

# Step-Wise Hierarchical Alignment Network for Image-Text Matching

Zhong Ji, Kexin Chen, Haoran Wang\*

School of Electrical and Information Engineering, Tianjin University, Tianjin, China

{jizhong, kxchen, haoranwang}@tju.edu.cn

## Abstract

Image-text matching plays a central role in bridging the semantic gap between vision and language. The key point to achieve precise visual-semantic alignment lies in capturing the fine-grained cross-modal correspondence between image and text. Most previous methods rely on single-step reasoning to discover the visual-semantic interactions, which lacks the ability of exploiting the multi-level information to locate the hierarchical fine-grained relevance. Different from them, in this work, we propose a **step-wise hierarchical alignment network** (SHAN) that decomposes image-text matching into multi-step cross-modal reasoning process. Specifically, we first achieve local-to-local alignment at fragment level, following by performing global-to-local and global-to-global alignment at context level sequentially. This progressive alignment strategy supplies our model with more complementary and sufficient semantic clues to understand the hierarchical correlations between image and text. The experimental results on two benchmark datasets demonstrate the superiority of our proposed method.

## 1 Introduction

Associating vision with language and exploring the correlations between them have attracted broad attention in the field of Artificial Intelligence during last decades. Thanks to the rapid development of deep learning [Krizhevsky *et al.*, 2012], plenty of areas have made tremendous progress in effectively bridging vision and language, such as visual question answering (VQA) [Antol *et al.*, 2015; Anderson *et al.*, 2018], image captioning [Zhang *et al.*, 2019; Feng *et al.*, 2019], image-text matching [Song and Soleymani, 2019; Wang *et al.*, 2020b], visual grounding [Wang *et al.*, 2020c], and zero-shot learning [Yu *et al.*, 2020]. In this paper, we focus on the task of image-text matching, which aims to provide flexible retrieval across samples from different modalities. It remains challenging since it is difficult for machines to understand the heterogeneous contents in images and texts so

\*indicates corresponding author

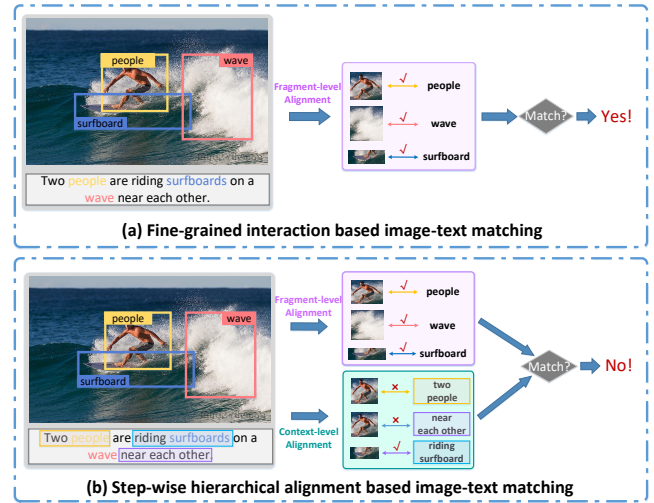


Figure 1: Illustration of the hard negative example which may cause mismatching by current fine-grained interaction learning methods, while our proposed step-wise hierarchical alignment network can not only explore the fragment-level alignment, but also take the context-level information into consideration so as to discriminate the hard negative pairs more accurately. Note that the displayed image-text pair is mismatched.

as to accurately infer the cross-modal similarity, especially when encountering hard negative example pairs which share similar semantics but convey slight differences (As shown in Fig.1).

To address this problem, early works [Frome *et al.*, 2013; Faghri *et al.*, 2017] usually map the whole images and texts into a shared embedding space, in which the cross-modal relevance can be directly measured by calculating the distance between heterogeneous samples. However, these global embedding based learning methods ignore that a visual or textual instance is actually composed of significant local fragments (e.g., salient regions in image and key words in text) and insignificant ones. Therefore, represent them from a global view may not fully exploit the fine-grained details in fragment-level and introduce some noisy background information. Based on this observation, another stride of approaches [Karpathy and Fei-Fei, 2015; Lee *et al.*, 2018] have been proposed to mitigate this issue by concentrating on the

fine-grained interactions between image regions and textual words. These fine-grained interaction based learning methods align image and text by aggregating the local similarities of all available region-word pairs. Afterwards, recent works resort to attention mechanism [Vaswani *et al.*, 2017; Nam *et al.*, 2017; Lee *et al.*, 2018; Hu *et al.*, 2019] to pinpoint the discriminative parts of fragment-level representations, which can neatly align image and text via discovering more fine-grained cross-modal correspondence and further leads to promising matching performance.

However, most existing fine-grained interaction learning based methods tend to discriminate image-text pairs according to the apparently distinct and distinguishable portions, which may consequently fail to distinguish those hard negative examples with similar semantic contents but slightly different context information. For example, as shown in Fig.1 (a), previous fine-grained interaction based methods may probably assign high matching score to the combination of the given image and the sentence “Two people are riding surfboards on a wave near each other.”, due to they both contain the salient foreground objects, *e.g.* “people”, “surfboard” and “wave”. The failure may be attributed to the inability of the previous methods, which only rely on single-step inference whilst pay less attention to the global context information that can provide some high-level clues between semantically similar samples. By contrast, when people want to precisely associate an image with a specific text, we typically not only attend to the fragment-level entities (“people”, “surfboard”, “wave”), but also leverage the context-level semantic information, *e.g.*, attributes (“two”) and relationships (“near each other”), as complementary hints to capture the complicated cross-modal correspondence. Accordingly, these observations motivate us to explore multi-step visual-semantic alignments for cross-modal correlation learning.

To circumvent the aforementioned issues, in this work, we propose a **Step-wise Hierarchical Alignment Network (SHAN)** for image-text matching, which performs hierarchical cross-modal alignment at two semantic levels, *i.e.*, **fragment-level alignment** and **context-level alignment**. Specifically, we measure the relevance of an image-text pair by performing three steps of alignments progressively. First of all, we employ cross-attention mechanism to align image regions with textual words to achieve fragment-level local-to-local (L2L) alignment, followed by adopting an adaptive cross-modal feature fusion strategy to produce the visual and textual global context representations by fully exploiting the intermediate information produced by last alignment process. Then, based on the generated context representation, we perform global-to-local (G2L) alignment to measure the cross-modal similarities at context-level. Finally, we further perform a global-to-global (G2G) cross-modal alignment by leveraging more context-level information. The main highlights of our proposed method are summarized as follows:

1. We propose a novel Step-wise Hierarchical Alignment Network to progressively explore the visual-semantic relevance for image-text matching, which leverages the multi-step hierarchical alignment to provide more com-

plementary cues to capture the fine-grained cross-modal semantic correlation.

2. We conduct experiments on two public datasets: *e.g.* Flickr30k and MS-COCO, and the quantitative experimental results validate that our model can achieve state-of-the-art performance on both datasets. Detailed visualization on L2L and G2L alignment step also qualitatively verify the effectiveness of our proposed method.

## 2 Related Work

A stride of typical solutions for cross-modal image-text matching [Kiros *et al.*, 2014; Ma *et al.*, 2015; Faghri *et al.*, 2017; Wang *et al.*, 2020a] can be classified as **global embedding based image-text matching**, which learns a joint embedding space for these heterogeneous data samples, so that the similarity of a given image-text pair can be directly compared. For example, [Kiros *et al.*, 2014] employed a convolution neural network (CNN) and an recurrent neural network (RNN) to extract the global representations for image and text respectively and utilized hinge-based triplet ranking loss to align these heterogeneous data. [Faghri *et al.*, 2017] further modified the ranking loss with hard negative mining and yielded encouraging cross-modal retrieval performance improvement.

Another line of studies [Lee *et al.*, 2018; Hu *et al.*, 2019; Wang *et al.*, 2019] can be summarized as **fine-grained interaction based image-text matching**, which focuses on exploring the fine-grained interactions between images and texts to obtain a deeper understanding of cross-modal correspondence. [Karpathy and Fei-Fei, 2015] developed a unified deep model to infer the latent alignment between textual segments and the implicit local regions of the image that they depict, and the global image-text similarity is finally calculated by aggregating all the region-word similarities. [Chen *et al.*, 2020] proposed an iterative matching scheme to capture the correspondences between image regions and textual words with multiple steps of alignments, which is mostly relevant to our work. However, they paid less attention to explore the context-level information in images and texts. By contrast, we perform visual-semantic alignment from both perspectives including fragment-level and context-level.

## 3 Our Method

We construct a Step-wise Hierarchical Alignment Network (SHAN) to progressively explore the semantic alignments between image and text. The overall framework is illustrated in Fig.2 (a). Next, we will introduce the details of our proposed SHAN that contains following three parts: 1) Features representation for image and text; 2) The implementation details of designed Step-wise Hierarchical Alignment model and 3) Objective function for model learning.

### 3.1 Feature Representation

#### Image Representation

For image feature extraction, we adopt an off-the-shelf object detection model (*i.e.* Faster R-CNN) which is pre-trained on Visual Genome dataset to detect  $k$  salient objects  $\mathbf{O} =$

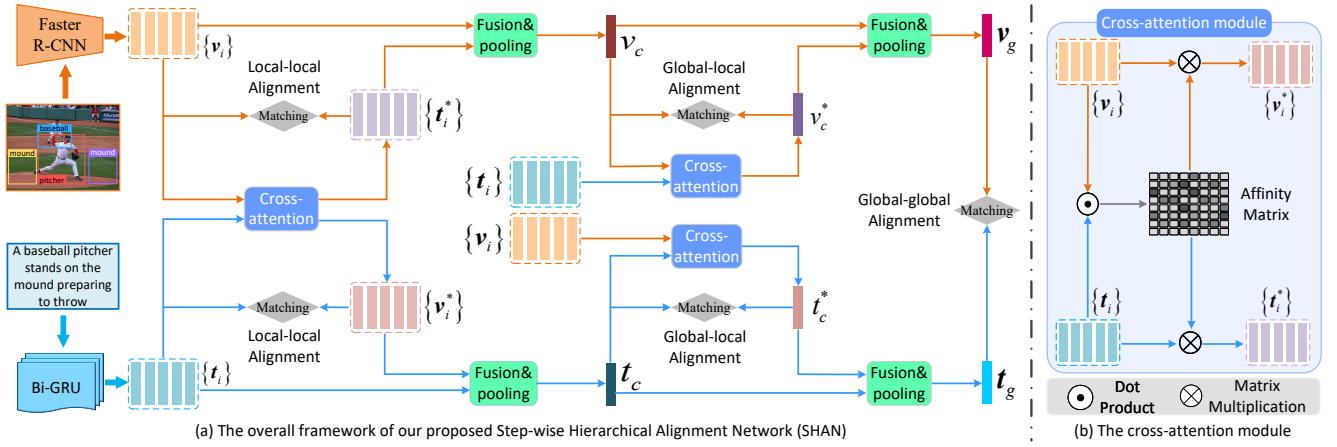


Figure 2: The framework of our proposed Step-wise Hierarchical Alignment Network (SHAN).

$\{\mathbf{o}_1, \dots, \mathbf{o}_k\} \in \mathbb{R}^{k \times 2048}$  in an image, then a fully-connect layer is followed to project them into  $D$ -dimension vectors:

$$\mathbf{v}_i = \mathbf{W}_v \mathbf{o}_i + b_v. \quad (1)$$

In our experiments, the number of extracted salient regions for each image is set to 36.

### Text Representation

We extract the word-level features  $\mathbf{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_n\}$  to represent a given text, where  $n$  is the number of words in the given sentence. Concretely, We first apply one-hot encoding on input words based on the whole vocabulary of the training dataset, then leverage a pre-trained word embedding model Glove and a learnable embedding layer to transform one-hot codes into fixed vectors and updatable vectors respectively. We then concatenate the two vectors to obtain the final word representations  $\mathbf{w}_j$ .

Next, we employ a bidirectional GRU to enhance the word-level representations by capturing the context information in the sentence from both forward and backward directions:

$$\begin{aligned} \vec{\mathbf{h}}_j &= \overrightarrow{GRU}(\mathbf{w}_j, \vec{\mathbf{h}}_{j-1}), \\ \overleftarrow{\mathbf{h}}_j &= \overleftarrow{GRU}(\mathbf{w}_j, \overleftarrow{\mathbf{h}}_{j+1}), \end{aligned} \quad (2)$$

where  $\vec{\mathbf{h}}_j$  and  $\overleftarrow{\mathbf{h}}_j$  denote the GRU hidden states for forward and backward directions respectively, note that we set the hidden dimension of GRU as  $D$ , therefore finally we can get word features  $\mathbf{T} = \{\mathbf{t}_1, \dots, \mathbf{t}_n\} \in \mathbb{R}^{n \times D}$  by averaging  $\vec{\mathbf{h}}_j$  and  $\overleftarrow{\mathbf{h}}_j$ :  $\mathbf{t}_j = \frac{\vec{\mathbf{h}}_j + \overleftarrow{\mathbf{h}}_j}{2}$ .

## 3.2 Step-wise Hierarchical Alignment

### Fragment-level L2L Alignment

In the first step, we carry out fine-grained fragment-level local-to-local matching between visual regions and textual words based on bidirectional cross-attention mechanism (see Fig.2 (b)). Mathematically, we first calculate the region-word affinity matrix as:

$$\mathbf{A} = (\tilde{\mathbf{W}}_v \mathbf{V})(\tilde{\mathbf{W}}_t \mathbf{T})^T, \quad (3)$$

where  $\mathbf{A} \in \mathbb{R}^{k \times n}$  is the region-word affinity matrix and  $\mathbf{A}_{ij}$  denotes the semantic similarity of the  $i$ -th region and the  $j$ -th word.

For region-queried cross-attention, each regional feature is used as a query to assign weights over the  $n$  words, then we can construct an **attended text-level representation**  $\mathbf{t}_i^*$  based on region feature  $\mathbf{v}_i$  by applying weighted combination over the  $n$  words, such process can be formulated as follows:

$$\mathbf{t}_i^* = \sum_{j=1}^n \alpha_{ij} \mathbf{t}_j, \quad \alpha_{ij} = \frac{\exp(\lambda \mathbf{A}_{ij})}{\sum_{j=1}^n \exp(\lambda \mathbf{A}_{ij})}, \quad (4)$$

where  $\lambda$  is the temperature parameter of the softmax function. Similarly, we can obtain word **attended image-level representations** by regarding each word feature as query:

$$\mathbf{v}_j^* = \sum_{i=1}^k \beta_{ij} \mathbf{v}_i, \quad \beta_{ij} = \frac{\exp(\lambda \mathbf{A}_{ij})}{\sum_{i=1}^k \exp(\lambda \mathbf{A}_{ij})}. \quad (5)$$

We then define the region-related L2L matching scores and word-related L2L matching scores as follows:

$$\begin{aligned} R_v(\mathbf{v}_i, \mathbf{t}_i^*) &= \frac{\mathbf{v}_i \mathbf{t}_i^{*T}}{\|\mathbf{v}_i\| \|\mathbf{t}_i^*\|}, i \in \{1, \dots, k\}, \\ R_t(\mathbf{t}_j, \mathbf{v}_j^*) &= \frac{\mathbf{t}_j \mathbf{v}_j^{*T}}{\|\mathbf{t}_j\| \|\mathbf{v}_j^*\|}, j \in \{1, \dots, n\}. \end{aligned} \quad (6)$$

Finally, the L2L matching score for the target image  $I$  and text  $T$  is obtained through taking average of all region-based matching scores and word-based matching scores as follows:

$$S_{L2L}(I, T) = \mu_1 \sum_{i=1}^k R_v(\mathbf{v}_i, \mathbf{t}_i^*) + (1 - \mu_1) \sum_{j=1}^n R_t(\mathbf{t}_j, \mathbf{v}_j^*), \quad (7)$$

where  $\mu_1$  is a hyper-parameter to balance the contributions of region-related L2L relevance and word-related L2L relevance.

### Context-level G2L Alignment

At the second step of our SHAN model, we implement a context-level global-to-local alignment mechanism to explore a deeper visual-semantic interaction. This alignment module

plays a key role in exploiting the context-level information as complementary clues to understand the semantic correlations between image and text. The global representations for image and text in our model are obtained by implementing **fusion and pooling operation** on original object/word features and they are both enhanced by cross-modal attention. Specifically, we regard the region-attended text-level representations  $\mathbf{T}^* = \{\mathbf{t}_1^*, \dots, \mathbf{t}_k^*\}$  as the aggregated message passed from textual to visual modality, and the word-attended image-level representations  $\mathbf{V}^* = \{\mathbf{v}_1^*, \dots, \mathbf{v}_n^*\}$  as the aggregated message passed from visual to textual modality. Then we fuse the original fragment features with aggregated message from another modality to obtain the global context representation.

Taking the visual regional features  $\mathbf{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_k\}$  and region-attended text-level representations  $\mathbf{T}^* = \{\mathbf{t}_1^*, \dots, \mathbf{t}_k^*\}$  as an example. We first perform **element-wise fusion operation** in a dynamic way as follows:

$$\begin{aligned} \mathbf{v}_i^c &= g * \mathbf{v}_i + (1 - g) * \mathbf{t}_i^*, \\ g &= \text{sigmoid}(\mathbf{W}_{gcat}(\mathbf{v}_i, \mathbf{t}_i^*) + b_g), \end{aligned} \quad (8)$$

where  $\mathbf{v}_i^c$  denotes the contextual information provided by the  $i$ -th region  $\mathbf{v}_i$  and its attended semantic knowledge  $\mathbf{t}_i^*$  from textual modality,  $g \in [0, 1]$  is a gate value to adaptively balance the importance of fused information from  $\mathbf{v}_i$  and  $\mathbf{t}_i^*$ ,  $cat()$  denotes concatenation on  $\mathbf{v}_i$  and  $\mathbf{t}_i^*$ . Empirically, we found that such an adaptive fusion strategy outperforms the traditional average pooling operation.

For the  $k$  regions in an image, we can acquire  $k$  fused visual contextual representations  $\mathbf{V}^c = \{\mathbf{v}_1^c, \dots, \mathbf{v}_k^c\} \in \mathbb{R}^{k \times D}$ . Similarly, for the  $n$  words in a sentence,  $n$  fused textual contextual representations  $\mathbf{T}^c = \{\mathbf{t}_1^c, \dots, \mathbf{t}_n^c\} \in \mathbb{R}^{n \times D}$  are also obtained. Then we use a simple **self-attention mechanism** to aggregate these separate features of  $k$  regions and  $n$  words into one global vector to represent the whole image and the whole sentence:

$$\begin{aligned} \gamma_v &= \text{softmax}(\hat{\mathbf{W}}_v \mathbf{V}^{cT}), \quad \mathbf{v}_c = \gamma_v \mathbf{V}^c, \\ \gamma_t &= \text{softmax}(\hat{\mathbf{W}}_t \mathbf{T}^{cT}), \quad \mathbf{t}_c = \gamma_t \mathbf{T}^c, \end{aligned} \quad (9)$$

where  $\gamma_v \in \mathbb{R}^{1 \times k}$  and  $\gamma_t \in \mathbb{R}^{1 \times n}$  denote the weights for convex combination generated by two linear projection layers respectively,  $\mathbf{v}_c \in \mathbb{R}^{1 \times D}$  is the global context representation for image  $I$ , and  $\mathbf{t}_c \in \mathbb{R}^{1 \times D}$  is the global context representation for text  $T$ .

Based on the generated visual and textual global context representations, similar to L2L alignment module, we then perform G2L alignment by resorting to bidirectional cross-attention mechanism. Specifically, for visual global-queried cross-modal attention,  $\mathbf{v}_c$  is used as an anchor vector to assign weights over  $n$  different words according to the semantic similarity between  $\mathbf{v}_c$  and  $n$  word features, then we use a convex combination to aggregate textual word features into a visual global-attended text-level representation:

$$\alpha_j = \frac{\exp(\text{sim}(\mathbf{v}_c, \mathbf{t}_j))}{\sum_{j=1}^n \exp(\text{sim}(\mathbf{v}_c, \mathbf{t}_j))}, \quad \mathbf{t}_c^* = \sum_{j=1}^n \alpha_j \mathbf{t}_j, \quad (10)$$

where  $\text{sim}(\cdot, \cdot)$  denotes the distance function to measure the similarity of the two vectors, and  $\text{sim}()$  represents the Cosine

distance function. Similarly, for textual global-queried cross-attention, we can obtain the textual global-attended image-level representation as follows:

$$\beta_i = \frac{\exp(\text{sim}(\mathbf{t}_c, \mathbf{v}_i))}{\sum_{i=1}^k \exp(\text{sim}(\mathbf{t}_c, \mathbf{v}_i))}, \quad \mathbf{v}_c^* = \sum_{i=1}^k \beta_i \mathbf{v}_i. \quad (11)$$

Then the visual global-related G2L matching score and textual global-related G2L matching score are derived as follows:

$$R_v(\mathbf{v}_c, \mathbf{t}_c^*) = \frac{\mathbf{v}_c \mathbf{t}_c^{*T}}{\|\mathbf{v}_c\| \|\mathbf{t}_c^*\|}, \quad R_t(\mathbf{t}_c, \mathbf{v}_c^*) = \frac{\mathbf{t}_c \mathbf{v}_c^{*T}}{\|\mathbf{t}_c\| \|\mathbf{v}_c^*\|}. \quad (12)$$

Afterwards, we define the final G2L matching score between image  $I$  and text  $T$  by summing up the two matching scores:

$$S_{G2L}(I, T) = \mu_2 \times R_v(\mathbf{v}_c, \mathbf{t}_c^*) + (1 - \mu_2) \times R_t(\mathbf{t}_c, \mathbf{v}_c^*), \quad (13)$$

where  $\mu_2$  is a hyper-parameter to balance the contributions of the two matching scores.

### Context-level G2G Alignment

On the top level of our SHAN model, we aim to further align the image and text from a global perspective. Specifically, the context-level global representations for image and text are obtained via fusing the intermediate semantic information produced during G2L alignment module, *i.e.* fusing  $\mathbf{v}_c$  and  $\mathbf{t}_c^*$  to obtain the final visual context-level global representation, and fusing  $\mathbf{t}_c$  and  $\mathbf{v}_c^*$  to obtain the **final textual context-level global representation**:

$$\begin{aligned} \mathbf{v}_g &= \mathbf{v}_c \oplus \mathbf{t}_c^*, \\ \mathbf{t}_g &= \mathbf{t}_c \oplus \mathbf{v}_c^*, \end{aligned} \quad (14)$$

where  $\oplus$  denotes element-wise sum. Our hypothesis is that after L2L and G2L alignment, the generated representations  $\mathbf{v}_c$  and  $\mathbf{t}_c$  have already captured rich implicit information regarding fine-grained visual-textual correlations, such a cross-attention based fusion strategy would further encourage the mutual information propagation between both modalities so as to learn more discriminative context-level features.

Then the overall G2G matching score between image  $I$  and text  $T$  is derived by computing the cosine distance of  $\mathbf{v}_g$  and  $\mathbf{t}_g$ :

$$S_{G2G}(I, T) = \frac{\mathbf{v}_g \mathbf{t}_g^T}{\|\mathbf{v}_g\| \|\mathbf{t}_g\|} \quad (15)$$

### 3.3 Learning Objective

As aforementioned, our model eventually acquire three matching scores (*e.g.*  $S_{L2L}(I, T)$ ,  $S_{G2L}(I, T)$  and  $S_{G2G}(I, T)$ ) at different levels for a given image-text pair  $< I, T >$ , and the final cross-modal similarity can be defined as follows:

$$S(I, T) = S_{L2L}(I, T) + S_{G2L}(I, T) + S_{G2G}(I, T). \quad (16)$$

Then, we train our model with a bidirectional hinge-based triplet ranking loss [Faghri *et al.*, 2017; Wang *et al.*, 2020b] that enforces the similarity of matched image-text pair to be higher than that of unmatched ones.

$$\begin{aligned} L(I, T) &= \max(0, m - S(I, T) + S(I, \hat{T})) \\ &\quad + \max(0, m - S(I, T) + S(\hat{I}, T)), \end{aligned} \quad (17)$$



Method	Text Retrieval			Image Retrieval			Rsum
	R@1	R@5	R@10	R@1	R@5	R@10	
SCAN [Lee <i>et al.</i> , 2018]	67.4	90.3	95.8	48.6	77.7	85.2	465.0
SAEM [Wu <i>et al.</i> , 2019]	69.1	91.0	95.1	52.4	81.1	88.1	477.0
RDAN [Hu <i>et al.</i> , 2019]	68.1	91.0	95.9	54.1	80.9	87.2	477.2
SGM [Wang <i>et al.</i> , 2020d]	71.8	91.7	95.5	53.5	79.6	86.5	478.6
CAAN [Zhang <i>et al.</i> , 2020]	70.1	91.6	<b>97.2</b>	52.8	79.0	87.9	478.6
SHAN-T2I	72.5	92.3	95.8	53.6	78.6	85.5	478.3
SHAN-I2T	70.6	91.7	95.5	50.5	77.1	85.2	470.6
SHAN-full	<b>74.6</b>	<b>93.5</b>	96.9	<b>55.3</b>	<b>81.3</b>	<b>88.4</b>	<b>490.0</b>

Table 1: Experimental results on Flickr30k testing set in terms of Recall@K (R@K).

Method	Text Retrieval			Image Retrieval			Rsum
	R@1	R@5	R@10	R@1	R@5	R@10	
SCAN [Lee <i>et al.</i> , 2018]	72.7	94.8	98.4	58.8	88.4	94.8	507.9
SAEM [Wu <i>et al.</i> , 2019]	71.2	94.1	97.7	57.8	88.6	94.9	504.3
RDAN [Hu <i>et al.</i> , 2019]	74.6	96.2	<b>98.7</b>	61.6	89.2	94.7	515.0
SGM [Wang <i>et al.</i> , 2020d]	73.4	93.8	97.8	57.5	87.3	94.3	504.1
CAAN [Zhang <i>et al.</i> , 2020]	75.5	95.4	98.5	61.3	<b>89.7</b>	95.2	515.6
SHAN-T2I	75.9	96.1	<b>98.7</b>	60.7	88.2	94.2	513.8
SHAN-I2T	73.0	95.8	97.9	58.5	87.3	94.0	506.5
SHAN-full	<b>76.8</b>	<b>96.3</b>	<b>98.7</b>	<b>62.6</b>	89.6	<b>95.8</b>	<b>519.8</b>

Table 2: Experimental results on MS-COCO 1K testing set in terms of Recall@K (R@K).

where  $m$  is a tuning margin parameter.  $\hat{T} = \arg\max_{t \neq T} S(I, t)$  and  $\hat{I} = \arg\max_{i \neq I} S(i, T)$  stand for contrastive negative samples in a mini-batch. The entire model is trained in an end-to-end manner.

## 4 Experiments

### 4.1 Dataset and Evaluation Metric

#### Dataset

Two benchmark datasets are used in our experiments to testify the performance of our method: (1) **Flickr30k** contains 31783 images and each image is annotated with 5 sentences. Following [Karpathy and Fei-Fei, 2015], we split the dataset into 1000 test images, 1000 validation images and 29000 training images. (2) **MS-COCO** is another large-scale image captioning dataset with 123287 images and each image is relative with 5 descriptions. We follow [Lee *et al.*, 2018] to split the dataset into 5000 images for validation, 5000 images for testing and the rest 113287 images for training. We report the matching performance by averaging the 5-folds over 1k test images.

#### Evaluation Metric

We take Recall@K (R@K,  $K = 1, 5, 10$ ) that describes the fraction of ground truth instance being retrieved at the top 1, 5, 10 results as the evaluation metric. Following previous works, we evaluate our approach on both text retrieval (find the most related text given an image) and image retrieval (find the matched image described by certain text) tasks. Additionally, the overall matching metric “Rsum” that sums over all recall values is also used to compare our model with other state-of-the-art methods.

### 4.2 Implementation Details

We present three variants of SHAN model to systematically investigate the effectiveness of our proposed approach: (1) SHAN-T2I applies hierarchical alignment only in text-to-image direction, the global representation for image in G2G alignment module is obtained simply by taking average over extracted regional features. (2) SHAN-I2T applies hierarchical alignment only in image-to-text direction, the global representation for text in G2G alignment module is obtained by taking average over word features. (3) SHAN-full applies hierarchical alignment both in text-to-image and image-to-text direction. We train our model with Adam optimizer for 30 epochs on each dataset. The dimension of joint embedding space for image regions and textual words are set to 1024, and the dimension of word embeddings is set to be 300, other parameters are empirically set as follows:  $\mu_1 = 0.3$ ,  $\mu_2 = 0.5$ ,  $\lambda = 15$  and  $m = 0.2$ .

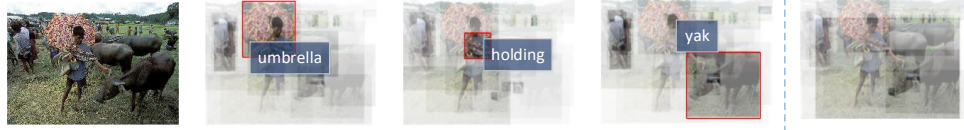
### 4.3 Comparison with State-of-the-art Methods

Table 1 and Table 2 display the performance comparison between our SHAN model and several state-of-the-art methods on Flickr30k and MS-COCO dataset, respectively. We can observe that our model achieves the best retrieval results on both datasets. Specifically, for text retrieval, our SHAN-full achieves a maximal improvement of 2.8% and 1.3% in terms of R@1 score on Flickr30k and MS-COCO, respectively. As for image retrieval, our best model reaches at 55.3% R@1 score and a 62.6% R@1 score on Flickr30k and MS-COCO respectively, which also surpasses other state-of-the-art methods. Similar performance improvement can be seen from other evaluation metrics (e.g. R@5, R@10).

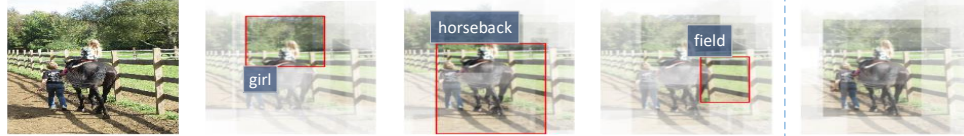
(a) A little girl holding a kitten next to a blue fence.



(b) A child holding a flowered umbrella and petting a yak.



(c) A trainer leads a girl on horseback to a field.



Step: 1

Step: 2

Figure 3: Visualization of attention weights in L2L and G2L alignment step. Specifically, at the first alignment step, each word feature is used as a query to assign weights on salient image regions, and at the second alignment step, textual context-level global features are leveraged as queries to assign different weights on salient image regions. Based on the generated cross-attention weights, we visualize the corresponding attended visual information produced by L2L and G2L alignment step progressively.

Method	Text Retrieval		Image Retrieval		Rsum
	R@1	R@5	R@1	R@5	
SHAN-T2I <sub>Glove</sub>	74.8	94.9	59.9	88.4	510.9
SHAN-T2I <sub>ΔGlove</sub>	74.9	95.2	59.7	88.0	510.7
SHAN-T2I <sub>random</sub>	75.0	96.0	59.6	87.6	510.2
SHAN-T2I <sub>L2L</sub>	72.4	94.5	58.2	87.5	504.4
SHAN-T2I <sub>L2L+G2L</sub>	74.9	95.5	59.3	88.4	510.5
SHAN-T2I	<b>75.9</b>	<b>96.1</b>	<b>60.7</b>	<b>88.2</b>	<b>513.8</b>

Table 3: Ablation experimental results on MS-COCO 1k testing set in terms of Recall@K (R@K).

#### 4.4 Ablation Studies

In this section, we perform ablation studies to investigate the contributions of each component in our proposed SHAN approach. We first validate the effect of different word embedding models by training SHAN-T2I with fixed Glove embedding only (SHAN-T2I<sub>Glove</sub>), fine-tuned Glove embedding only (SHAN-T2I<sub>ΔGlove</sub>) and a learnable embedding layer only (SHAN-T2I<sub>random</sub>). Then, we train SHAN-T2I by removing the G2L alignment module (SHAN-T2I<sub>L2L</sub>) and G2G alignment module (SHAN-T2I<sub>L2L+G2L</sub>) respectively to testify the impact of the proposed hierarchical alignment framework. The experimental results are shown in Table 3. From Table 3, we can draw the following conclusions: (1) The combination of pre-trained Glove embedding model with learnable embedding model can lead to better performance, in comparison to using Glove only or learnable embedding model only; (2) The proposed context-level global-to-local

alignment module and context-level global-to-global alignment module are both validated to be effective by providing considering performance boost.

#### 4.5 Visualization

From the qualitative results in Fig.3, we can observe that the words are appropriately matched with its corresponding depicted regions at alignment step 1. For example, in Fig. 3 (b), our model can well generate interpretable attention weights on nouns like “umbrella” and “yak”, as well as the actions like “holding”. It reveals that L2L alignment module is really conducive to bridging the gap between heterogeneous fragments. At step 2 of alignment, we can see that the meaningful image regions contributing to understanding the global context information are highlighted and trivial ones are suppressed. Overall, these observations confirm the effectiveness of our proposed alignment modules.

#### 5 Conclusion

In this paper, we proposed a Step-wise Hierarchical Alignment Network (SHAN) for image-text matching. Specifically, fragment-level local-to-local alignment, context-level global-to-local alignment and context-level global-to-global alignment were performed progressively to learn the visual-semantic correspondence between instances from different modalities. We conducted extensive experiments and ablation studies on two widely-used benchmark datasets, and the experimental results demonstrated our method can achieve competitive results.

## References

- [Anderson *et al.*, 2018] 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 *IEEE Conference On Computer Vision And Pattern Recognition*, pages 6077–6086, 2018.
- [Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *IEEE International Conference On Computer Vision*, pages 2425–2433, 2015.
- [Chen *et al.*, 2020] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *IEEE Conference On Computer Vision And Pattern Recognition*, pages 12655–12663, 2020.
- [Faghri *et al.*, 2017] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *ArXiv Preprint ArXiv:1707.05612*, 2017.
- [Feng *et al.*, 2019] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *IEEE Conference On Computer Vision And Pattern Recognition*, pages 4125–4134, 2019.
- [Frome *et al.*, 2013] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc’auelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances In Neural Information Processing Systems*, pages 2121–2129, 2013.
- [Hu *et al.*, 2019] Zhibin Hu, Yongsheng Luo, Jiong Lin, Yan Yan, and Jian Chen. Multi-level visual-semantic alignments with relation-wise dual attention network for image and text matching. In *International Joint Conference on Artificial Intelligence*, pages 789–795, 2019.
- [Karpathy and Fei-Fei, 2015] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference On Computer Vision And Pattern Recognition*, pages 3128–3137, 2015.
- [Kiros *et al.*, 2014] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *ArXiv Preprint ArXiv:1411.2539*, 2014.
- [Krizhevsky *et al.*, 2012] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks, 2012. *Advances In Neural Information Processing Systems*.
- [Lee *et al.*, 2018] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *European Conference On Computer Vision*, pages 201–216, 2018.
- [Ma *et al.*, 2015] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *IEEE International Conference On Computer Vision*, pages 2623–2631, 2015.
- [Nam *et al.*, 2017] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *IEEE Conference On Computer Vision And Pattern Recognition*, pages 299–307, 2017.
- [Song and Soleymani, 2019] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *IEEE Conference On Computer Vision And Pattern Recognition*, pages 1979–1988, 2019.
- [Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances In Neural Information Processing Systems*, pages 5998–6008, 2017.
- [Wang *et al.*, 2019] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *IEEE International Conference On Computer Vision*, pages 5764–5773, 2019.
- [Wang *et al.*, 2020a] Haoran Wang, Zhong Ji, Zhigang Lin, Yanwei Pang, and Xuelong Li. Stacked squeeze-and-excitation recurrent residual network for visual-semantic matching. *Pattern Recognition*, 105:107359, 2020.
- [Wang *et al.*, 2020b] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. Consensus-aware visual-semantic embedding for image-text matching. In *European Conference On Computer Vision (ECCV)*, pages 18–34, 2020.
- [Wang *et al.*, 2020c] Jingwen Wang, Lin Ma, and Wenhao Jiang. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *AAAI Conference On Artificial Intelligence*, pages 12168–12175, 2020.
- [Wang *et al.*, 2020d] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *The IEEE Winter Conference On Applications Of Computer Vision*, pages 1508–1517, 2020.
- [Wu *et al.*, 2019] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. Learning fragment self-attention embeddings for image-text matching. In *ACM International Conference On Multimedia*, pages 2088–2096, 2019.
- [Yu *et al.*, 2020] Yunlong Yu, Zhong Ji, Jungong Han, and Zhongfei Zhang. Episode-based prototype generating network for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14035–14044, 2020.
- [Zhang *et al.*, 2019] Wei Zhang, Bairui Wang, Lin Ma, and Wei Liu. Reconstruct and represent video contents for captioning via reinforcement learning. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 42(12):3088–3101, 2019.
- [Zhang *et al.*, 2020] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. Context-aware attention network for image-text retrieval. In *IEEE Conference On Computer Vision And Pattern Recognition*, pages 3536–3545, 2020.