. (2020) that PreLN has better gradient descent capability, faster convergence, and robustness, though the effect performance is not as good as PostLN. Takase et al. (2022) proposed a method that provides high stability and effective training by simple modifications to Post-LN, and the results are superior to Pre-LN. For this reason, researchers have also proposed alternative solutions: Xie et al. (2023) proposed a new Transformer architecture with Pre-Post-LN (PPLN), which integrates the connections in Post-LN and Pre-LN to avoid the limitations of both and solves the gradient vanishing problem induced by Post-LN and the representation collapse problem caused by Pre-LN.

Although existing research on the normalization methods for Transformers has been prolific and has achieved promising results, there may still be some potential limitations in methods based on BN and LN when addressing tasks that require higher levels of understanding and reasoning. In the context of VQA tasks, the shortcomings of BN and LN in the normalization process prevent them from fully meeting the demands of multimodal information fusion and deep network training. The batch dependency of BN and the issue of information loss between samples lead to poor performance when handling heterogeneous data, making it challenging to effectively manage differences in information between batch samples. This can result in the loss of valuable distinctions within individual samples, further affecting the model's ability to capture details and perform reasoning. Although LN performs relatively well under uneven batch conditions, its weakening effect on residual connections and the issue of gradient imbalance still limit the model's performance. To address these issues, further exploration of improved normalization methods or the integration of other optimization techniques may be a promising direction for enhancing the training effectiveness of VQA task models.

3. Methodology

Fig. 1 is the overall flowchart of the proposed Layer-residual Co-Attention Network (LRCN). The model has three main components: Question and image feature representation, Layer-residual Co-Attention Learning, and Multimodal fusion and answer prediction. Two structures will be chosen in the Layer-residual Co-Attention Learning stage: Layer-residual Stacking and Layer-residual Encoder–Decoder.

3.1. Question and image feature representation

The kernel of the VQA is the acquisition of a joint representation of images and text. Image and text are different levels of information about dimensionality and structure. Modal embeddings are often used to address this modal disparity and better encode and fuse images with language, where features are extracted from each modality independently and then mapped to a shared feature space.

Following Zhou et al. (2021), we first read the questions Q from the VQA v2 dataset and tokenize them into word tokens using a tokenization tool to obtain question features. If the question length is less than 14, padding with 0, while questions exceeding 14 are truncated. Subsequently, these word tokens are transformed into vectors using 300-D GloVe word embeddings pre-trained on a large corpus, resulting in a dimensionality of $m \times 300$. Finally, a single-layer LSTM network with 512 hidden units is employed to obtain the question feature Y, $Y \epsilon \mathbb{R}^{m \times 512}$, m is the number of words.

For the image features, we utilize the ResNext152 (Jiang et al., 2020) visual backbone network, pre-trained on the Visual Genome dataset, to extract grid visual features. First, the input images are padded to 16×16. Then, we apply a convolution operation with a kernel size of 2×2 and a stride of 2 to aggregate the input image features, resulting in a final resolution of 8×8 for the grid visual features, which are represented as $X \epsilon \mathbb{R}^{n \times 2048}$, where n = 64. In practice, we employ a linear transformation to ensure that the dimensionality of X aligns with the question feature Y; thus, the dimensionality of the grid visual features is $X \epsilon \mathbb{R}^{n \times 512}$.

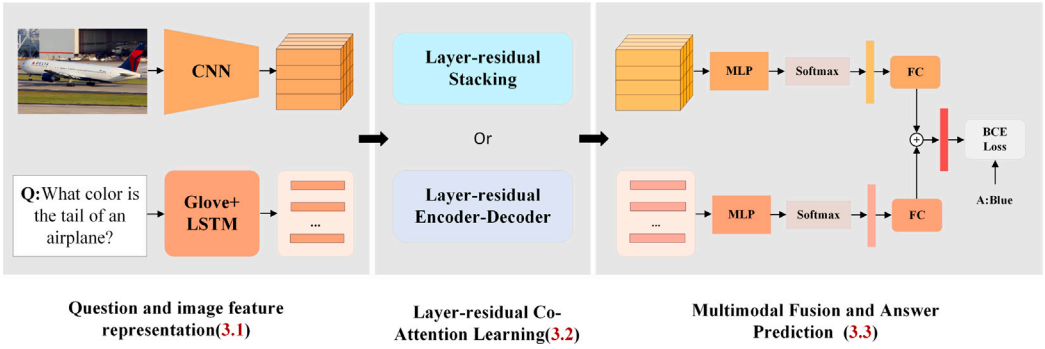

Fig. 1. Overall flowchart of the Layer-residual Co-Attention Network (LRCN).

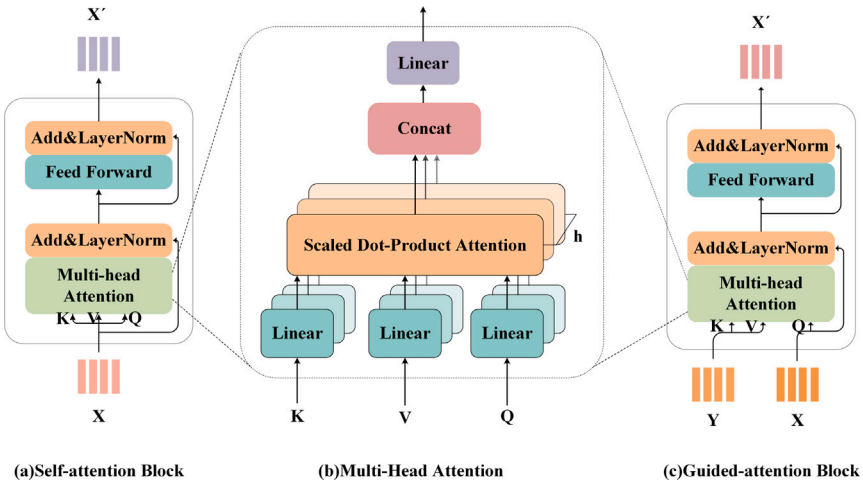

Fig. 2. Self-attention Block takes and the Guided-attention block.

3.2. Layer-residual co-attention learning

In the Layer-residual Co-Attention Learning stage, we have two alternative strategies for Layer-residual co-attention learning: Layer-residual stacking and Layer-residual encoder–decoder. The layer-residual stacking structures are divided into layer-residual pure-stacking and layer-residual co-stacking. Both structures consist of SA Block, GA Block, and Layer-residual mechanism.

3.2.1. Self-attention block and guided-attention block

SA Block focuses on intra-modal feature information, while GA Block obtains more accurate inter-modal feature information through the relationship between image and question. The purpose of the attention block is to sift through a large amount of information to find the essential features relevant to the task. The two attention blocks work in tandem to help the model focus on more critical image regions and question words, yielding fine-grained feature representations. Before presenting the attention Block, we introduce Multi-Head Attention (MHA), the core operator in the Block. For example, MHA is used to predict the state of charge of lithium-ion batteries, enhancing the recognition accuracy of hybrid networks through the multi-head attention mechanism (Li et al., 2024d). Similarly, in visual tracking tasks, the EMAT network achieves efficient feature fusion by optimizing the multi-head attention mechanism, thereby improving tracking accuracy (Wang et al., 2024). Compared to CNNs, MHA (see Fig. 2b) has a broader global view and can process information in parallel across different heads. Each head focuses on different information features, generating multiple independent subspaces that enhance the model's robustness and flexibility. Furthermore, compared to RNNs, MHA can simultaneously handle longer sequences, and due to its efficient parallel computing capabilities, MHA can globally fuse feature information, improving the model's representation ability and training efficiency for large-scale tasks.

We take the SA Block (see Fig. 2a) as an example. Given input $X = [u_1, \ldots, u_n] \epsilon \mathbb{R}^{n \times D}$ the input X is transformed into a query matrix $Q \epsilon \mathbb{R}^{n \times D_q}$, a key matrix $K \epsilon \mathbb{R}^{m \times D_k}$ and value matrix $V \epsilon \mathbb{R}^{m \times D_v}$. Specifically, the original query matrix, key matrix, and value matrix are projected into the same dimensional hidden-size projection matrix via fully connected layers for ease of operation, respectively:

$$Q = XW^Q, \; K = XW^K, \; V = XW^V \tag{1}$$

where $W^Q \epsilon \mathbb{R}^{D \times D_q}$, $W^K \epsilon \mathbb{R}^{D \times D_k}$, $W^V \epsilon \mathbb{R}^{D \times D_v}$ are trainable parameter matrices and $D_q = D_k = D_v$. The calculation formula of Self-attention (Vaswani et al., 2017) is defined as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{D_k}}\right) \tag{2}$$

where the softmax(.) function converts the similarity scores between queries and keys into a probability distribution, a process that ensures the stability of gradient computation and the robustness of numerical calculations and effectively captures the correlation of input features while suppressing interference from irrelevant features. This significantly improves the model's performance in handling complex inputs. D_k represents the scaling factor, which is used to prevent the values of QK^T from becoming too large, as this could cause the gradients in the softmax function to become very small, thus negatively impacting the model's training process.

Additionally, the primary reason we adopt MHA is that multiple heads operating in parallel can enhance the representative capacity of the attended features. To obtain different subspace of feature representations and improve the efficiency of operations, the query matrix

Q, key matrix K, and value matrix V entering the MHA are divided into h sub-matrices of the same size. Therefore, the input in the guide-attention of each head is:

$$Q^i = QW_i^Q,\ K^i = KW_i^K,\ V^i = VW_i^V \tag{3}$$

where $i \epsilon [1, 2, \ldots, h]$ is the linear subspace, $W_i^Q \epsilon \mathbb{R}^{D \times d_q}$, $W_i^K \epsilon \mathbb{R}^{D \times d_k}$, $W_i^V \epsilon \mathbb{R}^{D \times d_v}$ are the trainable mapping matrices for the ith head. Usually, $d_q = d_k = d_v = D/h$. After that, perform the h sub-matrices sets into h parallel heads containing independent scaled dot-product attention function, respectively:

$$\begin{cases} head_1 = Attention\left(Q^1, K^1, V^1\right) \\ \quad\vdots \\ head_i = Attention\left(Q^i, K^i, V^i\right) \\ \quad\vdots \\ head_h = Attention\left(Q^h, K^h, V^h\right) \end{cases} \tag{4}$$

The final results are obtained by concatenating the feature information calculated separately for the different subspaces and mapping them into the same matrix. Specifically, the multi-head attention mechanism learns diverse representations of input features from different subspaces through multiple parallel attention heads. Each head independently computes the similarity between Q, K, and V, capturing different levels of features or dependencies. The outputs of the heads are then concatenated and passed through a linear transformation to generate the final result:

$$MHA(Q, K, V) = Concat\left(head_1, head_2 ..., head_h\right) W^O \tag{5}$$

where $W^O \epsilon \mathbb{R}^{D_V \times D}$ is the projection learnable parameter matrix, usually with $D_V = D$. Overall, this mechanism enhances the model's representative capacity and improves its ability to capture complex features by attending to input data from multiple perspectives. Compared to single-head attention, it demonstrates more robust generalization and stability.

Self-attention and Guided-attention block consist of two sub-layers. The first is the MHA sub-layers, which apply residual connection (He et al., 2016) and layer normalization (Ba et al., 2016) after the input features performed the MHA operations to learn the relationships between the input features, and the second sub-layer is a fully connected Feed-Forward Network (FFN) module with a hidden layer. The equation is defined as follows:

$$X_l = LN\left(X_{l-1} + MHA(Q, K, V)\right) \tag{6}$$

$$FNN\left(X_L\right) = \sigma\left(X_L W_1 + b_1\right) W_2 + b_2 \tag{7}$$

where $X_l \epsilon \mathbb{R}^{n \times D}$ is the output feature of X, $LN(\cdot)$ represents layer normalization. σ denotes the ReLU function. The output features X_l of the MHA sub-layers are fed into FNN, and overfitting is prevented by the dropout (Srivastava et al., 2014) function. The model is eventually stabilized by $Add\&LN$, passing the features X_l as output to the next Block.

3.2.2. Layer-residual mechanism

Based on the theoretical foundation of related studies (He et al., 2020; Xiong et al., 2020), we introduce residual scores from the previous layer into each multi-head attention layer to mitigate overfitting, functioning similarly to regularization. This residual mechanism allows the model to incorporate information from the previous layer in each layer's feature representation, preventing excessive reliance on local features of the current layer and thereby reducing overfitting. The introduction of residual scores stabilizes gradient flow and reduces the freedom of independent learning at each layer, effectively preventing the model from falling into local minima during training. Furthermore, residual scores facilitate cross-layer feature sharing and reuse, enhancing the model's generalization ability. Next, we will explain the LRM in detail.

The Layer-residual Block uses a Post-LN style attention Block as the backbone. Since the two attention blocks have different attention directions (Section 3.2.1), we add skip edges to connect the same type of attention blocks to achieve the effect of the residual connection between layers, as shown in Fig. 3b. More formally, it adds PrevRe, the output X_l from the previous layer across a different type of attention block, as one additional input to the attention sub-layer in the current layer:

$$\begin{aligned} &X_l = LN\left(X_{l-1} + PrevRe + MHA\left(Q^{l-1}, K^{l-1}, V^{l-1}\right)\right) \\ &where\ PrevRe = X_{SA'l-1} \\ &or\ PrevRe = X_{GA'l-1} \end{aligned} \tag{8}$$

The LRM extracts rich feature information from the low layer to alleviate the diffuse phenomenon in the information transmission process, making the high layer output more fine-grained. Specifically, we add a direct residual connection between the output and input of each attention block. This connection does not involve any additional transformation; it simply passes the input data to the output of the next layer through addition. This design effectively preserves information from lower layers, ensuring that critical multimodal features are retained and not weakened as they propagate through deeper layers of the network.

As shown in Fig. 4a, we apply the Layer-residual Mechanism (LRM) in the SA unit of the MHA at layer l. Subsequently, by adding a skip edge, the features from the SA unit at layer $l-1$ are transferred to the SA unit at layer l. Based on Eq. (8), the LRM in the SA Block is defined as:

$$\begin{aligned} &X_{SA'l} = LN\left(MHA(Q, K, V) + X_{GA'l-1} + PrevRe\right) \\ &where\ Q = X_{GA'l-1}W_Q^{l-1},\ K = X_{GA'l-1}W_K^{l-1},\ V = X_{GA'l-1}W_V^{l-1}, \\ &PrevRe = X_{SA'l-1} \end{aligned} \tag{9}$$

where $W_Q^{l-1}, W_K^{l-1}, W_V^{l-1}$ are three different learnable parameter matrices.

As shown in Fig. 4b, we apply the Layer-Residual Mechanism (LRM) in the GA unit of the MHA at layer l. Subsequently, by adding a skip edge, the features from the GA unit at layer $l-1$ are transferred to the GA unit at layer l. Based on Eq. (8), the LRM in the GA Block is defined as:

$$\begin{aligned} &X_{GA'l} = LN\left(MHA(Q, K, V) + PrevRe + X_{SA'l}\right) \\ &where\ Q = X_{SA'l}W_Q^{l-1},\ K = YW_K^{l-1},\ V = YW_V^{l-1}, \\ &PrevRe = X_{GA'l-1} \end{aligned} \tag{10}$$

where $W_Q^{l-1}, W_K^{l-1}, W_V^{l-1}$ are three different learnable parameter matrices. Y denotes the question feature.

Additionally, taking LRM in SA as an example, the implementation of the LRM in SA is illustrated in Algorithm 1:

Algorithm 1 Pseudocode of LRM

Require: given the X^l as the input of the l layer of SA block
Ensure: X^{l+2} with the $PrevRe$ residual connection added.
1: Compute and output the attention result of the l layer
$X^{l+1} = X^l + MHA(Q, K, V)$ where $Q = X^l W_Q, K = X^l W_K, V = X^l W_V$ are the three different learnable parameter matrix
2: Compute and output the attention result of the $l+1$ layer $X^{l+2} = X^{l+1} + MHA(Q, K, V) + PrevRe$, $where\ PrevRe = X^l$
where $Q = X^{l+1}W_Q, K = X^{l+1}W_K, V = X^{l+1}W_V$ are the three different learnable parameter matrix

3.2.3. Layer-residual co-attention module

We proposed LRM is used for the Encoder–Decoder and Stacking structures, where the Stacking structures include Pure-Stacking and Co-Stacking, which constitute the Layer-residual Co-Attention module, as shown in Fig. 5. The pure-Stacking model refers to a structure in which the output of question self-attention is progressively fed to the guided attention up to the Lth layer. In contrast, the Encoder–Decoder model applies the fine-grained output obtained after encoding

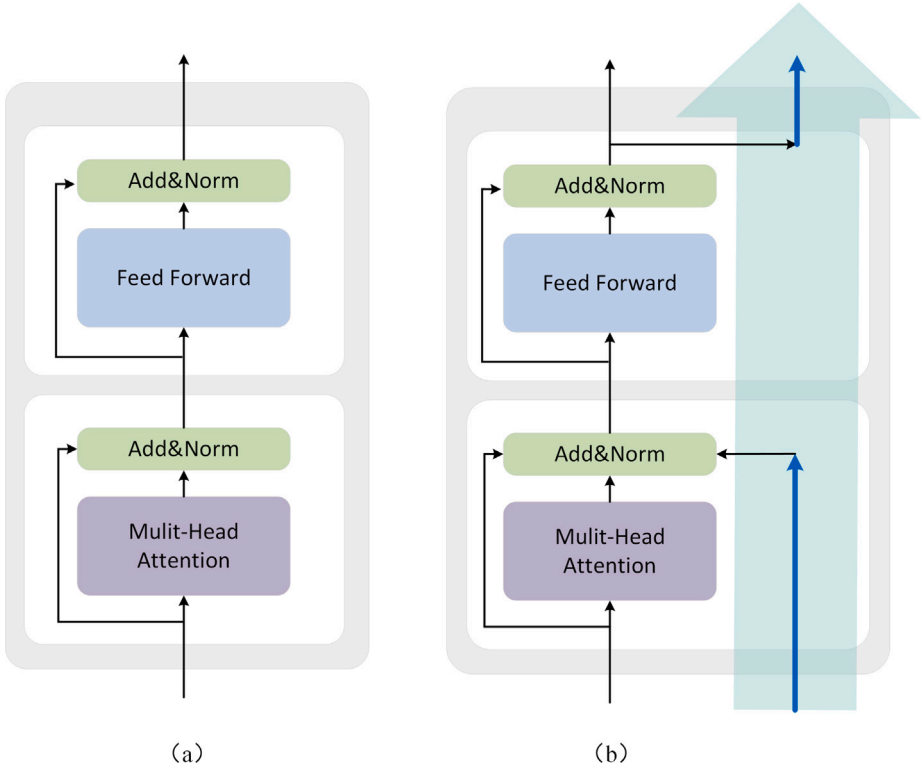

Fig. 3. (a) is the original Attention Block with Post-Ln style. (b) shows a Layer-residual Block. Our LRM creates a 'direct' path to propagate feature information under the same type of attention Block.

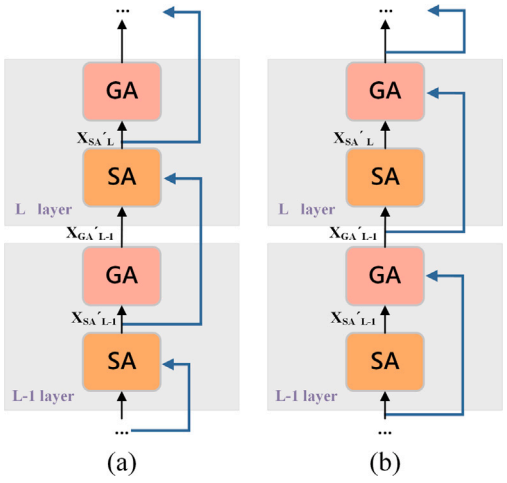

Fig. 4. (a) shows the structure when the LRM is used in the SA Block. (b) shows the structure when the LRM is used in the GA Block. The blue line is the 'direct' path that our LRM creates to propagate feature information. The lines where the GA accepts external input features are omitted here for simplicity.

through L layers of question self-attention to the guided attention units. Generally speaking, the Encoder–Decoder model feeds the output from the final SA(Y) unit into the GA(X, Y) unit. In contrast, the pure-stacking model learns from earlier SA(Y) and feeds into the GA(X, Y) unit. However, the learned self-attention from earlier SA(Y) units is inaccurate, and directly feeding this output into the GA unit could compromise the guidance of image learning. Therefore, we investigate the performance of both models after incorporating a layer-residual mechanism. Furthermore, to make the visual features guide the extraction of question features and obtain more high-level inter-modal interaction information, we optimize the pure-Stacking structure and propose a Co-Stacking design with richer linguistic information. This architecture captures critical modal information in VQA tasks layer by layer by cross-utilizing self-attention and guided attention mechanisms, which allows the model to begin focusing on core content information even during low-level interactions, thereby avoiding issues of poor interaction in subsequent layers due to insufficient granularity of low-level details. Although this structure may overlook the potential redundancy between the two attention mechanisms, the completeness of modal interaction is a significant focus for some complex multi-modal tasks (Chen et al., 2023). Therefore, some degree of potential redundancy is permissible.

We take the abovementioned textual feature Y and visual feature X as input and construct the LRC module by cascading LRC layers in depth (denoted as $LRC^{(1)}, \ldots, LRC^{(L)}$). As shown in Figs. 5a and b, the output $X^{(k-1)}$, $Y^{(k-1)}$ and $PrevRe^{(k-1)}$ are the input of the $LRC^{(k)}$ of layer k:

$$\left[X^{(k)}, Y^{(k)}, PrevRe^{(k)}\right] = LRC^{(k)}\left(\left[X^{(k-1)}, Y^{(k-1)}, PrevRe^{(k-1)}\right]\right) \tag{11}$$

Then, the output features $X^{(k)}$, $Y^{(k)}$ and $PrevRe^{(k)}$ are used as inputs to $LRC^{(k+1)}$. Iterating through layer L as described above, the final output visual feature $X^{(L)}$ and textual feature $Y^{(L)}$ are obtained as the output of the entire Layer-residual Co-Attention module:

$$\left[X^{(L)}, Y^{(L)}\right] = LRC^{(L)}\left(\left[X^{(L-1)}, Y^{(L-1)}, PrevRe^{(L-1)}\right]\right) \tag{12}$$

For the two structures Layer-Residual Encoder–Decoder (Fig. 5a) and Layer-Residual pure-Stacking (Fig. 5b), we set $X^{(0)} = X$, $Y^{(0)} = Y$ and $PrevRe^{(0)} = X$ for the first layer $LRC^{(1)}$.

As shown in Fig. 5c, the proposed Layer-residual Co-Stacking structure adds GA Blocks for the input textual features to be guided by visual features. Hence, add two Layer-residual transmission paths to the

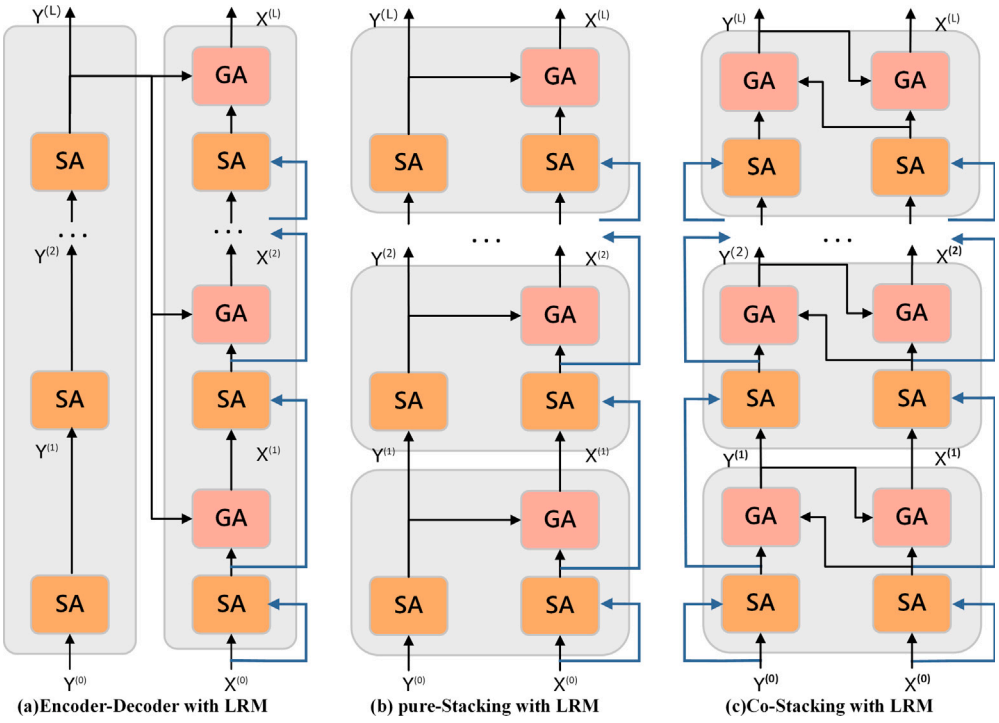

Fig. 5. The LRM is used for each of the three models consisting of cascaded LRC layers. The diagram depicts the LRM in SA Block, with the blue line indicating the LR transmission path.

model. Based on Eq. (11), the inter-layer recursive formula is defined as:

$$\begin{aligned}&\left[X^{(k)}, Y^{(k)}, PrevReX^{(k)}, PrevReY^{(k)}\right]\\&= LRC^{(k)}\left(\left[X^{(k-1)}, Y^{(k-1)}, PrevReX^{(k-1)}, PrevReY^{(k-1)}\right]\right)\end{aligned} \tag{13}$$

Based on Eq. (12), we can obtain the visual feature $X^{(L)}$ and textual feature $Y^{(L)}$ of $LRC^{(L)}$, the formula is as follows:

$$\left[X^{(L)}, Y^{(L)}\right] = LRC^{(L)}\left(\left[X^{(L-1)}, Y^{(L-1)}, PrevReX^{(L-1)}, PreReY^{(L-1)}\right]\right) \tag{14}$$

where $X^{(0)} = X$, $Y^{(0)} = Y$, $PrevReX^{(0)} = X$, and $PrevReY^{(0)} = Y$ for the first layer $LRC^{(1)}$.

3.3. Multimodal fusion and answer prediction

The core of multimodal fusion is to calculate the similarity or distance between images and questions. A widely adopted method is to map images and questions to a shared embedding space, calculate their similarity and answer predictions are made based on the proportion of similarity between the multimodal. The fine-grained visual features $X^{(L)}$ and textual features $Y^{(L)}$ output from the Layer-residual Co-Attention module are input to two Attention Pooling layers based on the weight summation approach. The attention weight formula is defined as:

$$\alpha = softmax\left(MLP\left(X^{(L)}\right)\right) \ and \ \beta = softmax\left(MLP\left(X^{(L)}\right)\right) \tag{15}$$

where $\alpha = \left[\alpha_1, \ldots, \alpha_m\right]$ refers to the visual attention weights and $\beta = \left[\beta_1, \ldots, \beta_n\right]$ contains the question attention weights, MLP is a two-layer nonlinear activation layer $(FC(D) - ReLU - Dropout(0.1) - FC(1))$. Afterward, the attention weights are multiplied by the corresponding features to obtain visual feature weights $\bar{X}\epsilon\mathbb{R}^D$ and textual feature weights $\bar{Y}\epsilon\mathbb{R}^D$, denoted as:

$$\bar{X} = \sum \alpha_i x_i^{(L)}, \ \bar{Y} = \sum \beta_j y_j^{(L)} \tag{16}$$

Next, the attended feature weights are fed into a linear fusion function to obtain the fused feature $Z\epsilon\mathbb{R}^{D_Z}$, and z is projected into the vector $f\epsilon\mathbb{R}^k$:

$$z = LayerNorm\left(\bar{X}W_x + \bar{Y}W_y\right), \ f = zW_z \tag{17}$$

where $W_x, W_y\epsilon\mathbb{R}^{D\times D_Z}$ and $W\epsilon\mathbb{R}^{D_Z\times k}$ are trainable parameter matrices. Next, we fed f into the nonlinear activation ReLU and sigmoid function to classify the answers. Finally, the Binary Cross-Entropy (BCE) (Teney et al., 2018) function is used as a loss function, and the formula is given as:

$$s = sigmoid\left(W^O ReLU\left(f\right)\right) \tag{18}$$

$$\xi = \sum_i^N \epsilon_i \log\left(s_i\right) + \left(1 - \epsilon_i\right)\log\left(1 - s_i\right) \tag{19}$$

Where s denotes the weight of the candidate's answer score, $\epsilon\epsilon\mathbb{R}^K$ denotes the ground-truth label and $\epsilon_i\epsilon[0,1]$ implies the match of the ith answer to the question. K corresponds to the size of the candidate's answer vocabulary.

4. Experimental analysis

This section is divided into four parts for description: the baseline dataset for experiments and validation metrics for the models are presented in Section 4.1, details of the experiments and the hyper-parameter settings are given in Section 4.2, the validity and rationality of our proposed LRM and LRCN model are verified by ablation experiments in Section 4.3. We compare the performance of the proposed model with existing models in Section 4.4, and lastly, the advantages of the proposed model are visualized in Section 4.5.

4.1. Datasets and evaluation metrics

VQA v2 (Goyal et al., 2019) is the most widely used benchmark dataset for VQA tasks, which is more balanced in terms of linguistic bias, consists of images derived from the MS-COCO dataset (Lin et al.,

2014) and 2,042,721 manually annotated question–answer pairs associated with the images. Each image includes at least three questions, containing ten answers and confidence levels from different annotators. The dataset is further divided into train, val, and test subsets:

- train (82,783 images with 443,757 QA pairs for training)
- val (40,504 images with 214,354 QA pairs for validation)
- test (81,434 images with 447,793 QA pairs for testing)

Additionally, the test subset is divided into test-dev and test-std sets in a ratio of 1:3 and used to verify model validity in the VQA challenge online. We use a standard accuracy-based voting mechanism to predict model response quality with the following formula:

$$Accuracy\left(a_i\right) = \min\left(\frac{\sum_i \prod (a = i)}{3}, 1\right) \tag{20}$$

where $a_1, \ldots, a_j$ represents answers from different annotators. The final accuracy results include three question types (Yes/No, Number, and Other) and Overall accuracy.

CLEVR (Johnson et al., 2017) is a diagnostic dataset for visual reasoning designed to better understand the visual reasoning capabilities of VQA systems. This dataset encompasses metrics for counting, comparison, and logical reasoning, comprising a total of 100,000 images and one million questions, of which 853,000 are unique. It is primarily divided into three subsets:

- train (70,000 images, 700,000 questions)
- validation (15,000 images, 150,000 questions)
- test (15,000 images, 150,000 questions)

4.2. Implementation details

Generally, we set the number of multi-heads h=8; each head has a dimension $D_h = D/h = 64$, and the batch size is set to 64. Following the strategy (Zhou et al., 2021), the size of the candidate answer set is 3129. To further verify the ability of the proposed LRM to deepen the model, we set the number of layers $L = [6, 8]$. For fairness in experimental comparisons, the Adam optimizer (Kingma & Ba, 2014) is used when training the models,

For VQA v2, the number of the question words is $m = 14$, and the number of the region features is $n = 64$; Hyper-parameters β_1, β_2 are set to 0.9 and 0.98; the learning rate is set to a minimum value of (2.5te-5, 1e-4), where t represents the current epoch, and the learning rate decays by 1/5 at the 10th and 12th epochs. In particular, we use a warm-up mechanism for the first 3 epochs to prevent the previous layers from being challenging to train due to the small learning rate. The models are trained for a total of 13 epochs, and to visually compare the performance of the different structural models, we test them on the test-std or test-std sets. In addition, applying a training set containing a train subset, a val subset, and an additional vg subset as an augmented dataset to complete training and the question–answer pairs of the vg subset are obtained from Visual Genome Krishna et al. (2017).

For CLEVR, the number of the question words is $m = 43$, and the number of grid feature n is set to 224. Similarly, the grid features are transformed into 512-D through a fully connected layer. The models are trained for a total of 16 epochs.

Additionally, we list the experimental environment settings to provide more detailed information, as shown in Table 1.

4.3. Ablation study

In this subsection, we depict the effectiveness of the LRCN through a series of ablation studies in the VQA v2 and CLEVR datasets. To ensure the experimental results, all models are trained on the train+val+vg subset and tested on the test-dev or test-std subset in the VQA v2, and on the train subset in the CLEVR.

Table 1
Experimental environment.

Hardware	
CPU	Montage Jintide(R) C6226R, 2.9GHZ
GPU	NVIDIA GeForce RTX 3090
SSD	1TB
RAM	128GB
Software	
System	Linux 5.15.0
Language	Python3.9, PyTorch2.1.2, CUDA 12.0

4.3.1. Ablation on the VQA v2

Layer-residual block in different LRCN variants:

In Fig. 5, we have designed three variants of LRCN on three distinct model stacking architectures, from left to right denoted as $LRCN_{ED-SA}$, $LRCN_{St-SA}$, and $LRCN_{CSt-SA}$, and the stacking structures employed are Encoder–Decoder, pure-Stacking, and Co-Stacking, respectively. Notably, these three variants act on the Layer residual SA Block. As a control group, Encoder–Decoder, pure-Stacking, and Co-Stacking structures are stacked by SA Block and GA Block, with the same architecture as LRCN without Blue line. Specifically, Encode-Decoder originates from our reference model TRAR base (Zhou et al., 2021), and subsequent experiments use the same visual feature extraction approach. Experimental results in Table 2 show that all three base models are largely improved after using the LRM. Even though the overall accuracy of Encoder–Decoder is already more effective, there is a 0.38% improvement. Furthermore, it can be seen that the LRM significantly affects the two Stacking style model enhancements, allowing them both to outperform TRAR.

Our proposed Layer-residual Co-Stacking structure achieves an increase of 0.83% in the overall accuracy result. Employing LRM allows both Layer-residual structures to have similar results, rather than one structure being significantly outperformed. The experimental results demonstrate the generality of LRM to VQA models, as the addition of textual co-attention is updated to the Co-Stacking structure, allowing the textual attention to match the keywords of the image information and improving the learning ability of the Stacking structure for multimodal tasks. Yu et al. (2019b) argued that the Stacking structure is less effective than the Encoder–Decoder because the early feeding of textual features instead of the final textual features compromises the image co-attention. In comparison, LRM mitigates this problem by bridging the high-layer semantic information with the underlying layer, which is why the Stacking structure is substantially improved.

Layer-residual in SA Block vs. Layer-residual in GA Block:

To verify the different effects of Layer-residual in SA and GA Block (Fig. 5a,b) on the models. We designed three sets of control experiments, using two LR attention blocks for each of the three structures. The $LRCN_{ED-GA}$, $LRCN_{St-GA}$, and $LRCN_{CSt-GA}$ represent the application of the Layer-residual GA Block to Encoder–Decoder, Pure-Stacking, and Co-Stacking, respectively, with the $LRCN_{ED-SA}$, $LRCN_{St-SA}$, and $LRCN_{CSt-SA}$ as control groups (the subscript 'SAGA' indicates that both LR attention blocks are used). Ablation results in Table 3 show that the Encoder–Decoder is more likely to answer correctly on the Y/N questions, while the Stacking structure is more likely to perform better on the Number and Other questions. It is their structure that determines the performance of different types of questions. Without relying on visual feature information, the Encoder–Decoder structure completes the textual features' keyword refinement first, making it easier for the model to focus on the question keywords and ignore the context around the keywords. Moreover, we also find that LR in GA amplifies this model learning bias effect, and the overall accuracy of LR in SA will be relatively higher. This is because the

Table 2
Ablation studies of our proposed variants on the test-dev of VQA v2. The values in "()" indicate the accuracy increase difference between the additional LRM and the non-LRM.

Models	Y/N (%)	Num (%)	Other (%)	All (%)
pure-Stacking	87.27	53.85	61.22	71.12
$LRCN_{St-SA}$	87.57(+0.3)	53.61(−0.24)	61.99(+0.77)	71.58(+0.46)
Co-Stacking	86.8	52.86	61.39	70.89
$LRCN_{CSt-SA}$	87.31(+0.51)	**54.43**(+1.57)	**62.33**(+0.94)	71.72(+0.83)
Encoder–Decoder	87.43	53.80	61.81	71.45
$LRCN_{ED-SA}$	**87.74**(+0.31)	53.99(+0.19)	62.29(+0.48)	**71.83**(+0.38)

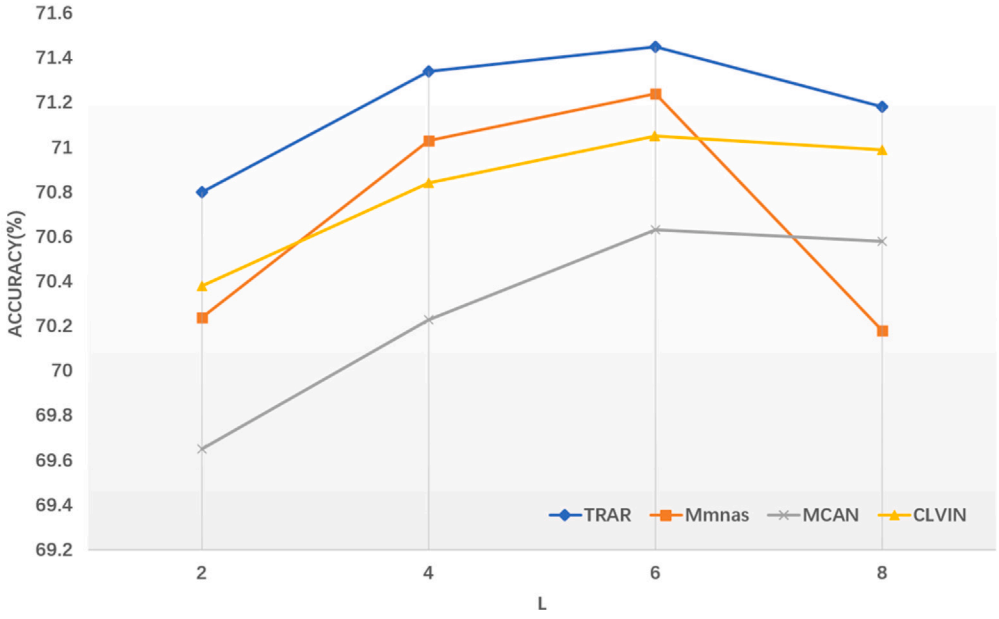

Fig. 6. Performance trend chart of some classic models under different layer depths.

Table 3
Ablation studies of our proposed Layer-residual Attention Block on the test-dev of VQA v2, where (layer) L = 6, (head) H = 8.

Models	Y/N (%)	Num (%)	Other (%)	All (%)
$LRCN_{St-GA}$	87.39	54.14	62.01	71.57
$LRCN_{St-SAGA}$	87.33	54.10	62.05	71.56
$LRCN_{St-SA}$	87.57	53.61	61.99	71.58
$LRCN_{CSt-GA}$	87.35	54.50	61.91	71.55
$LRCN_{CSt-SAGA}$	87.14	**54.65**	62.29	71.66
$LRCN_{CSt-SA}$	87.31	54.43	**62.33**	71.72
$LRCN_{ED-GA}$	**87.74**	53.52	62.00	71.64
$LRCN_{ED-SAGA}$	87.58	53.1	62.19	71.61
$LRCN_{ED-SA}$	**87.74**	53.99	62.29	**71.83**

Table 4
Ablation studies of deeper LRCNs (L) on the test-dev of VQA v2. The number of the head is 8.

Models	Y/N (%)	Num (%)	Other (%)	All (%)
$LRCN_{St-SA-6}$	87.57	53.61	61.99	71.58
$LRCN_{St-SA-8}$	87.68	**54.44**	61.94	71.69
$LRCN_{CSt-SA-6}$	87.31	54.43	**62.33**	71.72
$LRCN_{CSt-SA-8}$	87.47	**54.44**	62.20	71.73
$LRCN_{ED-SA-6}$	87.74	53.99	62.29	71.83
$LRCN_{ED-SA-8}$	**87.80**	54.21	62.28	**71.88**

attention properties of the two attention blocks (Section 3.2.1) are enhanced when LRM is applied. The GA unit involves a multimodal mixture of operations. Thus, LR in GA amplifies the model's ability to Co-attention. Besides, the SA unit only requires attention to the vital information of a single modal, and the residual information is closer to the original features $X^{(0)}$ and $Y^{(0)}$, making the system more stable and efficient in gradient backpropagation. However, there is no extra performance improvement from using both LR attention blocks simultaneously, possibly because the intensive low-layer extraction compromised the information inference capability of the features.

LRCN vs. Depth

Although traditional residual connection methods significantly improve gradient flow and mitigate the vanishing gradient problem in deep networks to some extent, they still fail to completely prevent the occurrence of gradient "shattering" as the number of layers increases. Specifically, in deep networks, after the input signal passes through multiple residual layers, the weights in the residual paths and the input signal accumulate increasing amounts of noise (primarily high-frequency or white noise), which leads to unstable gradients and blurred information transmission. For example, as shown in Fig. 6, in several VQA methods based on traditional residual connections, when the number of layers in the model architecture exceeds six, a noticeable decline in performance is observed, further validating this phenomenon.

Furthermore, to demonstrate that LRM can effectively alleviate this issue, Fig. 7 compares the performance results of $LRCN_{ED-SA}$ and TRAR under various types of questions across different layers.

As shown in Fig. 7, the performance of $LRCN_{ED-SA}$ is compared with the reference model TRAR for the different number of layers. $LRCN_{ED-SA}$ has an average accuracy improvement of 0.5% for the same number of layers except for the Number questions. In sync with

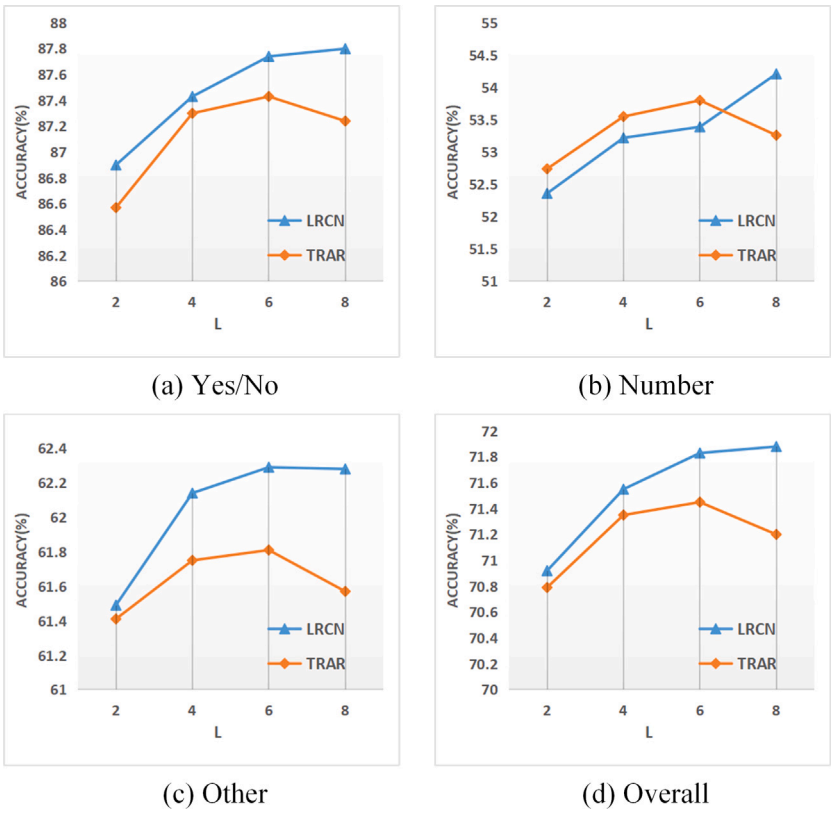

Fig. 7. The overall and per-type accuracies of the LRCNED-SA-L model compared with TRAR-L, where the number of layers. All the reported results are evaluated on the test-dev of VQA v2.

the increase of L, the gap continues to expand. Furthermore, most Transformer models (Guo et al., 2021; Rahman et al., 2021; Yu et al., 2019b; Zhou et al., 2021) fail to improve or even have degraded performance on the VQA task when the number of layers is more than 6. Compared to $L = 6$, some of the model's performance metrics at $L = 8$ are again improved, which shows that LRM mitigates the gradient fragmentation problem to some extent by reducing white noise. This is because it achieves more stable signal propagation by reducing the noise caused by the increasing number of layers. During forward propagation, the input signal can directly transfer the bottom feature information to the top layer through the LRM, thereby forming a constant mapping that effectively alleviates the network degradation problem.

Furthermore, the feedback signal can be passed directly into the bottom layer across the intermediate weight transformation matrix during backward propagation, preventing the attention dispersion problem. We further set up six sets of experiments to reduce the experimental period. The LR in SA is applied to the three LRCN variants to observe their detailed performances in L= [6, 8]. Table 4 shows that the accuracy of some metrics gradually increases with increasing L, which indicates the effectiveness of utilizing the LRM for deep models in the VQA v2 dataset.

LRCN vs Head(h)

Table 5 presents the ablation study results of $LRCN_{ED-SA}$ using different numbers of attention heads, showing that the number of heads has a significant impact on model performance. When the number of attention heads is less than 8, the average accuracy of the model is lower than that with H = 8. This is because the multi-head attention mechanism captures multi-level features and dependencies from different subspaces. However, when the number of attention heads exceeds 8, model performance does not improve but slightly decreases. This

Table 5
Ablation studies of $LRCN_{ED-SA}$ on the VQA v2 test-dev dataset with different number of heads (H) when L = 6.

heads (H)	Y/N (%)	Num (%)	Other (%)	All (%)
2	86.18	52.56	61.08	71.23
4	87.32	53.13	61.45	71.56
6	87.64	53.78	62.03	71.77
8	**87.74**	**53.99**	62.29	**71.83**
10	87.68	53.90	**62.33**	71.81

decline is attributed to the issues of information redundancy and dispersion caused by increasing the number of attention heads. To mitigate these negative effects, we introduced residual scores from the previous layer, which effectively enhanced model stability and generalization ability. Based on the comparative analysis of the experimental results, we finally set the number of attention heads to 8.

4.3.2. Ablation on the CLEVR

This section primarily discusses the ablation study of LRCN on the CLEVR dataset, as presented in Table 6. We investigate (1) the impact of layer residual blocks in different LRCN variants; (2) the effects of applying layer-residual mechanisms to the SA, GA, and SAGA blocks within the encoder–decoder structure; and (3) the influence of varying the number of layers while maintaining H = 8. St-SA denotes the use of a pure-stacking architecture with the application of a layer-residual mechanism in SA block, while CSt-SA indicates the use of a co-stacking architecture with the layer-residual mechanism in SA block.

Based on the analysis from Table 6, (1) when applying the layer-residual in SA block of different variants (rows 4–6), the performance of the LRCN model significantly surpasses the TRAR-Base, which demonstrates that our model is capable of executing reasoning tasks more effectively than TRAR-Base. Additionally, the performance of variant

Table 6
Ablation studies of LRCN on the CLEVR dataset.

Models	Overall	Count	Exist	Comp-Num	Query-Att	Comn-Att
TRAR-Base	98.54	96.34	99.24	98.60	99.43	98.93
$L = 6, H = 8$						
$LRCN_{St-SA}$	98.57	96.03	99.31	98.82	99.54	98.93
$LRCN_{CSt-SA}$	98.68	96.68	99.33	98.01	99.52	99.14
$LRCN_{ED-SA}$	98.63	96.38	99.33	99.01	99.51	99.14
$LRCN_{ED-GA}$	98.40	95.79	99.20	98.71	99.47	98.93
$LRCN_{ED-SAGA}$	98.53	96.07	99.25	98.89	99.49	99.02
$L = \{6, 8\}$, $H = 8$						
$LRCN_{CSt-SA-6}$	**98.68**	**96.68**	99.33	99.01	99.52	**99.14**
$LRCN_{CSt-SA-8}$	98.57	96.50	99.29	98.75	99.49	99.00
$LRCN_{ED-SA-6}$	98.63	96.38	99.33	99.01	99.51	**99.14**
$LRCN_{ED-SA-8}$	**98.65**	96.27	**99.35**	98.68	**99.55**	99.08

$LRCN_{St-SA}$ is inferior to $LRCN_{CSt-SA}$ and $LRCN_{ED-SA}$, further indicating that the Co-Stacking and Encoder–Decoder structures more effectively leverage crucial information from both text and images during the model's reasoning process. (2) Layer-Residual in SA block performs better compared to Layer-Residual in GA block, we hypothesize that the inference process of the model requires a more comprehensive integration of image information, whereas the GA block processes text and images simultaneously may introduce misleading image regions in the presence of textual information, which may in turn lead to the model in underperforming when dealing with noise. (3) We continue to explore the model's layer residual mechanism at a deeper level in the CLEVR dataset. As shown in the lower part of Table 6, as the number of layers L increases from 6 to 8, there is no significant degradation in the performance of $LRCN_{CSt-SA}$ and $LRCN_{ED-SA}$, while showing a slight upward trend on some metrics. This result illustrates that our proposed LRM in the CLEVR dataset is also able to mitigate the noise due to the increase in the number of layers to some extent. We hypothesize that despite the deepening of layers, the limited performance enhancement is primarily due to CLEVR primarily testing the model's understanding of structured visual information. Some studies (Shen et al., 2023, 2024; Yan et al., 2022) have indicated that a model's reasoning ability is closely related to the complexity of the structures present in the images. While LRCN's core design incorporates residual connections between layers to ensure more stable inter-layer information transmission, it lacks dedicated modeling of the local relationships among visual objects. Therefore, this may represent a limitation of the LRCN model. In contrast, LRCN outperforms TRAR-Base across all metrics, further demonstrating the effectiveness of LRM and its capacity to prevent the omission of critical information during the reasoning process.

4.3.3. Performance analysis

We discuss the computational cost of our model, including the parameters count (Params), computational complexity (FLOPs), the training time per epoch (Time), and overall accuracy, as detailed in Table 7. Based on the data in Table 7, we can draw the following conclusions: (1) Compared to the baseline model, our model achieves a consistently higher accuracy without introducing parameters or computational complexity. (2) $LRCN_{St-SA-6}$ and $LRCN_{ED-SA-6}$ have a slight advantage over models $LRCN_{CSt-SA-6}$ in terms of Params and FLOPs. This advantage may be attributed to the execution of GA in Co-Stacking. (3) our designed LRCN model does not introduce additional memory consumption, with an approximate runtime of 2900 s per epoch (training for a total of 13 epochs), while maintaining stable improvements in accuracy. This demonstrates the feasibility and rationality of LRCN.

4.4. Comparison with existing methods

To validate this work, we analyze the LRCN model against several state-of-the-art models on the VQA v2 and CLEVR datasets, which aims to demonstrate the validity and generalizability of our proposed model. The experimental results are shown in Tables 8 and 9, respectively. We compared the two LR attention Blocks (Section 4.3) and chose the more balanced LRs among the SA modules for the stacking, encoder–decoder and co-stacking structures.

Table 7
Comparison of Parameter count (Params), computational complexity (FLOPs) of the LRCN model on the VQA v2 dataset.

Models	Params (M)	FLOPs (G)	Time (s)	All (%)
baseline	51.61	3.67	2894	71.45
$LRCN_{St-SA-6}$	51.61	3.67	2956	71.58
$LRCN_{CSt-SA-6}$	65.71	5.67	3426	71.72
$LRCN_{ED-SA-6}$	51.61	3.67	2905	71.83

4.4.1. Comparison results on the VQA v2

Transformer architecture allows for the simultaneous processing of elements within an input sequence, and its self-attention mechanism enables the model to effectively capture dependencies in the input data. Consequently, Transformer and its variants have gradually become the prime choice for VQA tasks. Our LRCN model is also implemented based on the Transformer framework. To this end, we selected several representative Transformer-based VQA models for comparison, including MCAN, CLVIN, MMnasNet, TRAR, LAST-G, RRAF, MCAN-PA, SPCA-Net, VMAN, and RWSAN. Additionally, we included two classic non-Transformer methods, BUTD and MUAN, for comparison. The comparison results are presented in Table 8.

The BUTD (Anderson et al., 2018) model proposes a bottom-up and top-down attention approach used in subsequent mainstream models based on Fast-RCNN to extract visual features. MCAN (Yu et al., 2019b) and TRAR base (Zhou et al., 2021) share the same six-layer co-attention Transformer architecture but differ using different feature extraction methods. The MUAN (Yu et al., 2019a) model stacks ten layers of tensor-based Tucker decomposition blocks at depth to obtain intra- and inter-modal interactions. MUAN (Multi-Modal Attention Network) overcomes the limitations of BUTD and MCAN by simultaneously handling intra-modal relationships in both images and text, as well as performing effective cross-modal interactions, leading to more comprehensive feature extraction. Although MUAN's performance exceeds that of MCAN and BUTD, its complex network design may affect training efficiency. CLVIN (Chen et al., 2023) achieves a complete interaction of multimodal information based on MCAN, realizing a reasonable distribution of question word weight information. The MMnasNet (Yu et al., 2020) model adds task-specific heads to handle different modal tasks and uses a gradient-based NAS algorithm to search for a suitable structure for the task. RWSAN (Qin et al., 2023) proposes a Residual Weight Shared Attention Network by stacking residual weight shared attention layers and employing low-rank attention units for residual learning within each layer. VMAN (Song et al., 2024) enhances the attention mechanism by introducing a visually modified attention unit, which allows the model to acquire fine-grained query features and bolsters its reasoning capabilities. However, despite advancements made by both methods, they do not fully account for the overlooked low-level information compared to LRCN, and the computational complexity has also increased.

The LAST-G (Shen et al., 2023) model simultaneously models intra- and inter-window attention by setting up a local window of grid visual features. It is evident from the results that the LRCN outperforms the SOTA models in almost all types of accuracy. RRAF (Shen et al., 2024) focuses on modeling the relationships between visual objects in detail through position, appearance, and semantic features. This design makes RRAF particularly adept at reasoning about the relative positions or complex relationships of visual objects. In "Number" type questions, which often involve counting multiple objects and their relationships (e.g., "How many birds are in the picture?"), RRAF's ability to model

Table 8
Performance comparison of current state-of-the-art single models on the test-dev and test-std of VQA v2.

Model	Year	Test-dev (%)				Test-std (%)
		Yes/No	Number	Other	All	All
non-Transformer-based						
BUTD	2018	81.82	44.21	56.05	65.32	65.67
MUAN	2019	86.77	54.40	60.89	70.82	71.10
Transformer-based						
MCAN	2019	86.82	53.26	60.72	70.63	70.90
CLVIN	2023	87.29	53.53	61.15	71.05	71.35
MMnasNet	2020	87.27	55.68	61.05	71.24	71.46
TRAR	2021	87.43	53.80	61.81	71.45	–
LAST-G	2022	87.74	54.51	61.83	71.67	71.94
RRAF	2024	87.03	**55.39**	60.95	71.06	71.34
MCAN-PA	2022	86.99	54.86	61.09	71.05	71.52
SPCA-Net	2022	87.18	54.98	61.52	71.35	71.67
VMAN	2024	87.55	53.82	61.92	71.56	–
RWSAN	2023	86.45	52.18	60.38	70.19	–
$LRCN_{St-SA-6}$	–	87.57	53.61	61.99	71.58	71.71
$LRCN_{CSt-SA-6}$	–	87.31	54.43	**62.33**	71.72	71.77
$LRCN_{ED-SA-6}$	–	**87.74**	53.99	62.29	**71.83**	**72.12**

Table 9
Performance comparison with the state-of-the-art models on the CLEVR dataset. The best result is highlighted in bold, and the second best is underlined.

Model	Year	Overall	Count	Exist	Comp-Num	Query-Att	Comn-Att
TRAR-Base	2021	98.54	96.34	99.24	98.60	99.43	98.93
SANMT	2021	96.60	91.50	97.90	98.70	98.60	98.00
LAST-c	2022	**98.72**	**96.81**	99.31	98.77	99.52	**99.22**
RWSAN	2022	98.42	96.34	99.27	97.66	99.45	98.71
CLVIN	2023	98.67	96.65	99.23	98.81	99.51	99.06
CBNS	2024	98.40	99.80	98.40	99.00	98.40	99.20
$LRCN_{CSt-SA-6}$	–	98.68	96.68	**99.33**	**99.01**	99.52	99.14
$LRCN_{ED-SA-6}$	–	98.63	96.38	**99.33**	**99.01**	99.51	99.14
$LRCN_{St-SA-6}$	–	98.57	96.03	99.31	98.82	**99.54**	98.93

visual object relationships makes it excel in number reasoning tasks. LRCN introduces a simple yet effective skip connection by adding residual connections between layers, ensuring more stable information transfer across layers and reducing information loss during the propagation from lower to higher layers, thereby enhancing training stability. The core design of LRCN emphasizes the preservation of global information and cross-modal interaction, but it lacks specialized modeling for local relationships or unknown dependencies among visual objects. Therefore, for object counting or local reasoning in "Number" type questions, LRCN may struggle to capture these local details as precisely as LAST-G and RRAF. Moreover, the results show that these methods (SPCA-Net Yan et al., 2022, MCAN-PA Mao et al., 2022), like SPCA-Net, perform very well on "Number" type questions (e.g., 54.98% for SPCA-Net) but are less outstanding on other types.

From this, we can infer that, compared to models specifically designed for image reasoning such as RRAF, LRCN exhibits limitations in image reasoning tasks (e.g., Number type questions). However, LRCN enhances deep multimodal information transmission through LRM without increasing computational complexity, thereby gaining an advantage in tasks requiring global understanding and cross-modal information interactions, such as Yes/No and Other types of questions.

Notably, compared to the baseline model TRAR with the same number of layers, the 'All' accuracy of $LRCN_{ED}$ models on the test-dev is higher than 0.4%. Under the same conditions, $LRCN_{CSt}$ is 0.45% higher in the 'Other' type, 0.63% higher in the 'Number' type, and only 0.04% lower than LSAT-G. These results provide significant evidence of the effectiveness of LRCN. Particularly, to strike a balance between performance and model size, we chose only the model when L=6 as our primary model for comparison with SOTAs rather than the model when L=8, which has higher accuracy performance.

4.4.2. Comparison results on the CLEVR

To further assess the generalization capability of LRCN, we also validated LRCN on the CLEVR dataset, which is focused on visual reasoning abilities. We selected two LRCN variants $LRCN_{CSt-SA-6}$ and $LRCN_{ED-SA-6}$ as the final models for comparison. The comparison results on the CLEVR dataset are demonstrated as shown in Table 9. Analysis of the experimental results reveals that the LRCN model achieved competitive accuracy across nearly all metrics in the CLEVR dataset. CBNS (Bao et al., 2024) utilizes neural network learning and symbolic reasoning to achieve efficient VQA, and proposes a confidence-based neural-symbolic (CBNS) method for quantitatively evaluating neural symbolic reasoning based on uncertainty. LAST-c proposes a window self-attention mechanism for image self-attention modeling, enabling the model to effectively capture local regions of the image, reduce redundant information, and enhance the understanding of image context. Compared to LAST-c, our model exhibits limited performance improvement. We hypothesize that the reason lies in the critical importance of modeling rich structural information in datasets that test reasoning capabilities. However, the presence of abundant visual information in images necessitates meticulous processing to achieve superior reasoning performance. LRCN primarily introduces a simple yet effective layer residual mechanism to enhance information transmission. While LRCN has advantages in capturing multi-layer information, it does not specifically address structured visual information, which results in some aspects of its performance being inferior to that of LAST-c. However, it is undeniable that the final experimental results show that LRCN still has positive feedback for the model's inference ability.

4.5. Visualization analysis

This section validates the performance of our models by discussing some qualitative experimental results. As shown in Fig. 8, three typical

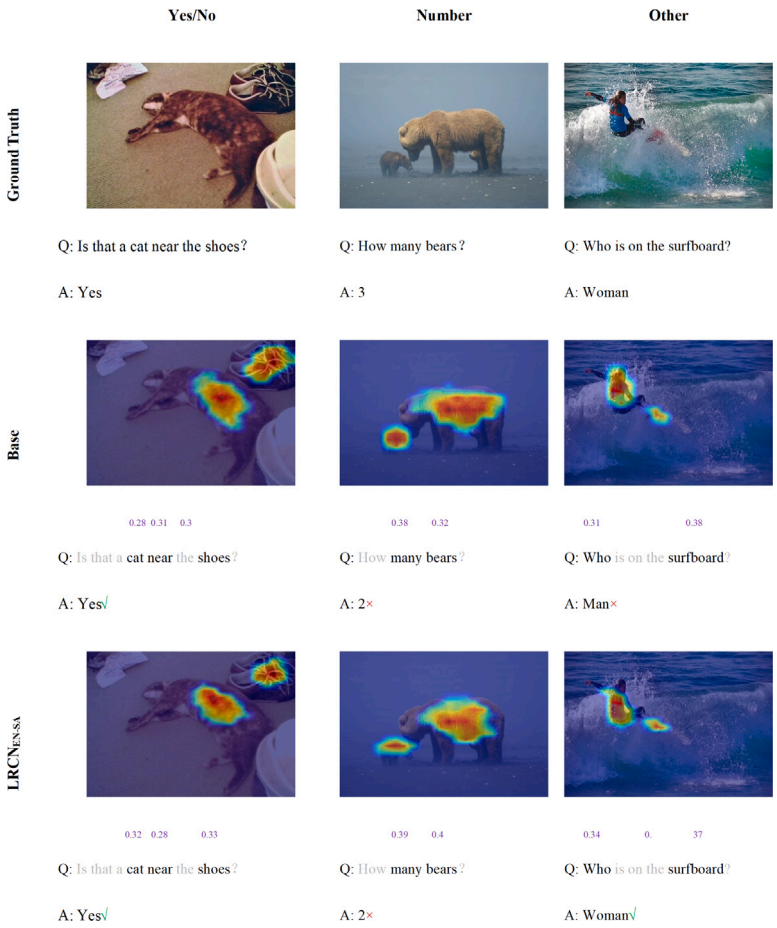

Fig. 8. Visualizations of the attentions in LRCN. In the learned visual attention maps, the highlighted areas obtain higher attention weights, and the value on the question word (by purple) indicates the weight of the question word.

samples are selected from the VQA dataset to provide the learned visual attention and the predicted answers, where the first row is the original image, question, and the ground truth answer; the second row denotes the result of the TRAR base model; and the last two rows are our proposed LRCN model with six layers. The highlighted part of the image represents the significant attention-weight relationship between the image area and the text, and the purple value on the question indicates the attention-weight of the question word. Although all three models give correct answer predictions for the first example, the Encoder–Decoder structure often focuses the question keywords as average and random. The Co-Stacking structure incorporates a co-attention of image-guided questions, which tends to give higher keyword weights to objects in the image. For example, when shoe objects and cat objects are present in a larger image area, co-attention accordingly assigns more attention weight to 'shoes' and 'cat'. For the second and third examples, we find that the $LRCN_{ED-SA}$ focuses more on attention areas than the base model due to the LRM allowing the model to have more efficient gradient propagation during the image feature processing phase, as the analysis is in Section 4.3. In the deep VQA model, the model is biased towards learning real objects in the low layers and abstract textures of objects in the high layers. When we bridge the information propagation between layers through the LRM, it allows the model to learn fine-grained texture features of higher dimensions and can pay attention to small regions to achieve better classification and prevent the phenomenon of attention diffusion.

5. Concluding remarks and future directions

In this work, we propose a lightweight and effective Layer-residual mechanism (LRM) to alleviate the information dispersion phenomenon in traditional attention networks. The LRM extracts rich underlying information by adding a skip edge between the same type of attention blocks to help the model produce fine-grained multimodal features. In addition, in the existing methods, such as pure-stacking, only visual features are inferred via guided attention, ignoring the need to use image content to navigate question keyword inference. Thus, we proposed a novel Co-Stacking structure that adds textual co-attention after textual feature self-attention to help the model focus on the question keywords that match the images. We introduced the LRM into the TRAR base model and proposed a Layer-residual Co-Attention Network (LRCN). Our comparative results on the VQA v2 and CLEVR datasets demonstrate the effectiveness and generalizability of the LRCN model. However, LRCN also exhibits certain limitations; for instance, in the CLEVR dataset, which focuses exclusively on structured visual information, the model's performance is relatively inferior compared to that on VQA v2. In future research, we aim to explore the adaptability of LRCN across various tasks, particularly in those with differing levels of complexity.

CRediT authorship contribution statement

Dezhi Han: Conceptualization, Methodology, Investigation, Funding acquisition, Supervision, Project administration. **Jingya Shi:** Conceptualization, Methodology, Software, Resources, Investigation, Writing – original draft. **Jiahao Zhao:** Conceptualization, Methodology, Software, Resources, Investigation, Writing – original draft. **Huafeng Wu:** Conceptualization, Methodology, Investigation, Funding acquisition, Supervision, Project administration. **Yachao Zhou:** Formal analysis, Investigation, Resources, Software. **Ling-Huey Li:** Conceptualization, Methodology, Software, Resources, Investigation, Writing –

original draft. **Muhammad Khurram Khan:** Writing – review, Validation, Visualization. **Kuan-Ching Li:** Writing – review & editing, Resources, Supervision.

Funding

This work is partly supported by the National Natural Science Foundation of China (Grant No. 52331012), the Natural Science Foundation of Shanghai, China under Grant 21ZR1426500 and the King Saud University, Riyadh, Saudi Arabia under project number (RSP2024R12).

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.

Data availability

The VQA2.0 dataset is licensed under CC BY 4.0. The data supporting this study's findings are openly available at https://visualqa.org/terms.html.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE conference on computer vision and pattern recognition* (pp. 6077–6086). Computer Vision Foundation / IEEE Computer Society.

Ba, L. J., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *CoRR*, arXiv:1607.06450.

Bao, Y., Xing, T., & Chen, X. (2024). Confidence-based interactable neural-symbolic visual question answering. *Neurocomputing*, *564*, Article 126991.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Lecture notes in computer science*: *vol. 12346*, *Computer vision - ECCV 2020 - 16th European conference, glasgow, UK, August 23-28, 2020, proceedings, part i* (pp. 213–229). Springer, http://dx.doi.org/10.1007/978-3-030-58452-8_13.

Chefer, H., Gur, S., & Wolf, L. (2021). Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *2021 IEEE/CVF international conference on computer vision* (pp. 387–396). IEEE, http://dx.doi.org/10.1109/ICCV48922.2021.00045.

Chen, C., Han, D., & Chang, C.-C. (2022). CAAN: Context-aware attention network for visual question answering. *Pattern Recognition*, *132*, Article 108980.

Chen, C., Han, D., & Chang, C.-C. (2024). MPCCT: multimodal vision-language learning paradigm with context-based compact transformer. *Pattern Recognition*, *147*, Article 110084.

Chen, C., Han, D., & Shen, X. (2023). CLVIN: complete language-vision interaction network for visual question answering. *Knowledge-Based Systems*, *275*, Article 110706.

Cho, J., Lu, J., Schwenk, D., Hajishirzi, H., & Kembhavi, A. (2020). X-LXMERT: paint, caption and answer questions with multi-modal transformers. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 8785–8805). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2020.EMNLP-MAIN.707.

Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In *2020 IEEE/CVF conference on computer vision and pattern recognition* (pp. 10575–10584). Computer Vision Foundation / IEEE, http://dx.doi.org/10.1109/CVPR42600.2020.01059.

Deng, J., Yang, Z., Chen, T., Zhou, W., & Li, H. (2021). Transvg: End-to-end visual grounding with transformers. In *2021 IEEE/CVF international conference on computer vision* (pp. 1749–1759). IEEE, http://dx.doi.org/10.1109/ICCV48922.2021.00179.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), (pp. 4171–4186). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/N19-1423.

Goyal, Y., Khot, T., Agrawal, A., Summers-Stay, D., Batra, D., & Parikh, D. (2019). Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, *127*, 398–414.

Guo, W., Zhang, Y., Yang, J., & Yuan, X. (2021). Re-attention for visual question answering. *IEEE Transactions on Image Processing*, *30*, 6730–6743. http://dx.doi.org/10.1109/TIP.2021.3097180.

Han, D., Pan, N., & Li, K.-C. (2020). A traceable and revocable ciphertext-policy attribute-based encryption scheme based on privacy protection. *IEEE Transactions on Dependable and Secure Computing*, *19*(1), 316–327.

Han, D., Zhou, H., Weng, T.-H., Wu, Z., Han, B., Li, K.-C., & Pathan, A.-S. K. (2023). LMCA: a lightweight anomaly network traffic detection model integrating adjusted mobilenet and coordinate attention mechanism for IoT. *Telecommunication Systems*, *84*(4), 549–564.

Han, D., Zhu, Y., Li, D., Liang, W., Souri, A., & Li, K.-C. (2021). A blockchain-based auditable access control system for private data in service-centric IoT environments. *IEEE Transactions on Industrial Informatics*, *18*(5), 3530–3540.

He, R., Ravula, A., Kanagal, B., & Ainslie, J. (2020). Realformer: Transformer likes residual attention. arXiv preprint arXiv:2012.11747.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition* (pp. 770–778). IEEE Computer Society, http://dx.doi.org/10.1109/CVPR.2016.90.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. http://dx.doi.org/10.1162/NECO.1997.9.8.1735.

Huang, X., Han, D., Weng, T.-H., Wu, Z., Han, B., Wang, J., Cui, M., & Li, K.-C. (2022). A localization algorithm for DV-hop wireless sensor networks based on manhattan distance. *Telecommunication Systems*, *81*(2), 207–224.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. R. Bach, & D. M. Blei (Eds.), *JMLR workshop and conference proceedings*: *vol. 37*, *Proceedings of the 32nd international conference on machine learning* (pp. 448–456). JMLR.org, URL http://proceedings.mlr.press/v37/ioffe15.html.

Jiang, H., Misra, I., Rohrbach, M., Learned-Miller, E. G., & Chen, X. (2020). In defense of grid features for visual question answering. In *2020 IEEE/CVF conference on computer vision and pattern recognition* (pp. 10264–10273). Computer Vision Foundation / IEEE, http://dx.doi.org/10.1109/CVPR42600.2020.01028.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., & Girshick, R. B. (2017). CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE conference on computer vision and pattern recognition* (pp. 1988–1997). Honolulu, HI, USA: IEEE Computer Society.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Li, Q., Chen, Y., He, X., & Huang, L. (2024). Co-training transformer for remote sensing image classification, segmentation and detection. *IEEE Transactions on Geoscience and Remote Sensing*.

Li, S., Gong, C., Zhu, Y., Luo, C., Hong, Y., & Lv, X. (2024). Context-aware multi-level question embedding fusion for visual question answering. *Information Fusion*, *102*, Article 102000. http://dx.doi.org/10.1016/J.INFFUS.2023.102000.

Li, J., Han, D., Weng, T.-H., Wu, H., Li, K.-C., & Castiglione, A. (2024). A secure data storage and sharing scheme for port supply chain based on blockchain and dynamic searchable encryption. *Computer Standards Interfaces*, Article 103887.

Li, J., Han, D., Weng, T.-H., Wu, H., Li, K.-C., & Castiglione, A. (2025). A secure data storage and sharing scheme for port supply chain based on blockchain and dynamic searchable encryption. *Computer Standards & Interfaces*, *91*, Article 103887.

Li, Z., Li, L., Chen, J., & Wang, D. (2024). A multi-head attention mechanism aided hybrid network for identifying batteries' state of charge. *Energy*, *286*, Article 129504.

Li, Y., Liang, W., Xie, K., Zhang, D., Xie, S., & Li, K.-C. (2023). LightNestle: Quick and accurate neural sequential tensor completion via meta learning. In *IEEE INFOCOM 2023 - IEEE conference on computer communications* (pp. 1–10). IEEE, http://dx.doi.org/10.1109/INFOCOM53939.2023.10228967.

Li, B., Wang, J., Zhao, M., & Zhou, S. (2022). Two-stage multimodality fusion for high-performance text-based visual question answering. In *Lecture notes in computer science*: *vol. 13844*, *Computer vision - ACCV 2022 - 16th Asian conference on computer vision, macao, China, December 4-8, 2022, proceedings, part IV* (pp. 658–674). Springer, http://dx.doi.org/10.1007/978-3-031-26316-3_39.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., & Gao, J. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Lecture notes in computer science*: *vol. 12375*, *Computer vision - ECCV 2020 - 16th European conference, glasgow, UK, August 23-28, 2020, proceedings, part XXX* (pp. 121–137). Springer, http://dx.doi.org/10.1007/978-3-030-58577-8_8.

Liang, W., Hu, Y., Zhou, X., Pan, Y., & Wang, K. I.-K. (2022). Variational few-shot learning for microservice-oriented intrusion detection in distributed industrial IoT. *IEEE Transactions on Industrial Informatics*, *18*(8), 5087–5095. http://dx.doi.org/10.1109/TII.2021.3116085.

Liang, W., Li, Y., Xie, K., Zhang, D., Li, K.-C., Souri, A., & Li, K. (2023). Spatial-temporal aware inductive graph neural network for C-ITS data recovery. *IEEE Transactions on Intelligence Transport System*, *24*(8), 8431–8442. http://dx.doi.org/10.1109/TITS.2022.3156266.

Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: common objects in context. In *Lecture notes in computer science*: *vol. 8693*, *Computer vision - ECCV 2014 - 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v* (pp. 740–755). Springer, http://dx.doi.org/10.1007/978-3-319-10602-1_48.

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI Open*, *3*, 111–132. http://dx.doi.org/10.1016/J.AIOPEN.2022.10.001.

Long, J., Liang, W., Li, K.-C., Wei, Y., & Marino, M. D. (2023). A regularized cross-layer ladder network for intrusion detection in industrial internet of things. *IEEE Transactions on Industrial Informatics*, *19*(2), 1747–1755. http://dx.doi.org/10.1109/TII.2022.3204034.

Manmadhan, S., & Kovoor, B. C. (2023). Object-assisted question featurization and multi-CNN image feature fusion for visual question answering. *International Journal of Intelligence and Information Technology*, *19*(1), 1–19. http://dx.doi.org/10.4018/IJIIT.318671.

Mao, A., Yang, Z., Lin, K., Xuan, J., & Liu, Y.-J. (2022). Positional attention guided transformer-like architecture for visual question answering. *IEEE Transactions on Multimedia*, *25*, 6997–7009.

Nguyen, B. X., Do, T., Tran, H., Tjiputra, E., Tran, Q. D., & Nguyen, A. (2022). Coarse-to-fine reasoning for visual question answering. In *IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 4557–4565). IEEE, http://dx.doi.org/10.1109/CVPRW56347.2022.00502.

Nguyen, N. H., Vo, D. T., Van Nguyen, K., & Nguyen, N. L.-T. (2023). Openvivqa: Task, dataset, and multimodal fusion models for visual question answering in vietnamese. *Information Fusion*, *100*, Article 101868.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), (pp. 1532–1543). ACL, http://dx.doi.org/10.3115/V1/D14-1162.

Qin, B., Hu, H., & Zhuang, Y. (2023). Deep residual weight-sharing attention network with low-rank attention for visual question answering. *IEEE Transactions on Multimedia*, *25*, 4282–4295. http://dx.doi.org/10.1109/TMM.2022.3173131.

Qiu, C., Liu, Z., Song, Y., Yin, J., Han, K., Zhu, Y., Liu, Y., & Sheng, V. S. (2023). Rtunet: Residual transformer unet specifically for pancreas segmentation. *Biomedical Signal Processing and Control*, *79*, Article 104173.

Rahman, T., Chou, S.-H., Sigal, L., & Carenini, G. (2021). An improved attention for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1653–1662).

Shen, X., Han, D., Guo, Z., Chen, C., Hua, J., & Luo, G. (2023). Local self-attention in transformer for visual question answering. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *53*(13), 16706–16723.

Shen, X., Han, D., Zong, L., Guo, Z., & Hua, J. (2024). Relational reasoning and adaptive fusion for visual question answering. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *54*(6), 5062–5080.

Song, X., Han, D., Chen, C., Shen, X., & Wu, H. (2024). Vman: visual-modified attention network for multimodal paradigms. *Visual Computer*, 1–18.

Sood, E., Kögel, F., Müller, P., Thomas, D., Bâce, M., & Bulling, A. (2023). Multimodal integration of human-like attention in visual question answering. In *IEEE/CVF conference on computer vision and pattern recognition* (pp. 2648–2658). IEEE, http://dx.doi.org/10.1109/CVPRW59228.2023.00265.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Takase, S., Kiyono, S., Kobayashi, S., & Suzuki, J. (2022). B2t connection: Serving stability and performance in deep transformers. arXiv preprint arXiv:2206.00330.

Tan, H., & Bansal, M. (2019). LXMERT: learning cross-modality encoder representations from transformers. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 5099–5110). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/D19-1514.

Teney, D., Anderson, P., He, X., & van den Hengel, A. (2018). Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *2018 IEEE conference on computer vision and pattern recognition* (pp. 4223–4232). Computer Vision Foundation / IEEE Computer Society, http://dx.doi.org/10.1109/CVPR.2018.00444.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30: annual conference on neural information processing systems 2017* (pp. 5998–6008).

Wang, J., Lai, C., Wang, Y., & Zhang, W. (2024). EMAT: Efficient feature fusion network for visual tracking via optimized multi-head attention. *Neural Networks*, *172*, Article 106110.

Wang, C., Shen, Y., & Ji, L. (2022). Geometry attention transformer with position-aware LSTMs for image captioning. *Expert Systems with Applications*, *201*, Article 117174. http://dx.doi.org/10.1016/J.ESWA.2022.117174.

Wu, H., Wang, F., Mei, X., Liang, L., Han, B., Han, D., Weng, T.-H., & Li, K.-C. (2024). A novel fuzzy control path planning algorithm for intelligent ship based on scale factors. *Journal of Supercomputing*, *80*(1), 202–225.

Xie, S., Zhang, H., Guo, J., Tan, X., Bian, J., Awadalla, H. H., Menezes, A., Qin, T., & Yan, R. (2023). ResiDual: Transformer with dual residual connections. *CoRR*, http://dx.doi.org/10.48550/ARXIV.2304.14802, arXiv:2304.14802.

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., & Liu, T. (2020). On layer normalization in the transformer architecture. In *International conference on machine learning* (pp. 10524–10533). PMLR.

Yan, F., Silamu, W., Li, Y., & Chai, Y. (2022). SPCA-net: a based on spatial position relationship co-attention network for visual question answering. *Visual Computer*, *38*(9), 3097–3108.

Yang, Z., Xuan, J., Liu, Q., & Mao, A. (2022). Modality-specific multimodal global enhanced network for text-based visual question answering. In *IEEE international conference on multimedia and expo* (pp. 1–6). IEEE.

Yu, Z., Cui, Y., Yu, J., Tao, D., & Tian, Q. (2019). Multimodal unified attention networks for vision-and-language interactions. arXiv preprint arXiv:1908.04107.

Yu, Z., Cui, Y., Yu, J., Wang, M., Tao, D., & Tian, Q. (2020). Deep multimodal neural architecture search. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 3743–3752).

Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., & Yan, S. (2022). MetaFormer is actually what you need for vision. In *IEEE/CVF conference on computer vision and pattern recognition* (pp. 10809–10819). IEEE, http://dx.doi.org/10.1109/CVPR52688.2022.01055.

Yu, Z., Yu, J., Cui, Y., Tao, D., & Tian, Q. (2019). Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6281–6290).

Zhang, H., Li, R., & Liu, L. (2022). Multi-head attention fusion network for visual question answering. In *2022 IEEE international conference on multimedia and expo* (pp. 1–6). IEEE.

Zhou, Y., Ren, T., Zhu, C., Sun, X., Liu, J., Ding, X., Xu, M., & Ji, R. (2021). Trar: Routing the attention spans in transformer for visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2074–2084).