Refining Visual Perception for Decoration Display: A Self-Enhanced Deep Captioning Model

Longfei Huang hlf@njust.edu.cn
Xiangyu Wu wxy_yyjhl@njust.edu.cn
Jingyuan Wang wangjingyuan357@njust.edu.cn
Weili Guo wlguo@njust.edu.cn
Yang Yang*[*] yyang@njust.edu.cn
Nanjing University of Science and Technology

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

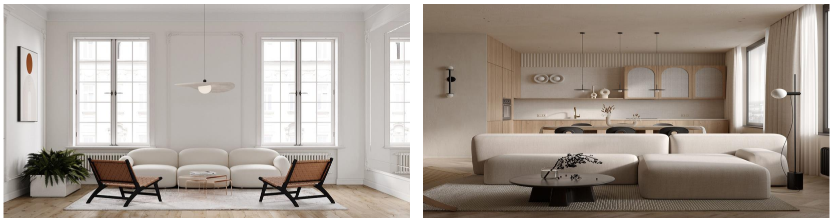
Traditional decoration displays usually include renderings and corresponding descriptions to give users a deeper understanding and feeling. Nevertheless, describing massive renderings undoubtedly requires a lot of manpower. Thanks to the development of artificial intelligence, especially deep learning techniques, image captioning has been developed to automatically generate captions for given images. However, the defect of exploring "perceptive" words (e.g., bright, capacious, and comfortable, etc) is exposed when transferring existing captioning approaches to the decoration display task. To address this issue, in this paper, we propose a self-enhanced deep captioning model, which generates the captions with visual perception using the designed Self-Enhanced Transformer (SET). In detail, SET first pre-trains the scene-aware encoder, which employs the multi-task-based multi-modal transformer to enhance the perceptive semantics of the visual representations. Then, SET combines the pre-trained encoder with the transformer decoder for fine-tuning and designs a knowledge-enhanced module on the top of the decoder to adaptively fuse the decoded representations and retrieved language cues for making more suitable word prediction. In experiments, we first validate SET on the MS-COCO dataset, and we achieve at least 0.6 improvements on the CIDEr-D score. Furthermore, to address the effectiveness of SET on the decoration display task, we collect a new dataset called DecorationCap. We present a thorough empirical analysis to verify the generality of SET and find that SET surpasses other comparison methods with at least 6.8 improvements on the CIDEr-D score.

Keywords: Cross-modal Learning, Image Captioning, Decoration Display, Transformer

1. Introduction

With the development of the Internet, more and more real estate users choose to browse and shop online. Generally, to provide better user understanding and experience, companies will hire professionals to describe the renderings. However, describing a large amount of decoration renderings will undoubtedly consume a huge number of manpower and material resources. Fortunately, with the development of artificial intelligence, especially deep learning techniques, image captioning has been researched. The task of image captioning is to learn a mapping function from visual features to natural language features, thereby automatically generating captions for the given images.

[*] Corresponding author



A pure white color scheme, high ceilings, and tall windows build an ethereal living room decor scheme. A white living room rug underlines a curvaceous white sofa design, which is too cushiony to resist.

Smooth creamy walls paint a restful cocoon around the living room, where a modern modular sofa design divides the open floor plan with a softly rounded silhouette.

DLCT: A living room is mainly white with white sofa, which is simple and atmospheric.

DLCT: A living room is mainly gray and white, with white sofa, simple and atmospheric.

RSTNet: A living room sofa background wall is painted white to make the space more light.

RSTNet: A living room is mainly gray, with white walls, making the space look light.

Figure 1: Examples of decoration displays. The two renderings describe living rooms with different styles. “perceptive” words are marked in red. DLCT Luo et al. (2021) is the captioning model, and RSTNet Zhang et al. (2021b) is the captioning model considering perceptive words.

Early image captioning methods mainly followed retrieval or template-based approaches Gupta et al. (2012), which accomplished sentence generation by retrieving existing captions or relying on hand-coded language structures. However, the expressiveness of these approaches is limited considering the inflexibility.

The decoration display task emphasizes the descriptions with “perceptive” words (e.g., bright, capacious, and comfortable, etc). Several approaches Yang et al. (2019); Zhang et al. (2021b) attempted to introduce context knowledge to promote the generation of perceptive words. These methods only consider additional textual knowledge to assist inference and ignore the matching with visual priors, which is easy to cause inductive bias. For example, when seeing the relationship “person on bike”, these methods naturally replace “on” with “ride” and infer “person riding bike on a road”, even though the ”road” does not exist in the image.

Considering the semantic gap between vision and language, many words have no direct visual representation Yang et al. (2022). Therefore, traditional image captioning models always fail to predict “perceptive” words due to the overemphasis on object description, leading to poor prediction and understanding. For example, as shown in Figure 1, there exist “perceptive” words such as “ethereal”, “curvaceous”, and “modern” when describing the living rooms, but the method, i.e., DLCT Luo et al. (2021), only provides a straight caption, rather than an understandable sentence. To address this issue, several approaches Yang et al. (2019); Zhang et al. (2021b) attempted to introduce context knowledge to promote the generation of perceptive words. These methods tried to integrate the context language knowledge into the model, but they are limited to expanding perceptive words such as “with” and “a”, leading to the monotony problem. For example, as shown in Figure 1, RSTNet Zhang et al. (2021b) generates simple and monotonous perceptive descriptions for different styles of living room renderings, e.g., “the white wall makes the space more light”.

To address these challenges, we propose a novel Self-Enhanced Transformer (SET), which aims to enhance the perceptive semantics into visual embedding during encoding and augment the knowledge adaptively during decoding. In detail, SET includes two core modules: 1) Scene-Aware encoder. We first pre-train a multi-modal transformer encoder with the image-sentence pair as input, which aims to enhance the perceptive semantics into the visual representation with the multi-task losses. 2) Knowledge-Enhanced module. We adopt the pre-trained encoder with a decoder for fine-tuning, in which we design a novel knowledge-enhanced module on the top of the decoder to adaptively fuse the retrieved language cues from the constructed domain knowledge graph to obtain more accurate predictions. Consequently, the proposed SET can address the defect of exploring perceptive words from the perceptive image encoding and knowledge-enhanced generation, which is particularly prominent in the automatic description of decoration display and other captioning tasks.

2. RELATED WORK

2.1. Image Captioning

Image captioning aims to automatically generate natural language descriptions for images. Early works designed template-based methods Kuznetsova et al. (2012). Inspired by the encoder-decoder technique Ramos et al. (2023) in the NLP technique, many approaches have designed encoder-decoder-based methods Fu et al. (2024). Furthermore, with the success of attention mechanism, Luo et al. (2021) introduced both regional features and grid features into the attention module to supplement fine-grained details and context information. However, traditional image captioning models always decode based on global or local visual representations, whereas many words have no direct visual representation Yang et al. (2022), leading to the generation failure of perceptive words. To enhance the generation of perceptive words, several attempts tried to introduce the context knowledge into the decoding Yang et al. (2019); Zhang et al. (2021b). Zhang et al. (2021b) built a BERT-based language model to extract language context and proposed an adaptive attention module for word prediction. However, they only focus on extending generated words, which ignores the problem that even the same visual object can be expressed in various perceptive words under different scenes, thus leading to the monotony problem.

2.2. Transformer Models

To process the limitation of RNN-based methods, Vaswani et al. (2017); Chen et al. (2021) proposed the Transformer with self-attention mechanisms and acquired great success in neural language processing NLP task. Following this idea, many attempts have transferred transformer into computer vision Yang et al. (2023b, 2024) and multi-modal learning Yang et al. (2023a). For example, Li et al. (2020) used object tags detected in images as anchor points to ease the learning of alignments; Yu et al. (2021) incorporated structured knowledge obtained from scene graphs to learn joint representations of vision and language. Meanwhile, many approaches have also adopted the transformer for image captioning tasks. For example, Herdade et al. (2019) introduced transformer architecture into image captioning, and used the self-attention mechanism to model the spatial relationship between regional features; Pan et al. (2020) introduced bilinear pooling into the attention module of a base transformer, which selectively used visual information for multi-modal reasoning; Ji et al.

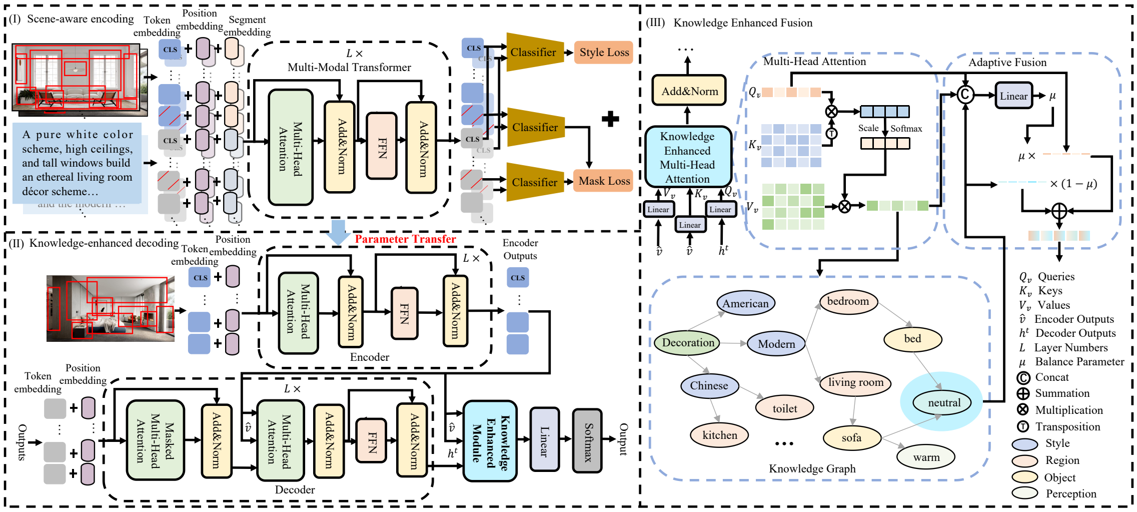


Figure 2: An illustration of SET. Considering the input, the image is represented with region feature representations by Faster R-CNN Ren et al. (2015), and the sentence is represented with word feature representations with a BERT-based language model. SET first pre-train scene-aware encoder, which aims to enhance the perceptive semantics of the visual representations. Then, SET adopts the pre-trained encoder with a decoder for fine-tuning, in which a novel knowledge-enhanced module is designed to adaptively fuse the decoded representation and retrieved language cues. As a result, SET can get more perceptive captions.

(2021) designed a global enhancement module in the base transformer architecture, which can capture the global features to guide the caption generation.

3. Proposed Method

3.1. Preliminaries

In an image captioning scenario, without any loss of generality, we define the image-sentence pair as $(\mathbf{v}, \mathbf{w})$, where $\mathbf{v}$ denotes the image, $\mathbf{w}$ represents the corresponding sentence. The image captioning task is to learn a mapping function for automatically generating sentences using the input images, thereby the sentence acts as the supervision in the learning process. As shown in Figure 2, SET generates image captions with visual perception by effectively fusing perceptive information in both encoding and decoding processes. Specifically, the framework is composed of two components: 1) Scene-Aware encoder and 2) Knowledge-Enhanced decoder. Considering the reproducibility and effectiveness, inspired by Su et al. (2020), the input representations of image regions and sentence words are extracted by Faster R-CNN and BERT, respectively.

3.2. Self-Enhanced Transformer

To further improve the visual representations containing textual semantics, we turn to adopt the multi-modal transformer as the encoder during pre-training, and then transfer the pre-trained encoder to the classical encoder-decoder framework for fine-tuning.

Encoder. To build the multi-modal transformer, inspire by Su et al. (2020), we concatenate the image-sentence pair as a long sequence $\mathbf{x} = \{[IMG], \Phi(\mathbf{v}^1), \Phi(\mathbf{v}^2), \cdots, \Phi(\mathbf{v}^{S_v}), [TXT], \mathbf{w}^1, \mathbf{w}^2, \cdots, \mathbf{w}^{S_w}\} \in \mathcal{R}^{(S_v+S_w+2)\times d}$, where $d = d_w$ is the dimension of cross-modal common feature space, $\Phi(\cdot) \in \mathcal{R}^{d_v \times d_w}$ denotes the linear mapping function that maps the visual representations to the common feature space. The special tokens $[IMG]$ and $[TXT]$ are defined to learn the global representations for image/text modalities. In the multi-head attention layer, the input representations can be used to compute three matrices: Q, K, and V corresponding to queries, keys, and values. The dot-product similarity between queries and keys determines attention distributions:

$$\begin{aligned} Q = \mathbf{x}W_Q, \quad K = \mathbf{x}W_K, \quad V = \mathbf{x}W_V, \\ A = \frac{QK^\top}{\sqrt{d_M}} \quad Att(\mathbf{x}) = softmax(A)V, \end{aligned} \tag{1}$$

where $Q \in \mathcal{R}^{(S_v+S_w+2)\times d_M}$, $K \in \mathcal{R}^{(S_v+S_w+2)\times d_M}$, $V \in \mathcal{R}^{(S_v+S_w+2)\times d_M}$, and $W_Q \in \mathcal{R}^{d\times d_M}$, $W_K \in \mathcal{R}^{d\times d_M}, W_V \in \mathcal{R}^{d\times d_M}$ are learnable matrices. Multi-head attention comprises M parallel heads, and $d_M = d/M$. Results of each head are concatenated and passed through a linear transformation to construct the output, i.e., $MultiAtt(\mathbf{x}) = [Att(\mathbf{x})_1, \cdots, Att(\mathbf{x})_M]$ W_M, where $W_M \in \mathcal{R}^{d\times d}$ is the learnable parameter. The FFN is a fully-connected network: $FFN(MultiAtt(\mathbf{x})) = \max(0, MultiAtt(\mathbf{x})W_1 + b_1)W_2 + b_2$, where W_1 and W_2 are matrices for linear transformation, b_1 and b_2 are the bias terms. Meanwhile, each sub-layer is followed by dropout, shortcut connection He et al. (2016), and layer normalization Ba et al. (2016). Note that the position features of the imaging modality are designed according to Li et al. (2020) using the location of the region, i.e., each image region position is represented $[\frac{a_1}{WI}, \frac{c_1}{HE}, \frac{a_2}{WI}, \frac{c_2}{HE}, \frac{(c_2-c_1)(a_2-a_1)}{WI\times HE}]W_{lo}$, where WI, HE are the width and height of the input image, and the last value represents the fraction of image covered, $W_{lo} \in \mathcal{R}^{5\times d}$ is the linear mapping function. The position features of the text modality are designed according to the original method Vaswani et al. (2017). Two types of segment embedding, i.e., $SE_1 \in \mathcal{R}^d$ and $SE_2 \in \mathcal{R}^d$, are defined to separate input elements from different sources, SE_1 denotes tokens from image and SE_2 denotes tokens from the sentence. The learned segment embedding is added to every input element to indicate which segment it belongs to. Finally, we can acquire the global representations from the $[IMG]$ and $[TXT]$ token, i.e., $\hat{\mathbf{v}}^{[IMG]}, \hat{\mathbf{w}}^{[TXT]}$, and individual representations from other tokens.

Scene-Aware Encoder. As shown in Figure 2, to improve the visual representations containing textual semantics, we design a multi-task loss to pre-train the encoder.

$$L = \ell_S + \lambda_1 \ell_M, \tag{2}$$

where ℓ_M employs the masking loss to constrain the cross-modal consistency, and the ℓ_S loss adopts the style prediction loss to enhance the representation learning. λ_1 is the balance parameter.

For masking loss, following Su et al. (2020), we sample image/text tokens and mask them (i.e., using $[MASK]$ tokens) with 15% probability. ℓ_M aims to predict mask token labels (the class label of the image region is predicted by the pre-trained Faster R-CNN and the class label of the word is constructed by the whole vocabulary as one-hot form) based on their surrounding contexts (including the contextual image regions and words):

$$\ell_M = \sum_{(\mathbf{v},\mathbf{w})} \big(\sum_{i \in \mathcal{D}_v} \ell(\mathbf{y}^i, f_v(\hat{\mathbf{v}}^i)) + \sum_{j \in \mathcal{D}_w} \ell(\mathbf{y}^j, f_w(\hat{\mathbf{w}}^j)) \big), \tag{3}$$

where $\mathcal{D}_v$ and $\mathcal{D}_w$ denote the mask token set of image and sentence. $\hat{\mathbf{v}}^i$ and $\hat{\mathbf{w}}^j$ represent the output representations of masked tokens, $\mathbf{y}^i$ and $\mathbf{y}^j$ denote the corresponding class labels. ℓ utilizes the cross-entropy loss here, $f_v(\cdot)$ and $f_w(\cdot)$ denote the image classifier and text classifier, respectively. Moreover, to ensure the global representations of decoration rendering with the same style, we add a style prediction loss:

$$\ell_S = \sum_{(\mathbf{v},\mathbf{w})} \ell(\max(g(\hat{\mathbf{v}}^{[IMG]}), g(\hat{\mathbf{w}}^{[TXT]})), \mathbf{y}^s(\mathbf{v},\mathbf{w})), \tag{4}$$

where $\mathbf{y}^s(\mathbf{v},\mathbf{w}) \in \mathcal{R}^{14}$ represents the style class label, e.g., European style, Chinese style, etc. $g(\cdot)$ is the style classifier.

Decoder. Considering that the parameters in the scene-aware encoder are sharable (i.e., we embed semantic perception into visual representations), we transfer the pre-trained encoder to the classical encoder-decoder framework for fine-tuning. In detail, as shown in Figure 2 (II), we only input the image regions into the encoder and use the word representation generated by the decoder at the last moment and visual region output representations as the input of first decoder layer. The interaction between visual and language representations is completed by using the cross-modal multi-head attention mechanism: $h^t = Decoder(\hat{\mathbf{v}}, \hat{\mathbf{w}}^{<t})$, where $\hat{\mathbf{v}} = \{\hat{\mathbf{v}}^1, \hat{\mathbf{v}}^2, \cdots, \hat{\mathbf{v}}^{S_v}\}$ is the output set of transformer encoder, $\hat{\mathbf{w}}^{<t} = \{\hat{\mathbf{w}}^1, \hat{\mathbf{w}}^2, \cdots, \hat{\mathbf{w}}^{t-1}\}$ is word sequence representations of the partially generated sentence. In detail, we adopt the same cross-attention mechanism as Li et al. (2021), in which we adopt the image region output as the keys (i.e., K) and values (i.e., V), and the partially ground-truth sentence as the queries (i.e., Q). The dot-product similarity between queries and keys provides attention distribution for determining which visual regions to focus on for decoding. The h^t is the hidden state output by the transformer decoder to predict the current word $\hat{\mathbf{w}}^t$.

Knowledge-Enhanced Module. To further fuse the perceptive semantics in the decoding process, we develop the knowledge-enhanced module on top of the classical transformer decoder. In this module, we aim to retrieve similar language cues from the constructed knowledge graph as the complementary information, and then adaptively fuse the retrieved language cues and hidden state output for word prediction. In detail, as shown in the Figure 2 (III), we construct the knowledge graph (KG) of the decoration domain following Che et al. (2020), and initialize the entity representation through the pre-trained scene-aware encoder. To retrieve the knowledge cues, we first embed the visual output representations:

$$\begin{aligned} Q_v &= h^t W_{Q_v}, \quad K_v = \hat{\mathbf{v}} W_{K_v}, \quad V_v = \hat{\mathbf{v}} W_{V_v}, \\ A &= \frac{Q_v K_v^\top}{\sqrt{d_M}} \quad Att(\hat{\mathbf{v}}) = softmax(A) V_v, \\ \bar{\mathbf{v}} &= [Att(\hat{\mathbf{v}})_1, \cdots, Att(\hat{\mathbf{v}})_M] W \end{aligned} \tag{5}$$

where $W_{Q_v} \in \mathcal{R}^{d \times d_M}, W_{K_v} \in \mathcal{R}^{d \times d_M}, W_{V_v} \in \mathcal{R}^{d \times d_M}, W \in \mathcal{R}^{d \times d}$ are learnable matrices. Then, we utilize $\bar{\mathbf{v}}$ to retrieved O most similar entities according to dot-product function $sc(\bar{\mathbf{v}}, \mathbf{e}^o)$, $\mathbf{e}^o$ denotes the $o-$th entity in the KG, and the overall representation can be formulated as $\hat{\mathbf{e}}^t = \sum_{o \in \mathcal{D}_O} \mathbf{e}^o sc(\bar{\mathbf{v}}, \mathbf{e}^o)$, $\mathcal{D}_O$ denotes the retrieved cues set. Lastly, instead of predicting a word using the hidden state h^t directly, we combine language cue representation $\hat{\mathbf{e}}^t$, visual output representation $\bar{\mathbf{v}}$, and the hidden state Q_v together to measure the contribution of visual signals and language signals for each word prediction:

$$\mu = [\bar{\mathbf{v}}, Q_v, \hat{\mathbf{e}}^t] W_\mu, \quad \hat{h}^t = \mu Q_v + (1 - \mu) \hat{\mathbf{e}}^t \tag{6}$$

where W_μ is a fully connected network to predict μ.

Training. To train the model, we first minimize the cross-entropy loss (i.e., ℓ_{XE}) following a standard practice of image captioning, with ground-truth caption $\mathbf{y}_w$ and prediction $\hat{\mathbf{y}}_w$: $\ell_{XE}(\theta) = -\sum_{t=1}^{S_w} \log p_\theta(\mathbf{y}_w^t | \mathbf{y}_w^{1:t-1})$. Then, we directly optimize the non-differentiable metric with self-critical sequence training Rennie et al. (2017): $\ell_{RL}(\theta) = -\mathbb{E}_{\mathbf{y}_w^{1:S_w}} \; p_\theta[r(\mathbf{y}_w^{1:S_w})]$, where $\mathbf{y}_w^{1:S_w}$ denotes the target ground-truth sequence. The parameters θ of the network define a policy p_θ. The reward $r(\cdot)$ is a sentence-level metric for the generated sentence and the ground-truth, which is always represented by the score of captioning metric (e.g., CIDEr-D Vedantam et al. (2015)).

4. Experiments

4.1. Experimental Setup

To demonstrate the effectiveness of SET, we conduct the experiments on two datasets, i.e., the public MS-COCO dataset Lin et al. (2014), and specifically collect the DecorationCap dataset. We firstly adopt the popular MS-COCO dataset as most captioning methods Huang et al. (2019); Cornia et al. (2020); Zhang et al. (2021b), we use $\hat{\mathbf{v}}$ to hierarchically retrieve the language cue representation. DecorationCap dataset is a real-world dataset for decoration display tasks, one of the largest residential service websites in China. In detail, the DecorationCap dataset contains 119,789 decoration cases, in which each case includes 4 scenes (i.e., living room, kitchen, bedroom, and bathroom). Each scene contains various rendering-description pairs with different perspectives. Therefore, the DecorationCap dataset has entirely 838,717 rendering-description pairs, and the descriptions are Chinese. Note that different scenes in the same case usually keep the same decoration style. Considering the DecorationCap dataset has no ground-truths of bounding boxes for each image, we manually annotate 600 images with labelme [1], and pre-trained the Faster R-CNN with

[1] http://labelme.csail.mit.edu/Release3.0/

novel semi-supervised methods Sohn et al. (2020). Following the MS-COCO dataset, we divide 5,000 rendering-description pairs as the validation set, 5,000 rendering-description pairs as the test set, and the rest as the training set.

4.2. Implementations

In experiments, the S_v is set as 36 according to most traditional methods Anderson et al. (2018). We set the maximum length of the sentence as $S_w = 100$, and the excessive parts are removed. The temperature scale parameter $\tau = 0.5$, retrieved knowledge cues $O = 3$, the balance parameters λ_1 is searched in $\{0.01, 0.1, 1, 10\}$ to find the best settings. In all experiments, the batch size is set to 32. The optimization method is Adaptive Moment Estimation (Adam), with a base learning rate of 2×10^{-5}, weight decay of 10^{-4}, learning rate warmed up over the first 8,000 steps, and linear decay of the learning rate. The ratio of dropout is 0.1 and the maximal number of epochs is 25. We run the following experiments on NVIDIA TITAN X GPU. We will publish the dataset sooner.

Pre-training. During pre-training, we use 838,717 rendering-description pairs, of which 5,000 are used as validation set, 5,000 were used as test set, and the rest is training set.

Pre-trained encoder transfer. In the pre-training phase, we concatenate the image-sentence pair as a long sequential input to enhance the perceptive semantics into the visual representation, using the multi-task loss. Then, in the fine-tuning phase, we can directly transfer the pre-trained encoder with only the visual regions as input.

KG construction. Inspired by Zhang et al. (2021a), to introduce perceptual semantics more accurately, we construct a knowledge graph in the decoration domain. In detail, we construct a perceptiveness knowledge graph, which mainly includes the objects (i.e., modified words) and their attributes (i.e., perceptive words), the construction of the knowledge graph can be summarized using three steps: 1) The tokenization step, which uses the functions provided in Che et al. (2020) to tokenize each sentence description into discrete words. 2) Modification relationship extraction step, which explores the modification relationship between discrete words, such as for "warm sofa", the modifier is "warm" (i.e., perceptive word), and the modified word is "sofa". We use the dependency analysis function provided in Che et al. (2020) to extract the attributives and their headwords in each sentence description. 3) The hierarchical construction step, which obtains the style and scenes corresponding to each image by exploring the given prior knowledge of the DecorationCap dataset, such as Chinese style and living room.

The comparison models fall into three categories: 1) traditional deep supervised captioning methods: SCST Rennie et al. (2017), Up-Down Anderson et al. (2018), and AoANet Huang et al. (2019). 2) Deep captioning methods considering perceptive words: SGAE Yang et al. (2019), RSTNet Zhang et al. (2021b). 3) Transformer based methods: ORT Herdade et al. (2019), M^2 Transformer Cornia et al. (2020), X-Transformer Pan et al. (2020) and DLCT Luo et al. (2021) and SmallCap Ramos et al. (2023)

.

Moreover, we conduct extra ablation studies to evaluate each term in our proposed SET: 1) Transformer-based, we remove both the pre-training and knowledge-enhanced module and only adopt the transformer-based encoder and decoder for training. 2) w/o KEM, we remove the knowledge-enhanced module in the decoding process. 3) w/o Pre-training, we remove the pre-training process of the encoder. 4) w/o Pre-training+, we only pre-train the multi-modal encoder to initialize the entity representations in the knowledge graph, without transferring to the fine-tuning phase. 5) w/o ℓ_M, we remove the mask prediction loss in L for pre-training. 6) w/o ℓ_S, we remove the style prediction loss in L for pre-training. 7) Hard μ, we add a sigmoid operator on μ for hard combination. For evaluation, we use different metrics, including BLEU(B@N), METEOR(M), ROUGE-L(R), CIDEr-D(C), and SPICE(S), to evaluate the proposed method and comparison methods.

Table 1: The performances of various methods on MS-COCO.

	Cross-Entropy Loss						**CIDEr-D Score Optimization**					
	B@1	B@4	M	R	C	S	B@1	B@4	M	R	C	S
SCST	-	30.0	25.9	53.4	99.4	-	-	34.2	26.7	55.7	114.0	-
Up-Down	77.2	36.2	27.0	56.4	113.5	20.3	79.8	36.3	27.7	56.9	120.1	21.4
AoANet	77.4	37.2	28.4	57.5	119.8	21.3	80.2	38.9	29.2	58.8	129.8	22.4
ORT	-	-	-	-	-	-	80.5	38.6	28.7	58.4	128.3	22.6
M^2 Transformer	-	-	-	-	-	-	80.8	39.1	29.2	58.6	131.2	22.6
X-Transformer	77.3	37.0	28.7	57.5	120.0	21.8	80.9	39.7	29.5	59.1	132.8	**23.4**
DLCT	-	-	-	-	-	-	81.4	39.8	29.5	59.1	133.8	23.0
SmallCap	-	37.2	28.3	-	121.8	21.5	-	-	-	-	-	-
SGAE	77.6	36.9	27.7	57.2	116.7	20.9	80.8	38.4	28.4	58.6	127.8	22.1
RSTNet	-	-	-	-	-	-	81.1	39.3	29.4	58.8	133.3	23.0
Transformer based	76.5	36.0	27.1	56.5	113.2	20.3	79.2	36.4	27.8	56.9	123.0	21.1
w/o Pre-training	76.4	36.0	27.0	56.3	113.1	20.2	79.0	36.2	27.4	56.5	122.8	21.1
w/o Pre-training+	76.9	36.6	27.5	57.0	117.3	21.0	80.1	37.4	28.7	57.2	128.5	22.1
w/o KEM	77.6	37.0	28.5	57.5	120.1	21.5	81.3	39.4	29.4	58.7	133.3	22.7
w/o ℓ_M	77.4	36.8	28.2	57.2	120.0	20.3	80.9	39.5	29.0	58.4	132.8	22.5
w/o ℓ_S	77.8	37.3	28.8	58.0	121.7	21.8	81.5	40.1	29.6	59.1	133.5	22.7
Hard μ	77.7	37.2	28.7	57.7	121.0	21.7	81.4	39.5	29.5	58.9	133.5	22.8
SET	**80.0**	**37.8**	**29.2**	**58.4**	**122.6**	**22.0**	**81.8**	**40.5**	**29.7**	**59.5**	**134.4**	23.1

4.3. Experimental Results

4.4. Results on MS-COCO dataset

Table 1 presents the quantitative comparison results on the MS-COCO dataset with other methods. For fairness, all the models are first trained under cross-entropy loss and then optimized for CIDEr-D score as Huang et al. (2019). "-" represents the results that have not been given in the raw paper. The results reveal that: 1) The captioning models that consider the perceptive words are competitive to or worse than transformer-based models. There are two possible reasons: a) most sentences in the MS-COCO dataset are usually a brief introduction to an event, without considering the visual perception. Thereby, there is not much difference between transformer-based methods and deep captioning models considering perceptive words. b) SGAE even performs worse, because the backbones of the encoder and decoder of SGAE are not transformer-based structures, thereby the modeling ability is limited. 2) SET performs the best on most criteria considering various optimization, except the SPICE on CIDEr-D Score Optimization. This phenomenon validates that the multi-task-based pre-training and knowledge-enhanced module can not only promote the learning of perceptive semantics but also affect the encoding of common semantics for visual representation, which verifies the generality of SET. 3) The performance improvement of SET on the MS-COCO dataset is lower than the DecorationCap dataset. For this reason, most sentences in the MS-COCO dataset are usually a simple introduction to an event without considering visual perception. Meanwhile, the results of the ablation study have a similar phenomenon with DecorationCap datasets.

Table 2: The performances of various methods on DecorationCap.

	Cross-Entropy Loss								**CIDEr-D Score Optimization**							
	B@1	B@2	B@3	B@4	M	R	C	S	B@1	B@2	B@3	B@4	M	R	C	S
SCST	4.9	2.4	1.2	0.6	3.9	9.3	4.1	1.2	8.1	4.0	1.9	0.9	5.1	11.1	5.6	1.7
Up-Down	5.4	2.7	1.8	0.8	4.0	9.8	4.3	1.5	8.6	4.1	1.9	1.2	5.3	11.9	5.7	1.9
AoANet	9.1	4.3	2.1	1.1	4.3	10.4	5.3	2.3	10.2	4.5	2.5	1.7	6.4	12.5	6.2	2.8
ORT	9.3	4.3	2.4	1.3	4.7	10.3	5.4	2.4	10.5	4.5	2.7	1.7	6.5	12.3	6.5	2.9
M^2 Transformer	11.7	5.8	3.0	1.6	6.5	12.2	5.5	2.7	15.7	6.6	3.2	1.7	7.2	13.1	6.8	3.2
X-Transformer	11.7	5.9	3.2	1.7	6.8	12.7	5.9	2.9	15.4	6.6	3.3	1.9	7.5	13.5	7.1	3.6
DLCT	12.5	6.4	3.5	2.2	7.0	13.4	6.5	3.3	16.2	7.1	3.5	2.5	7.6	14.0	7.8	3.9
SGAE	9.5	4.5	2.6	1.5	4.7	10.9	5.7	3.2	11.0	4.8	3.0	2.0	6.8	12.7	6.7	3.5
RSTNet	13.1	6.6	3.5	2.5	7.0	13.6	7.1	3.7	16.5	7.5	3.7	2.7	7.9	14.8	8.1	4.5
Transformer based	9.0	3.5	2.0	1.2	4.2	10.2	5.1	2.3	10.1	4.0	2.4	1.5	6.1	12.3	6.0	2.7
w/o Pre-training	8.7	3.2	2.0	1.1	4.0	9.8	5.0	2.0	9.4	3.6	3.3	1.4	5.7	12.0	5.9	2.4
w/o Pre-training+	10.4	4.6	3.6	1.9	4.8	11.2	7.6	4.1	12.5	6.3	4.1	1.8	6.5	13.1	9.5	4.8
w/o KEM	12.8	7.5	4.7	2.3	5.7	13.5	9.5	4.0	14.9	8.5	5.3	2.7	7.2	13.7	11.5	4.5
w/o ℓ_M	13.4	7.0	4.2	2.3	5.5	12.4	7.2	3.9	15.5	8.2	6.0	2.4	7.5	15.0	10.0	4.3
w/o ℓ_S	15.9	9.7	6.0	3.6	7.9	17.5	12.0	4.5	17.3	10.0	6.3	3.9	8.3	18.4	14.0	5.1
Hard μ	14.1	9.0	5.7	3.3	6.8	15.4	10.3	4.2	16.2	9.0	6.0	3.5	7.7	16.5	12.4	5.0
SET	**18.2**	**10.7**	**6.9**	**4.5**	**8.5**	**18.9**	**13.4**	**5.1**	**18.9**	**11.4**	**7.2**	**4.8**	**9.2**	**19.7**	**14.9**	**5.7**

Specific Domain DecorationCap dataset. Table 2 presents the quantitative comparison results on the DecorationCap dataset with other methods. We use the code given in the original paper to retrain the models with the DecorationCap dataset. For fairness, all the models are also first trained under cross-entropy loss and then optimized for CIDEr-D score as Huang et al. (2019). From the results, we find that: 1) all methods have severe performance degradation on the DecorationCap dataset, for the reason that the sentences in the DecorationCap dataset are longer and more complex than the MS-COCO dataset, affecting the generations of captions. 2) Transformer-based approaches also perform better than traditional deep supervised captioning methods, which validates the effectiveness of the Transformer. 3) The captioning models that consider the perceptive words are competitive with transformer-based models, especially the RSTNet performs better than transformer-based approaches. 3) SET acquires more obvious advantages compared with state-of-the-art baselines. The phenomenon validates that SET can well model the perceptive words.

Ablation Study. The bottom of Table 2 record the results of ablation study. The results reveal that: 1) "w/o Pre-training+" performs better than the "w/o Pre-training" and "Transformer based", which validates the effectiveness of multi-modal transformer pre-training and knowledge-enhanced module. 2) "w/o KEM" performs better than "w/o Pre-training" and "Transformer based", which validates that the pre-training encoder is vital for the SET framework. Because multi-modal pre-training can not only provide a betterinitialized encoder but also enhance the quality of the constructed knowledge graph. 3) “Hard μ” performs worse than soft μ, which validates the effectiveness of soft adaptive fusion. 4) SET performs the best compared with other variants, which verifies that each module in SET can contribute to the modeling.

4.5. Diversity and Distinctiveness Properties

We conduct more experiments with the properties used in Wang et al. (2020, 2022). According to the original setting, we set the captioning number N=1 and neighbor number K=5 for the DecorationCap dataset, N=5, and K=5 for the MS-COCO dataset. Note that N=5 in the MS-COCO dataset because each image has 5 captions in MS-COCO. In result, CIDErBtw=72.6/74.8 (i.e., the distinctiveness property), Self-CIDEr=57.1/55.2 (i.e., the diversity property) for SET/DLCT on Decoration, and CIDErBtw =5.2/6.7 (i.e., the distinctiveness property), Self-CIDEr =21.4/13.4 (i.e., the diversity property) for SET/DLCT on MS-COCO. Note that the smaller the distinctiveness metric, the better. The results show that our proposed SET can also generate diverse and distinctive captions compared with state-of-the-art captioning methods.

Table 3: Performance of SET with different values of temperature parameter on MS-COCO.

	Cross-Entropy Loss								CIDEr-D Score Optimization							
	B@1	B@2	B@3	B@4	M	R	C	S	B@1	B@2	B@3	B@4	M	R	C	S
τ=0.1	79.8	61.7	48.1	37.5	29.0	58.3	122.3	21.9	81.7	66.0	51.4	40.0	29.5	59.1	133.7	22.8
τ=0.5	**80.0**	**62.0**	**48.2**	**37.8**	**29.2**	**58.4**	**122.6**	**22.0**	**81.8**	**66.2**	**51.8**	**40.5**	**29.7**	**59.5**	**134.4**	**23.1**
τ=1.0	79.7	61.8	48.0	37.3	28.9	58.1	122.1	21.6	81.5	65.7	51.0	39.6	29.2	58.7	133.1	22.5
τ=2.0	79.5	61.7	47.7	37.2	28.9	58.0	122.0	21.6	81.3	65.4	49.7	39.5	29.2	58.5	132.8	22.4
τ=5.0	79.1	61.5	47.3	37.0	28.5	57.6	121.4	21.3	81.0	65.1	49.2	39.2	29.0	58.1	132.4	22.1

Table 4: Performance of SET with different values of temperature parameter on DecorationCap.

	Cross-Entropy Loss								CIDEr-D Score Optimization							
	B@1	B@2	B@3	B@4	M	R	C	S	B@1	B@2	B@3	B@4	M	R	C	S
τ=0.1	18.0	10.4	6.8	4.4	8.1	18.4	13.0	5.0	18.6	11.2	7.1	4.6	9.0	19.3	14.2	5.6
τ=0.5	**18.2**	**10.7**	**6.9**	**4.5**	**8.5**	**18.9**	**13.4**	**5.1**	**18.9**	**11.4**	**7.2**	**4.8**	**9.2**	**19.7**	**14.9**	**5.7**
τ=1.0	18.0	10.2	6.5	4.2	7.9	18.1	12.7	4.8	18.3	11.0	6.8	4.4	8.7	18.7	13.9	5.4
τ=2.0	17.7	10.0	6.1	3.9	7.6	17.9	12.5	4.6	18.0	10.8	6.6	4.3	8.6	18.4	13.7	5.2
τ=5.0	17.3	9.5	5.7	3.5	7.3	17.5	12.1	4.3	17.8	10.3	6.2	4.0	8.3	18.0	13.2	4.9

4.6. Parameter Analysis

Influence of Temperature Parameter. To explore the influence of temperature parameters, i.e., τ, we conduct more experiments. In detail, we tune the τ in $\{0.1, 0.5, 1, 2, 5\}$ and record the results from two datasets respectively in Table 3 and Table 4. The results reveal that the performance of SET increases firstly, and then decreases with the increasing of τ. The reason is that small τ makes the distribution sharper, which can help the model learn from hard negatives Wang and Liu (2021), but if the τ is too small (e.g., $\tau = 0.1$), the model will pay more attention to difficult negatives, which may cause the semantically similar instances far away and affects the performance.

Table 5: Performance of SET with different numbers of retrieved language cues on MS-COCO.

	Cross-Entropy Loss								CIDEr-D Score Optimization							
	B@1	B@2	B@3	B@4	M	R	C	S	B@1	B@2	B@3	B@4	M	R	C	S
O=1	77.9	61.7	47.8	37.5	29.0	58.1	122.3	21.9	81.7	65.7	51.3	40.2	29.6	59.2	134.1	22.9
O=2	77.9	61.8	47.9	37.7	29.1	58.2	122.5	21.9	81.8	65.9	51.5	40.4	29.7	59.4	134.3	22.9
O=3	**80.0**	**62.0**	**48.2**	**37.8**	**29.2**	**58.4**	**122.6**	**22.0**	**81.8**	**66.2**	**51.8**	**40.5**	**29.7**	**59.5**	**134.4**	**23.1**
O=4	77.8	61.5	47.6	37.4	29.0	58.1	122.1	21.8	81.5	65.4	51.0	40.0	29.5	59.0	133.7	22.8
O=5	77.6	61.3	47.2	37.1	28.7	57.8	121.8	21.7	81.2	65.0	50.7	39.7	29.0	58.8	133.4	22.5

Table 6: Performance of SET with different numbers of retrieved language cues on DecorationCap.

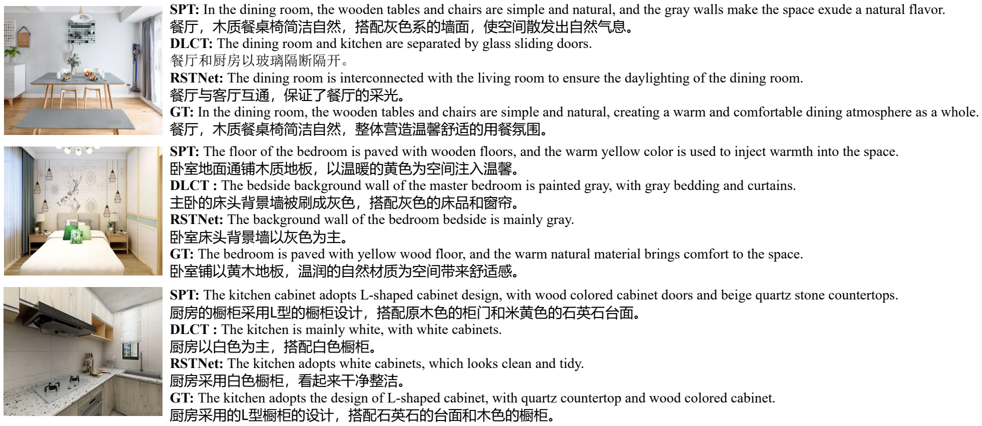
	Cross-Entropy Loss								CIDEr-D Score Optimization							
	B@1	B@2	B@3	B@4	M	R	C	S	B@1	B@2	B@3	B@4	M	R	C	S
O=1	17.6	10.4	6.5	4.0	8.2	18.4	12.9	4.9	18.5	11.1	6.8	4.3	8.8	19.0	14.5	5.3
O=2	17.9	10.5	6.8	4.3	8.3	18.8	13.1	4.9	18.7	11.3	7.0	4.7	9.1	19.4	14.7	5.5
O=3	**18.2**	**10.7**	**6.9**	**4.5**	**8.5**	**18.9**	**13.4**	**5.1**	**18.9**	**11.4**	**7.2**	**4.8**	**9.2**	**19.7**	**14.9**	**5.7**
O=4	17.5	10.3	6.5	3.8	8.0	18.1	12.7	4.7	18.2	11.0	6.7	4.1	8.5	18.7	14.2	5.2
O=5	17.2	10.0	6.1	3.5	7.7	17.8	12.5	4.5	17.9	10.7	6.5	3.9	8.2	18.4	14.0	5.0

Influence of Retrieved Language Cues. We also conduct experiments to validate the influence of retrieved language cues on the two datasets. In detail, we incorporate cues with different numbers (i.e., $O \in \{1, 2, 3, 4, 5\}$) to empirically investigate the impact on generation. Table 5 and Table 6 depict the results, which reveal that the performance of SET increases firstly, and then decreases on various criteria. The reason may be that more cues can even bring noise.

4.7. Case Study

To explore the effectiveness of SET on caption generation, we provide examples of the DecorationCap dataset compared with other methods, i.e., DLCT and RSTNet. “GT” denotes the human-annotated ground-truth. We also provide the English description for convenient reading. Using the first case in Figure 3 as an example, the DLCT (i.e., Transformer-based method) only describes the placement of dining room objects. The RSTNet can improve the generation of perceptive words, but the description is inaccurate and misses the point. In contrast, SET can not only accurately capture the layouts in the dining room, e.g., “tables”, “chairs”, “wall”, etc, but also describe the overall understanding and feeling with suitable perceptive words, e.g., “wooden tables”, “simple and natural”, and “gray walls make the space feel natural”, which can well match human-annotated ground-truth.

Moreover, we provide more visualizations to validate the effectiveness of the knowledge-enhanced module. In detail, we provide the top retrieved knowledge cues from the con-



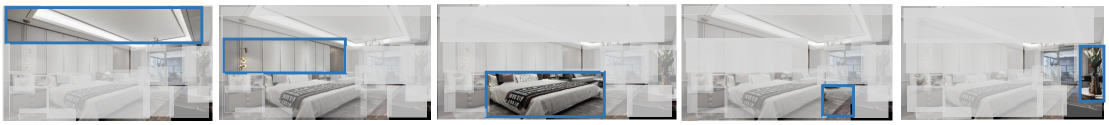
SPT: In the dining room, the wooden tables and chairs are simple and natural, and the gray walls make the space exude a natural flavor.
餐厅，木质餐桌椅简洁自然，搭配灰色系的墙面，使空间散发出自然气息。
DLCT: The dining room and kitchen are separated by glass sliding doors.
餐厅和厨房以玻璃隔断隔开。
RSTNet: The dining room is interconnected with the living room to ensure the daylighting of the dining room.
餐厅与客厅互通，保证了餐厅的采光。
GT: In the dining room, the wooden tables and chairs are simple and natural, creating a warm and comfortable dining atmosphere as a whole.
餐厅，木质餐桌椅简洁自然，整体营造温馨舒适的用餐氛围。

SPT: The floor of the bedroom is paved with wooden floors, and the warm yellow color is used to inject warmth into the space.
卧室地面通铺木质地板，以温暖的黄色为空间注入温馨。
DLCT : The bedside background wall of the master bedroom is painted gray, with gray bedding and curtains.
主卧的床头背景墙被刷成灰色，搭配灰色的床品和窗帘。
RSTNet: The background wall of the bedroom bedside is mainly gray.
卧室床头背景墙以灰色为主。
GT: The bedroom is paved with yellow wood floor, and the warm natural material brings comfort to the space.
卧室铺以黄木地板，温润的自然材质为空间带来舒适感。

SPT: The kitchen cabinet adopts L-shaped cabinet design, with wood colored cabinet doors and beige quartz stone countertops.
厨房的橱柜采用L型的橱柜设计，搭配原木色的柜门和米黄色的石英石台面。
DLCT : The kitchen is mainly white, with white cabinets.
厨房以白色为主，搭配白色橱柜。
RSTNet: The kitchen adopts white cabinets, which looks clean and tidy.
厨房采用白色橱柜，看起来干净整洁。
GT: The kitchen adopts the design of L-shaped cabinet, with quartz countertop and wood colored cabinet.
厨房采用的L型橱柜的设计，搭配石英石的台面和木色的橱柜。

Figure 3: Examples of captions generated by SET and baselines on DecorationCap dataset, GT denotes ground-truth.

structed knowledge graph according to the $sc(\bar{\mathbf{v}}, \mathbf{e}^o)$, the results are recorded in Figure 4. The blue box region indicates the image region with the highest attention.



Top 3 retrieval words:	Top 3 retrieval words:	Top 3 retrieval words:	Top 3 retrieval words:	Top 3 retrieval words:
简洁的 (Simple) 0.3478	灰色的 (Grey) 0.3387	舒服的 (Comfortable) 0.3382	深灰的 (Dark grey) 0.3405	清新的 (Clear) 0.3489
现代的 (Modern) 0.3378	简洁的 (Simple) 0.3316	舒适的 (Comfortable) 0.3365	现代的 (Modern) 0.3347	精致的 (Exquisite) 0.3283
大气的 (Generous) 0.3144	现代的 (Modern) 0.3297	灰色的 (Grey) 0.3253	优雅的 (Elegant) 0.3248	绿色的 (Green) 0.3228

Figure 4: The visualization of image region retrieval perceptive words. We outline the image region with the maximum attention weight in blue and retrieve the top-3 perceptive words with the largest similarity.

5. Conclusion

In this paper, we focus on the decoration display task, which plays an important role in online housing services. To solve the defect of exploring perceptive words when transferring existing captioning approaches to the decoration display task, we propose a self-enhanced deep captioning model, which generates the captions with visual perception using the Self-Enhanced Transformer (SET). In detail, SET first pre-trained a scene-aware encoder with a multi-modal transformer, which aimed to preliminarily enhance the perceptive semantics of the visual representations. Then, SET combines the pre-trained encoder with a transformer decoder for fine-tuning and designs a knowledge-enhanced module on the top of the decoder to adaptively fuse the decoded representations and retrieved language cues for word prediction. In experiments, we validate the effectiveness of SET on both the public dataset and a specific domain dataset DecorationCap.

6. Acknowledgments

This work is partially supported by National Key RD Program of China (2022YFF0712100), NSFC (62276131), Natural Science Foundation of Jiangsu Province of China under Grant (BK20240081), the Fundamental Research Funds for the Central Universities (No.30922010317, No.30923011007).

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, Utah, US, 2018.

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016.

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. N-LTP: A open-source neural chinese language technology platform with pretrained models. *CoRR*, abs/2009.11616, 2020.

Jia Chen, Jiaxin Mao, Yiqun Liu, Ziyi Ye, Weizhi Ma, Chao Wang, Min Zhang, and Shaoping Ma. A hybrid framework for session context modeling. *ACM Trans. Inf. Syst.*, 39: 30:1–30:35, 2021.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, pages 10575–10584, Seattle, WA, 2020.

Zhongtian Fu, Kefei Song, Luping Zhou, and Yang Yang. Noise-aware image captioning with progressively exploring mismatched words. In *AAAI*, pages 12091–12099, Vancouver, Canada, 2024.

Ankush Gupta, Yashaswi Verma, and C. V. Jawahar. Choosing linguistics over vision to describe images. In *AAAI*, pages 606–612, Toronto, Ontario, 2012.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, Las Vegas, US, 2016.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. In *NeurIPS*, pages 11135–11145, British Columbia, UK, 2019.

Lun Huang, Wenmin Wang, Jie Chen, and Xiaoyong Wei. Attention on attention for image captioning. In *ICCV*, pages 4633–4642, Seoul, Korea, 2019.

Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Gen Luo, Yongjian Wu, Yue Gao, and Rongrong Ji. Improving image captioning by leveraging intra- and inter-layer global representation in transformer network. In *AAAI*, pages 1655–1663, Virtual, 2021.

Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. Collective generation of natural image descriptions. In *ACL*, pages 359–368, Jeju Island, Korea, 2012.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, pages 9694–9705, Virtual, 2021.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, pages 121–137, Glasgow, Scotland, 2020.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, Zurich, Switzerland, 2014.

Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. Dual-level collaborative transformer for image captioning. In *AAAI*, pages 2286–2293, Virtual Event, 2021.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *CVPR*, pages 10968–10977, Washington, US, 2020.

Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. Smallcap: Lightweight image captioning prompted with retrieval augmentation. In *CVPR*, pages 2840–2849, Vancouver, BC, 2023.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, Montreal, Canada, 2015.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, pages 1179–1195, Honolulu, HI, 2017.

Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *CoRR*, abs/2005.04757, 2020.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: pre-training of generic visual-linguistic representations. In *ICLR*, Addis Ababa, Ethiopia, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NuerIPS*, pages 5998–6008, California, US, 2017.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, Massachusetts, US, 2015.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *CVPR*, pages 2495–2504, Virtual, 2021.

Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B. Chan. Compare and reweight: Distinctive image captioning using similar images sets. In *ECCV*, pages 370–386, Glasgow, UK, 2020.

Qingzhong Wang, Jia Wan, and Antoni B. Chan. On diversity in image captioning: Metrics and methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44:1035–1049, 2022.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, pages 10685–10694, California, US, 2019.

Yang Yang, Hongchen Wei, Hengshu Zhu, Dianhai Yu, Hui Xiong, and Jian Yang. Exploiting cross-modal prediction and relation consistency for semi-supervised image captioning. *IEEE Transactions on Cybernetics*, 54:890–902, 2022.

Yang Yang, Ran Bao, Weili Guo, De-Chuan Zhan, Yilong Yin, and Jian Yang. Deep visual-linguistic fusion network considering cross-modal inconsistency for rumor detection. *Sci. China Inf. Sci.*, 66:222102, 2023a.

Yang Yang, Yurui Huang, Weili Guo, Baohua Xu, and Dingyin Xia. Towards global video scene segmentation with context-aware transformer. In *AAAI*, pages 3206–3213, Washington, DC, 2023b.

Yang Yang, Jinyi Guo, Guangyu Li, Lanyu Li, Wenjie Li, and Jian Yang. Alignment efficient image-sentence retrieval considering transferable cross-modal representation learning. *Frontiers Comput. Sci.*, 18:181335, 2024.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *AAAI*, pages 3208–3216, Virtual, 2021.

Ningyu Zhang, Qianghuai Jia, Shumin Deng, Xiang Chen, Hongbin Ye, Hui Chen, Huaixiao Tou, Gang Huang, Zhao Wang, Nengwei Hua, and Huajun Chen. Alicg: Fine-grained and evolvable conceptual graph construction for semantic search at alibaba. In *KDD*, pages 3895–3905, Virtual, 2021a.

Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *CVPR*, pages 15465–15474, Virtual, 2021b.