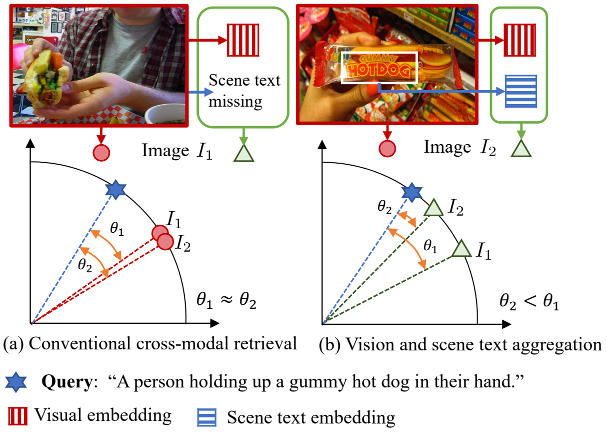
ViSTA: Vision and Scene Text Aggregation for Cross-Modal Retrieval

Mengjun Cheng[2*] Yipeng Sun[1*†] Longchao Wang[1] Xiongwei Zhu[1]
Kun Yao[1] Jie Chen[2] Guoli Song[3] Junyu Han[1] Jingtuo Liu[1] Errui Ding[1] Jingdong Wang[1]
Department of Computer Vision Technology (VIS), Baidu Inc.[1]
School of Electronic and Computer Engineering, Peking University,[2] Peng Cheng Laboratory[3]
{sunyipeng, wanglongchao, zhuxiongwei, yaokun01, hanjunyu, liujingtuo}@baidu.com
mjcheng@stu.pku.edu.cn, {chenj, songgl}@pcl.ac.cn, {dingerrui, wangjingdong}@baidu.com

Abstract

*Visual appearance is considered to be the most important cue to understand images for cross-modal retrieval, while sometimes the scene text appearing in images can provide valuable information to understand the visual semantics. Most of existing cross-modal retrieval approaches ignore the usage of scene text information and directly adding this information may lead to performance degradation in scene text free scenarios. To address this issue, we propose a full transformer architecture to unify these cross-modal retrieval scenarios in a single **V**ision and **S**cene **T**ext **A**ggregation framework (ViSTA). Specifically, ViSTA utilizes transformer blocks to directly encode image patches and fuse scene text embedding to learn an aggregated visual representation for cross-modal retrieval. To tackle the modality missing problem of scene text, we propose a novel fusion token based transformer aggregation approach to exchange the necessary scene text information only through the fusion token and concentrate on the most important features in each modality. To further strengthen the visual modality, we develop dual contrastive learning losses to embed both image-text pairs and fusion-text pairs into a common cross-modal space. Compared to existing methods, ViSTA enables to aggregate relevant scene text semantics with visual appearance, and hence improve results under both scene text free and scene text aware scenarios. Experimental results show that ViSTA outperforms other methods by at least **8.4**% at Recall@1 for scene text aware retrieval task. Compared with state-of-the-art scene text free retrieval methods, ViSTA can achieve better accuracy on Flicker30K and MSCOCO while running at least three times faster during the inference stage, which validates the effectiveness of the proposed framework.*



(a) Conventional cross-modal retrieval (b) Vision and scene text aggregation

Figure 1. Given a text query, two images are close in visual semantics for (a) conventional cross-modal retrieval. By considering visual appearance and scene text information, e.g.,“gummy hotdog”, into one framework, (b) the proposed **V**ision and **S**cene **T**ext **A**ggregation (ViSTA) approach enables to distinguish the semantic difference between images I_1 and I_2 ($\theta_2 < \theta_1$), and can be also adapted to conventional scene text free scenarios.

1. Introduction

As one of the most important multi-modal understanding tasks, cross-modal retrieval attracts much attention due to its valuable applications, e.g., news search and product retrieval. Cross-modal text-to-image retrieval [10, 11, 22] aims to return the most relevant candidate based on the relevance between the text content of a query and the visual appearance of an image. The performance of this retrieval task is largely improved by better visual representation and detailed image-text alignment [3, 22, 25, 28].

In recent years, following the success of BERT [7] in natural language modeling, transformer-based single encoder architectures [5, 15, 16, 20, 23, 26, 29, 36, 44, 47, 49] are adopted to fuse images and text, and image-text pretraining for fine-tuning becomes the mainstream paradigm in modeling visual-language tasks, significantly boosting

[*] Equal Contributions. This work is done when Mengjun Cheng is a research intern at Baidu Inc.

[†] Corresponding author.

the cross-modal retrieval performance. However, these approaches with deep interactions between images and text are orders of magnitudes slower and hence impractical for the large-scale cross-modal retrieval task. As dual-encoder architectures, CLIP [37], ALIGN [18] and WenLan [17] exploit cross-modal contrastive pre-training by encoding images and text separately, which allows that image and text features can be computed in an offline setting to efficiently calculate similarities between large-scale image-text pairs. Even though the performance of the cross-modal retrieval task is greatly improved by the million-scale image-text contrastive pre-training [37], it is still difficult and ineffective to learn specific fine-grained visual concepts, e.g., the scene text semantics from images [37]. More recently, a new cross-modal retrieval task [31] is proposed to enable the usage of scene text in an image together with its visual appearance. Specifically, an image in this task is paired with the corresponding scene text features to help to determine the similarity between the query's textual content and the image's visual appearance plus scene text. Benefiting from exploiting additional scene text features, this model can improve the cross-modal retrieval accuracy than those exploiting only visual appearance. Nevertheless, in a real-world image corpus, there are only a fraction of images containing scene text instances. The model designed for the scene text aware retrieval task might fail to generate reliable similarities between the query and images without scene text instances, and can not be adapted to the conventional scene text free retrieval task.

To overcome this issue, we propose an effective **Vi**sion and **S**cene **T**ext **A**ggregation (ViSTA) framework to tackle both scene text aware and scene text free cross-modal retrieval tasks. Specifically, ViSTA utilizes a full transformer design to directly encode image patches and fuse scene text embedding to learn an aggregated visual representation. To enforce each modality focusing on its most important features, we propose a novel token based aggregation approach by sharing the necessary scene text information only through the fusion token. To tackle the modality missing problem of scene text, we further develop dual contrastive supervisions to strengthen the visual modality, and embed both image-text pairs and fusion-text pairs into a common cross-modal space. Compared to existing fusion methods, ViSTA enables to aggregate relevant scene text semantics with visual appearance, and hence improve results under both scene text free and scene text aware scenarios.

The contributions of this paper are three-fold. **1**) We propose a full transformer architecture to effectively aggregate vision and scene text, which is applicable in both scene text aware and scene text free retrieval scenarios. **2**) We propose a fusion token based transformer aggregation design to exchange the relevant information among visual and scene text features, and dual contrastive losses to enhance visual features. **3**) The proposed cross-modal retrieval framework can remarkably surpass existing methods for the scene text aware retrieval task and achieve better performance than state-of-the-art approaches on scene text free retrieval benchmarks as well.

To the best of our knowledge, it is the first time to solve scene text free and scene text aware cross-modal retrieval tasks with a vision and scene text aggregated transformer.

2. Related Work

Cross-modal retrieval aims to return relevant images or text descriptions given text or an image query. Most approaches learn a joint cross-modal embedding space to produce closer representation for semantically relevant image and text pairs [10, 11, 33]. Since the deep learning era, the visual representation for cross-modal retrieval has been consistently improved from grid-based CNN (convolution neural network) [10] to a pre-trained object detector [22, 25]. In the meantime, finer image-text alignment approaches are developed, e.g., attention mechanisms, iterative matching, and graph-based relationship reasoning between image features and text embedding [3, 8, 22, 25, 28]. Most of these approaches rely on RoI (region-of-interest) features extracted from a pre-trained Faster-RCNN detector on the Visual Genome (VG) dataset [21], which limits the performance on the out-of-domain visual concepts. By contrast, ViSTA directly takes image patches as the input and builds upon the recent contrastive image-text pre-training paradigm, which is capable of achieving better performance by end-to-end training at a much faster inference speed.

Vision language pre-training has become a mainstream paradigm in multi-modal understanding, which can remarkably boost the performance on various vision and language tasks, e.g., cross-modal retrieval and visual question answering (VQA), etc. Most of these approaches utilize transformer based architectures, which can be categorized as single-encoder and dual-encoder pre-training. The single encoder architectures [5, 15, 16, 20, 23, 26, 29, 36, 41, 42, 44, 47, 49] are adopted to fuse images and text with the multimodal transformer for interactions, performing high accuracy in various downstream tasks. To speed up the inference stage and adapt to more visual categories, grid-based image features [15, 16] and newly proposed patch-based image embedding methods [20, 24, 44] are utilized for end-to-end training, which directly take image pixels or patches and text embedding as the input. However, the computation cost of these approaches is still huge and impractical for the large-scale cross-modal retrieval task. Instead, dual-encoder architectures [17, 18, 37] encode images and text separately, making it possible to calculate similarities of image-text pairs in the linear time complexity. Even though the performance of the cross-modal retrieval task is greatly improved by the million-scale image-text contrastive pre-

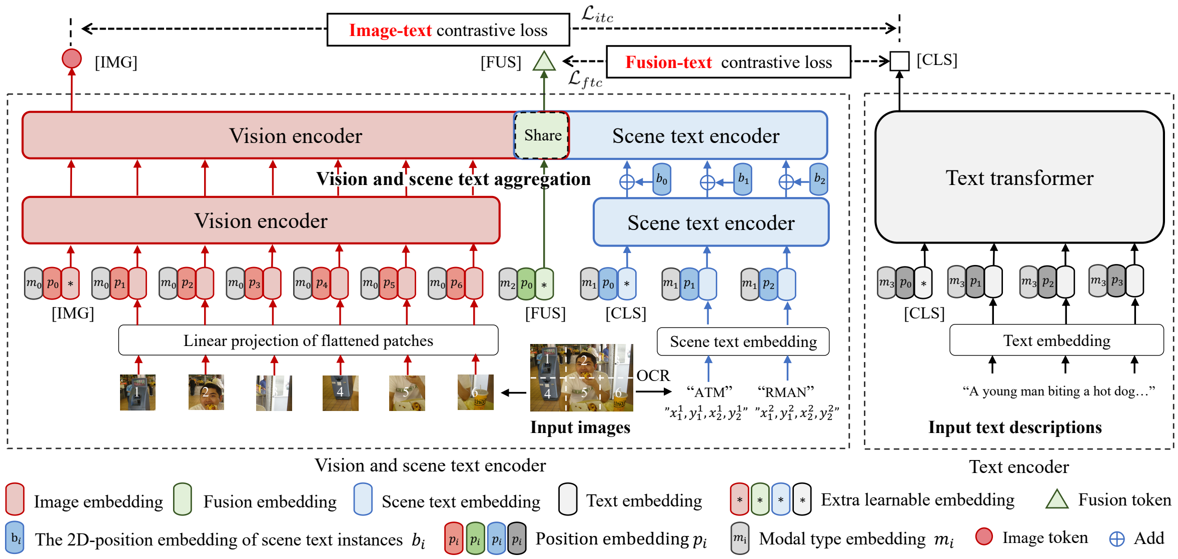


Figure 2. The proposed **V**ision and **S**cene **T**ext **A**ggregation (ViSTA) framework for cross-modal retrieval. With the proposed fusion token based vision scene text aggregation layer, ViSTA learns a common cross-modal space by a dual-encoder transformer architecture, supervised by dual contrastive losses between image-text pairs and fusion-text pairs, respectively.

training [37], it is still difficult and ineffective to learn specific fine-grained visual concepts, e.g., the scene text semantics from images [37]. By contrast, ViSTA incorporates vision and scene text into a full transformer based dual-encoder architecture, taking image patches, scene text, and text queries as the input for unified cross-modal retrieval.

Scene text in vision and language receives much attention as the extension of previous applications, e.g., text-based image caption [39, 45] and Text-VQA [2, 40, 45, 48, 50]. All these approaches utilize OCR (optical character recognition) results to form scene text embedding [2, 12, 40, 45], following the typical architecture of single-stream transformer [29] with RoI region features. Other works [12] [43] for scene text retrieval tasks aim to return images that contain the query word, and a CNN based fusion approach [1] integrates scene text and visual appearance to improve the performance for fine-grained image classification in specific scenarios. More recently, StacMR [31] introduces scene text aware cross-modal retrieval (StacMR) considering scene text as an additional modality, which utilizes GCN (graph convolution network) to obtain context representation of images and scene text for final fusion. Different from all these methods, ViSTA utilizes full transformer blocks to encode image patches and scene text with mid-level fusion, which can be adapted to both scene text aware and scene text free scenarios.

3. Approach

The overall architecture of our proposed ViSTA framework is developed as a dual-encoder architecture as shown in Fig. 2, which makes it practical for large-scale cross-modal retrieval. To achieve a strong feature representation for better retrieval accuracy, we adopt a full transformer design to encode images, scene text, and text query by uni-modal encoders, respectively, before feeding them for further aggregation and calculating the cross-modal contrastive losses. The whole model including vision, scene text, and text encoders is end-to-end trainable, which allows better generalization beyond RoI features by cross-modal pre-training [16] [15] [20] [44]. In order to fuse visual features with relevant scene text semantics, we propose a fusion token based aggregation approach, which shares the relevant information across these two modalities only through the fusion token. As a result, this token can see all the information at each transformer layer and can be used for fusion-text contrastive learning. Since scene text instances do not often appear in images and in some cases the correlation between scene text and images might be weak in visual semantics. Therefore, to enhance the visual representation rather than over-fit to the noisy scene text features, we also utilize image token at the last layer for effective image-text contrastive learning. With such designs, ViSTA can be effectively adapted to both scene text aware and scene text free retrieval scenarios.

Problem formulation. Given a set of image and text pairs, the vision and scene text encoder aims to encode an image I and recognize the scene text appearing in this image. The scene text instances contain a set of N_o detected words and locations by an OCR model as $\mathcal{O} = \{\mathbf{o}_j^{word}, \mathbf{o}_j^{bbox}\}_{j=1}^{N_o}$. If there is no scene text detected in the image, $\mathcal{O}$ can be an empty set as $\emptyset$. In the scene text aware text-to-image retrieval task [31], the model is required to generate a similarity score $S(q, I)$ between a text query q and each image I based on the relevance of the query's textual content and the image's visual features $\mathcal{V}$ together with its scene text features $\mathcal{O}$. In the scene text free text-to-image retrieval task, which is the same as the conventional text-to-image retrieval, scene text instances do not appear in images. Therefore, these images are only sorted by the relevance between the visual appearance and the content of text query.

3.1. Vision and Scene Text Encoders

Following the success of vision transformer [9], the vision encoder directly takes image patches as the input. By slicing an image into multiple patches, a patch sequence $\mathbf{P} = [\mathbf{p}_1, \cdots, \mathbf{p}_{N_p}]$ is used to form a simple linear projection of pixels before feeding into transformers. A positional embedding is added to each patch token to encode the position information. Besides, the embedding of the devised special token [IMG] is inserted in $\mathbf{P}$. The vision encoder is built upon a stack of L_v standard transformer layers. Let us denote the input sequence of the l-th vision transformer layer by $\mathbf{V}_l$. The output sequence of the l-th layer severs as the input sequence of the next layer, calculated as

$$\begin{aligned} \mathbf{Y}_l &\leftarrow \mathrm{MHSA}(\mathrm{LN}(\mathbf{V}_l)) + \mathbf{V}_l \\ \mathbf{V}_{l+1} &\leftarrow \mathrm{MLP}(\mathrm{LN}(\mathbf{Y}_l)) + \mathbf{Y}_l, \end{aligned} \tag{1}$$

where $\mathrm{MHSA}(\cdot)$ denotes the multi-head self-attention layer, $\mathrm{MLP}(\cdot)$ denotes the multi-layer perception layer, and $\mathrm{LN}(\cdot)$ denotes the layer normalization. The input of the first transformer block, $\mathbf{V}_1$, is just the patch sequence $\mathbf{P}$. Finally, the output of the last vision transformer layer, $\mathbf{V}_{L_v}$, serves as the visual features $\mathcal{V} = \{\mathbf{v}_j\}_{j=1}^{N_v}$. To be specific, the j-th item in $\mathbf{V}_{L_v}$ corresponds to the $\mathbf{v}_j$ as $\mathbf{v}_j = \mathbf{V}_{L_v}[:, j]$.

Similar to the vision encoder, the scene text encoder is a stack of L_s standard transformer layers. The input scene text embedding is mainly obtained from the OCR results by Google API [13] and encoded in tokens. The input token from these OCR results is combined with modal type $\mathbf{S}^{type}$ and position embedding $\mathbf{S}^{token\_id}$ as

$$\mathbf{S}_{init} = \mathrm{Embedding}(\mathbf{o}^{word}) + \mathbf{S}^{type} + \mathbf{S}^{token\_id}. \tag{2}$$

Following the previous method in Text-VQA [14], the scene text embedding encoded by BERT [7] can be further combined with the 4-dimensional location information of OCR tokens using normalized bounding box coordinates $\mathbf{o}^{bbox}$ and can be formulated as

$$\mathbf{S}_0 = \mathbf{BERT}(\mathbf{S}_{init}) + \mathrm{F}_{linear}(\mathbf{o}^{bbox}), \tag{3}$$

where F_{linear} linearly projects the normalized coordinates into the two-dimensional position embedding with the same size as the encoded scene text tokens.

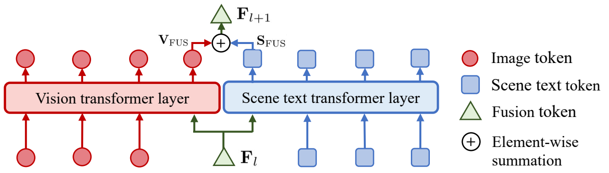


Figure 3. The vision scene text aggregation layer. The fusion token shared between two modalities exchanges the relevant information to learn a scene text aggregated visual representation.

3.2. Vision and Scene Text Aggregation

Since scene text appearing in images may provide valuable information while in most cases images do not contain any scene text information, the semantic relevance between scene text and visual appearance varies case by case, which might be weak in correlations. Therefore, it is challenging to aggregate these two different modalities into a unified visual representation for effective cross-modal retrieval.

To handle both scene text aware and scene text free cross-modal retrieval tasks, it is necessary for the visual tower to learn corresponding final features of image modality for matching. Therefore we use different tokens, an image token or a fusion token, to get the final features based on whether the OCR recognition result is none or not in the training stage. In the scene text free scenarios, our visual tower degenerates to a pure vision encoder model as in Section 3.1 and outputs the image feature of [IMG] token as final features. In the scene text aware scenarios, we use the scene text encoder to learn semantic features of scene text. And as shown in Fig 2, our visual tower simply adds L_f layers of vision and scene text aggregation layer to make mid-level fusion in image modality and outputs fusion features from extra fusion token [FUS] as final features.

As shown in the detailed structure of Fig. 3, the vision scene text aggregation layer is composed of a vision transformer layer and a scene text transformer layer from two encoders. To exchange the relevant information of vision and scene text, the two layers are added with a new token, which is a shared special fusion token [FUS]. We denote the input image token and scene text token of l-th vision encoder and scene text encoder in the aggregation stage by $\mathbf{V}_l$ and $\mathbf{S}_l$. And the input fusion token of l-th vision and scene text aggregation is denoted by $\mathbf{F}_l$. The workflow of the vision transformer layer of Eq. 1 in the aggregation stage is updated to

$$\begin{aligned} \mathbf{Y}_l &\leftarrow \mathrm{MHSA}(\mathrm{LN}([\mathbf{V}_l; \mathbf{F}_l])) + [\mathbf{V}_l; \mathbf{F}_l] \\ [\mathbf{V}_{l+1}; \mathbf{V}_{\mathrm{FUS}}] &\leftarrow \mathrm{MLP}(\mathrm{LN}(\mathbf{Y}_l)) + \mathbf{Y}_l, \end{aligned} \tag{4}$$

where $\mathbf{V}_{\mathrm{FUS}}$ is the output image feature corresponding to the fusion token. Same is the workflow of the scene text transformer layer in the aggregation stage as

$$\begin{aligned} \mathbf{Y}_l &\leftarrow \mathrm{MHSA}(\mathrm{LN}([\mathbf{S}_l; \mathbf{F}_l])) + [\mathbf{S}_l; \mathbf{F}_l] \\ [\mathbf{S}_{l+1}; \mathbf{S}_{\mathrm{FUS}}] &\leftarrow \mathrm{MLP}(\mathrm{LN}(\mathbf{Y}_l)) + \mathbf{Y}_l, \end{aligned} \tag{5}$$

Table 1. Dataset split for the evaluation of cross-modal retrieval tasks. Note that $*$ indicates that the CTC-5K test samples have been excluded from the MSCOCO train split.

Task	Pre-training	Fine-tuning	Test
Scene text aware	VG	Flickr30K + TC + CTC train	CTC-1K, 5K
Conventional scene text free	SBU + GCC + VG + MSCOCO*	Flickr30K train	Flickr30K test
		MSCOCO* train	MSCOCO-5K test

Table 2. Model settings at various scales.

Model	Vision	Scene text	Input
	encoder	encoder	size
ViSTA-S	12 layers, 6 heads	BERT-mini	224× 224
ViSTA-B	12 layers, 12 heads	BERT-Base	384× 384
ViSTA-L	12 layers, 24 heads	BERT-Base	384× 384

where $\mathbf{S}_{\text{FUS}}$ is the output scene text feature corresponding to the fusion token. The input fusion features of the next layer are calculated by their element-wise summation as shown in Fig. 3, defined as $\mathbf{F}_{l+1} = \mathbf{V}_{\text{FUS}} + \mathbf{S}_{\text{FUS}}$. In this way, visual features $\mathbf{V}$ and scene text features $\mathbf{S}$ are learned through independent transformer layers, respectively. The special fusion token [FUS] plays the role of the bridge of two encoders as it is shared in two encoders. Due to the vision and scene text aggregation layers, the learning of image features and scene text features is affected by each other by the indirect fusion token. A similar bottleneck attention structure for video classification [34] fuses video patches and sound by averaging the prediction of the two modalities. Instead of updating shared tokens twice, ViSTA directly adds the predicted fused tokens from vision and scene text transformer layers, forming the fusion token during the aggregation process. To further consider the modality missing problem of scene text, we propose additional image-text contrastive loss to enhance the visual representation together with the fusion-text contrastive loss. Therefore, both image-text pairs and fusion-text pairs contain the information of visual appearance and share only the relevant part of the information within scene text through shared tokens, which aims to benefit scene text aware cross-modal learning.

Fusion feature embedding. We consider the fusion token as another modality and therefore add a different modal type embedding to the randomly initialized [FUS] token embedding, which can be calculated as

$$\mathbf{F}_0 = \mathbf{F}^{init} + \mathbf{F}^{type} + \mathbf{F}^{token\_id}, \tag{6}$$

where $\mathbf{F}_0$ is the first input fusion features of vision and scene text aggregation layers.

3.3. Cross-Modal Contrastive Learning

Conventional scene text free and scene text aware image-text retrieval are two different tasks calling for visual features only and fused visual-semantic features respectively, which correspond with the output features of [IMG] and [FUS] tokens. The final features are constructed into image-text pairs or fusion-text pairs with text queries. We introduce dual contrastive learning losses to embed both image-text pairs and fusion-text pairs into a common cross-modal space. The total loss is

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{itc} + (1-\alpha)\mathcal{L}_{ftc}, \tag{7}$$

where $\mathcal{L}_{itc}$ and $\mathcal{L}_{ftc}$ are image-text contrastive loss and fusion-text contrastive loss. Note that α is the parameter to trade off between these losses and is set to 0.9 as default.

For N image and text pairs as a batch, the fusion-text contrastive loss aims to maximize the similarity between N matched pairs and minimize the similarity between the last $N^2 - N$ incorrect pairs, formulated as

$$\mathcal{L}_{ftc} = \frac{1}{2}(\mathcal{L}_{f2t} + \mathcal{L}_{t2f}). \tag{8}$$

The fusion-text contrastive learning aims to minimize the symmetric loss between the fused token and text [CLS] as

$$\begin{aligned} \mathcal{L}_{f2t} &= -\frac{1}{N}\sum_{i=1}^{N} log \frac{\exp(f_i^\top t_i/\sigma)}{\sum_{j=1}^{N}\exp(f_i^\top t_j/\sigma)} \\ \mathcal{L}_{t2f} &= -\frac{1}{N}\sum_{i=1}^{N} log \frac{\exp(t_i^\top f_i/\sigma)}{\sum_{j=1}^{N}\exp(t_i^\top f_j/\sigma)}, \end{aligned} \tag{9}$$

where f_i and t_j are the normalized embedding of fusion features in the i-th pairs and that of text in the j-th pairs, respectively. The temperature parameter σ is a trainable variable and its initial value is set to 0.07 as default [18]. Same as $\mathcal{L}_{ftc}$, the image-text contrastive loss is formulated as

$$\mathcal{L}_{itc} = \frac{1}{2}(\mathcal{L}_{i2t} + \mathcal{L}_{t2i}), \tag{10}$$

where

$$\begin{aligned} \mathcal{L}_{i2t} &= -\frac{1}{N}\sum_{i=1}^{N} log \frac{\exp(v_i^\top t_i/\sigma)}{\sum_{j=1}^{N}\exp(v_i^\top t_j/\sigma)} \\ \mathcal{L}_{t2i} &= -\frac{1}{N}\sum_{i=1}^{N} log \frac{\exp(t_i^\top v_i/\sigma)}{\sum_{j=1}^{N}\exp(t_i^\top v_j/\sigma)}. \end{aligned} \tag{11}$$

Note that v_i is the normalized embedding of the i-th image. In the training stage, if the extracted OCR result is None, the $\mathcal{L}_{ftc}$ loss would not be added to the total loss.

Table 3. Comparisons with the state-of-the-art scene text aware approaches on CTC.

Model	CTC-1K						CTC-5K					
	Image-to-text			Text-to-image			Image-to-text			Text-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
SCAN [22]	36.3	63.7	75.2	26.6	53.6	65.3	22.8	45.6	54.3	12.3	28.6	39.9
VSRN [25]	38.2	67.4	79.1	26.6	54.2	66.2	23.7	47.6	59.1	14.9	34.7	45.5
STARNet [31]	44.1	74.8	82.7	31.5	60.8	72.4	26.4	51.1	63.9	17.1	37.4	48.3
ViSTA-S	**52.5**	**77.9**	**87.2**	**36.7**	**66.2**	**77.8**	**31.8**	**56.6**	**67.8**	**20.0**	**42.9**	**54.4**

Table 4. Comparisons with other approaches on Flickr30K and MSCOCO in terms of zero-shot retrieval.

Model	Time (ms)	Flickr30k (1K)						MS-COCO (5K)					
		Image-to-text			Text-to-image			Image-to-text			Text-to-image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ViL-BERT [29]	˜900	31.9	61.1	72.8	-	-	-	-	-	-	-	-	-
Unicoder-VL [23]	˜925	64.3	85.8	92.3	48.4	76.0	85.2	-	-	-	-	-	-
ImageBERT [36]	˜900	70.7	90.2	94.0	54.3	79.6	87.5	44.0	71.2	80.4	32.3	59.0	70.2
UNITER-B [5]	˜900	**80.7**	**95.7**	98.0	66.2	88.4	92.9	-	-	-	-	-	-
ViLT-B [20]	˜15	73.2	93.6	96.5	55.0	82.5	89.8	56.5	82.6	89.6	40.4	70.0	81.1
ViSTA-B	˜17	75.3	93.8	97.5	59.5	84.3	90.3	60.7	85.8	92.3	44.8	72.8	82.5
ViSTA-L	˜40	79.2	95.4	**98.1**	**67.0**	**88.7**	**93.1**	**63.9**	**87.1**	**93.0**	**47.4**	**75.0**	**84.0**

4. Experiments

We conduct experiments on two downstream cross-modal retrieval benchmarks to validate the effectiveness of the proposed approach. The scene text aware cross-modal retrieval task is evaluated on the COCO-Text Captioned (CTC) [31] dataset, and the conventional cross-modal retrieval experiments are conducted on the Flickr30K [46] and MSCOCO [19] benchmarks, including image-to-text and text-to-image retrieval tasks reported in Tab. 3 and Tab. 5. We also analyze the effectiveness of structures of the proposed ViSTA and show some cases in the ablation study.

Datasets. All the pre-training, fine-tuning, and test settings of different tasks are reported in Tab. 1. The setting of scene text aware cross-modal retrieval task follows [31]. In conventional scene text free cross-modal retrieval task, four publicly available datasets including Microsoft COCO (MSCOCO) [27], Visual Genome (VG) [21], SBU Captions (SBU) [35], and Google Conceptual Captions (GCC) [38] datasets are used for pre-training. Since the CTC dataset is also constructed from MSCOCO, all images of CTC-5K test are contained in MSCOCO train set. Therefore, for evaluation purpose on the CTC dataset, we remove the duplicate images from the MSCOCO dataset and denote it as MSCOCO*. For the evaluation metric, all these experiments are evaluated in terms of the percentage of containing a matched pair in the top returns, i.e., R@1, R@5, and R@10, respectively.

Implementation details. For a fair comparison, we implement several versions of models with different scales, as shown in Tab. 2. For all experiments, we use the AdamW optimizer with a base learning rate of 1e-4 and augmentation of random horizontal flipping and random augmentation [16]. We pre-train for 80 epochs on 40 NVIDIA Tesla V100 GPUs and finetune for another 10 epochs on 8 Tesla V100 GPUs. For scene text free cross-modal retrieval task, we pre-train ViSTA-B and ViSTA-L on combined datasets, i.e., SBU, CC, VG and MSCOCO* for fair comparisons with previous approaches. Note that images in the CTC train set are all included in the train set of MSCOCO.

4.1. Scene Text aware Cross-Modal Retrieval

For fair comparisons in scene text aware retrieval, we evaluate models on CTC-1K and CTC-5K test sets, respectively, strictly following the previous train and test split [31]. As shown in Tab. 3, Our ViSTA-S model performs a large improvement of 8.4% and 5.4% on R@1 in scene text aware image-text retrieval task on CTC-1k. Compared to STARNet [31] which uses GCN to get the representation of scene text for fusion, we use BERT to refine it. And the self-attention operators on vision encoders learn the long-range dependence in images and help our ViSTA model to learn the relationship between patches. The vision and scene text aggregation layers learn the joint distribution of modalities of vision and scene text and refine the representation space.

4.2. Scene Text Free Cross-Modal Retrieval

For conventional image-text retrieval, we measure zero-shot and fine-tuned performance on the Karpathy & Fei-Fei split of MS-COCO and Flickr30K [46] and compare with state-of-the-art methods in Tab. 4 and Tab. 5, respectively. All the settings are the same as the pre-training stage. When fine-tuning on Flickr30K, we use the fine-tuned weight on COCO-5K as the initial weight. Following the efficient framework of dual-tower and patch projection operator, our model has a comparable speed with ViLT [20] and better performance, as shown in Tab. 5. And our large-scale model ViSTA-L achieves superior results than state-of-the-

Table 5. Comparisons with state-of-the-art approaches on fine-tuning Flicker30K and MSCOCO benchmarks.

Model	Time (ms)	Flickr30K (1K)						MS-COCO (5K)					
		Image-to-text			Text-to-image			Image-to-text			Text-to-image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
SCAN [22]	-	67.4	90.3	95.8	48.6	77.7	85.2	50.4	82.2	90.0	38.6	69.3	80.4
VSRN [25]	-	71.3	90.6	96.0	54.7	81.8	88.2	53.0	81.1	89.4	40.5	70.6	81.1
IMRAM [3]	-	74.1	93.0	96.6	53.9	79.4	87.2	53.7	83.2	91.0	39.7	69.1	79.8
GSMN [28]	-	76.4	94.3	97.3	57.4	82.3	89.0	-	-	-	-	-	-
SGRAF [8]	-	77.8	94.1	97.4	58.5	83.0	88.8	57.8	-	91.6	41.9	-	81.3
Vil-BERT [29]	~920	58.2	84.9	91.5	-	-	-	-	-	-	-	-	-
Unicoder-VL [23]	~925	86.2	96.3	99.0	71.5	91.2	95.2	62.3	87.1	92.8	48.4	76.7	85.9
UNITER-B [5]	~900	85.9	97.1	98.8	72.5	92.4	96.1	64.4	87.4	93.1	50.3	78.5	87.2
ERNIE-ViL-B [47]	~920	86.7	97.8	99.0	74.4	92.7	95.9	-	-	-	-	-	-
VSEinfty [4]		88.4	98.3	99.5	74.2	93.7	96.8	66.4	89.3	-	51.6	79.3	-
PCME [6]	-	-	-	-	-	-	-	44.2	73.8	83.6	31.9	62.1	74.5
Miech et al [32]	-	-	-	-	72.1	91.5	95.2	-	-	-	-	-	-
12-in-1 [30]	-	-	-	-	67.9	89.6	94.2	-	-	-	-	-	-
Pixel-BERT-X [16]	~160	87.0	**98.9**	99.5	71.5	92.1	95.8	63.6	87.5	93.6	50.1	77.6	86.2
SOHO [15]	-	86.5	98.1	99.3	72.5	92.7	96.1	66.4	88.2	93.8	50.6	78.0	86.7
H Xue et al. [44]	-	87.0	98.4	99.5	73.5	93.1	96.4	-	-	-	-	-	-
Pixel-BERT-R [16]	~60	75.7	94.7	97.1	53.4	80.4	88.5	59.8	85.5	91.6	41.1	69.7	80.5
ViLT-B [20]	~15	83.5	96.7	98.6	64.4	88.7	93.8	61.5	86.3	92.7	42.7	72.9	83.1
ViSTA-B	~17	84.8	97.4	99.0	68.9	91.1	95.1	63.9	87.8	93.6	47.8	75.8	84.5
ViSTA-L	~40	**89.5**	98.4	**99.6**	**75.8**	**94.2**	**96.9**	**68.9**	**90.1**	**95.4**	**52.6**	**79.6**	**87.6**

art methods at a low speed. Our model is not affected in those datasets when the modality of scene text is missing and still performs well on downstream tasks due to the fusion token based vision and scene text aggregation.

4.3. Ablations

To validate the effectiveness of the proposed vision and scene text aggregation layers for the visual tower, we conduct ablation experiments on the CTC dataset. We fix the text tower with BERT-mini and implement different visions of the visual tower. As shown in Tab. 6, only using scene text information encoded by GCN or BERT-mini is insufficient for cross-modal retrieval. Compare with the architecture in STARNet [31], incorporating vision transformer in cross-modal retrieval, e.g., ViT-S, can achieve better performance due to the improved visual representation. Compared with results of only using the visual modality, ViSTA with scene text embedding can remarkably improve the performance by 5.5%/2.1% in R@1 on CTC-1K, which is contributed by the effective vision and scene text aggregation.

Table 6. Ablation study on the impact of modality aggregation.

Model	Visual	Scene text	CTC-1K					
			Image-to-text			Text-to-image		
			R@1	R@5	R@10	R@1	R@5	R@10
GCN		✓	10.8	20.2	25.4	4.4	11.3	15.6
BERT-mini		✓	24.3	35.4	40.8	9.6	17.8	22.6
RoI + GCN [31]	✓	✓	44.1	74.8	82.7	31.5	60.8	72.4
ViT-S + GCN	✓	✓	47.2	74.2	84.2	33.2	63.6	75.4
ViSTA-S	✓		47.0	73.8	84.3	34.6	63.4	75.3
ViSTA-S	✓	✓	**52.5**	**77.9**	**87.2**	**36.7**	**66.2**	**77.8**

Table 7. Ablation study on the number of fusion layers.

The number of fused layers	CTC-1K					
	Image-to-text			Text-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10
$L_f = 1$	48.2	74.3	85.0	35.6	64.8	76.8
$L_f = 2$	52.2	77.0	86.3	35.4	64.8	76.2
$L_f = 4$	**52.5**	**77.9**	**87.2**	**36.7**	**66.2**	**77.8**

Table 8. Ablation study on the impact of various fusion strategies.

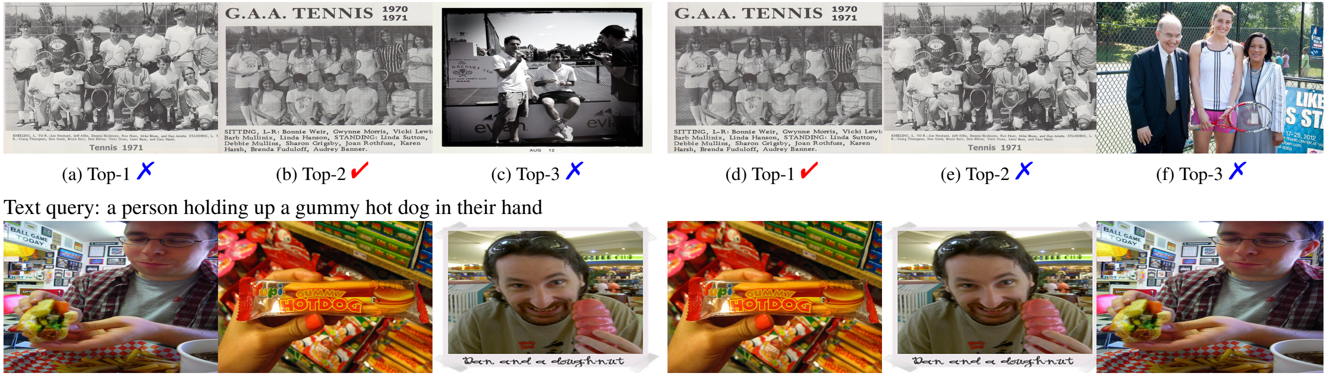
Fusion strategies	CTC-1K					
	Image-to-text			Text-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10
Global attention	48.4	75.5	86.5	34.7	64.3	76.2
Cross attention	50.5	74.4	84.1	31.1	59.8	72.9
Fusion token	**52.5**	**77.9**	**87.2**	**36.7**	**66.2**	**77.8**
Late fusion	49.2	73.4	85.8	34.9	65.0	76.7

Table 9. Ablation study on the impact of loss functions.

$\mathcal{L}_{ftc}$	$\mathcal{L}_{itc}$	CTC-1K					
		Image-to-text			Text-to-image		
		R@1	R@5	R@10	R@1	R@5	R@10
✓		46.6	71.3	82.4	30.3	58.7	71.4
✓	✓	**52.5**	**77.9**	**87.2**	**36.7**	**66.2**	**77.8**

We also conduct several experiments to validate the effectiveness of the proposed architecture. Tab. 7 shows that the model can improve the results with an increased number of fusion layers. Tab. 8 shows the results between the proposed fusion token based method, multi-modal transformer with global attention [41] and cross-attention [29] as well as late fusion strategy for comparisons. As shown in the Tab. 9, our proposed dual contrastive learning performs better than a single fusion based contrastive loss. Two separate contrastive losses for scene text aware scenarios help to maintain effective cross-modal features when the scene text

Text query: A tennis team was featured in a newspaper in 1970 or 1971.



(a) Top-1 ✗ (b) Top-2 ✓ (c) Top-3 ✗ (d) Top-1 ✓ (e) Top-2 ✗ (f) Top-3 ✗

Text query: a person holding up a gummy hot dog in their hand

(g) Top-1 ✗ (h) Top-2 ✓ (i) Top-3 ✗ (j) Top-1 ✓ (k) Top-2 ✗ (l) Top-3 ✗

Figure 4. Examples of the text-to-image retrieval task for comparisons between results with and without scene text. Note that text queries with the corresponding top returned images are shown in (a) to (l). The first three columns show the retrieved results of ViSTA-S without scene text embedding, and the last three columns show the results of ViSTA-S. (best view in colors).

(a) Top three returned results by ViSTA:
1) A man eating a Nathans chili cheese dog in front of an **ATM**. ✓
2) A man eats a hot dog at a fast food place. ✗
3) The guy is eating a doughnut at a doughnut shop. ✗

Top three returned results of ViSTA w/o scene text:
1) The guy is eating a doughnut at a doughnut shop. ✗
2) A man eats a hot dog at a fast food place. ✗
3) A young man biting a hot dog sitting at a table at a fast food court. ✓

(b) Top three retrieved results by ViSTA:
1) A **STA LUCIA** bus is driving down the road. ✓
2) A bus sits in the parking lot outside of Piccadilly Gardens. ✗
3) A charter bus with two stories heading to some where. ✗

Top three returned results of ViSTA w/o scene text:
1) A bus pull into a small parking lot space. ✗
2) A charter bus with two stories heading to some where.✗
3) A bus sits in the parking lot outside of Piccadilly Gardens. ✗

(c) Top three returned results by ViSTA:
1) The arriving passengers on the Ethiopian airliner are deplaning on the runway. ✗
2) A **China** Airlines airliner is parked at an airport near another jet. ✓
3) A large continental jet sitting on a tarmac at an airport. ✗

Top three returned results of ViSTA w/o scene text:
1) Commercial Lufthansa air plane parked at an airport. ✗
2) The arriving passengers on the Ethiopian airliner are deplaning on the runway. ✗
3) An Aegean Airlines airplane on an airport runway. ✗

Figure 5. Examples of image-to-text retrieval for comparisons between the top returned results with and without scene text.

information is noisy or missing.

Qualitative comparisons. For visual comparisons, we also report some examples to illustrate the effectiveness of our method. Our model benefits from the scene text information in learning visual features. As shown in Fig. 4, based on the query of ”tennis”, ”1970”, and ”1971”, our ViSTA model matches the correct images while the ViTSA without scene text embedding retrieves a confusing result. And in the second example, the ”gummy hotdog” is perfectly retrieved. For the text retrieval task, shown in Fig. 5, the scene text extracted from images has semantic information and is contained in retrieved results with ViSTA while it does not well without scene text embedding.

5. Conclusions and Discussions

We have proposed an effective vision and scene text aggregation transformer to learn a scene text enhanced visual representation for cross-modal learning, unifying conventional and scene text aware cross-modal retrieval tasks in a single framework. To handle images where scene text does not appear, we propose a fusion token based aggregation approach, sharing relevant information only through the fusion token, and a dual contrastive learning approach to enhance the visual features as well. Experimental results show the superior performance of ViSTA on both scene text aware retrieval and scene text free retrieval methods, which demonstrates the effectiveness of the proposed framework.

Note that the proposed approach can be also applied in other vision and language tasks when scene text is necessary as an additional modality. The contribution from scene text aggregation also depends on the percentage of images containing relevant scene text semantics and the correlation between visual appearance and scene text in a specific task.

Broader impacts. Since the proposed approach can be trained with a large amount of image and text pairs collected from the web, further data analysis, balancing and cleaning should be taken in production to mitigate the negative social impacts caused by distribution bias and mislabeled data.

References

[1] Xiang Bai, Mingkun Yang, Pengyuan Lyu, Yongchao Xu, and Jiebo Luo. Integrating scene text and visual appearance for fine-grained image classification. *IEEE Access*, 6:66322–66335, 2018. 3

[2] Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gómez i Bigorda, Marçal Rusiñol, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, pages 4290–4300. IEEE, 2019. 3

[3] 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 *CVPR*, pages 12652–12660. Computer Vision Foundation / IEEE, 2020. 1, 2, 7

[4] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *CVPR*, pages 15789–15798. Computer Vision Foundation / IEEE, 2021. 7

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: universal image-text representation learning. In *ECCV (30)*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer, 2020. 1, 2, 6, 7

[6] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *CVPR*, pages 8415–8424. Computer Vision Foundation / IEEE, 2021. 7

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics, 2019. 1, 4

[8] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *AAAI*, pages 1218–1226. AAAI Press, 2021. 2, 7

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net, 2021. 4

[10] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: improving visual-semantic embeddings with hard negatives. In *BMVC*, page 12. BMVA Press, 2018. 1, 2

[11] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomás Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013. 1, 2

[12] Lluís Gómez, Andrés Mafla, Marçal Rusiñol, and Dimosthenis Karatzas. Single shot scene text retrieval. In *ECCV (14)*, volume 11218 of *Lecture Notes in Computer Science*, pages 728–744. Springer, 2018. 3

[13] Google. Cloud Vision API, 2020(accessed June 3, 2020). 4

[14] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *CVPR*, pages 9989–9999. Computer Vision Foundation / IEEE, 2020. 4

[15] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *CVPR*, pages 12976–12985. Computer Vision Foundation / IEEE, 2021. 1, 2, 3, 7

[16] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *CoRR*, abs/2004.00849, 2020. 1, 2, 3, 6, 7

[17] Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, Zongzheng Xi, Yueqian Yang, Anwen Hu, Jinming Zhao, Ruichen Li, Yida Zhao, Liang Zhang, Yuqing Song, Xin Hong, Wanqing Cui, Dan Yang Hou, Yingyan Li, Junyi Li, Peiyu Liu, Zheng Gong, Chuhao Jin, Yuchong Sun, Shizhe Chen, Zhiwu Lu, Zhicheng Dou, Qin Jin, Yanyan Lan, Wayne Xin Zhao, Ruihua Song, and Ji-Rong Wen. Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *CoRR*, abs/2103.06561, 2021. 2

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 2021. 2, 5

[19] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676, 2017. 6

[20] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR, 2021. 1, 2, 3, 6, 7

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017. 2, 6

[22] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV (4)*, volume 11208 of *Lecture Notes in Computer Science*, pages 212–228. Springer, 2018. 1, 2, 6, 7

[23] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344. AAAI Press, 2020. 1, 2, 6, 7

[24] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *CoRR*, abs/2107.07651, 2021. 2

[25] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, pages 4653–4661. IEEE, 2019. 1, 2, 6, 7

[26] 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 (30)*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer, 2020. 1, 2

[27] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV (5)*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014. 6

[28] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *CVPR*, pages 10918–10927. Computer Vision Foundation / IEEE, 2020. 1, 2, 7

[29] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pages 13–23, 2019. 1, 2, 3, 6, 7

[30] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, pages 10434–10443. Computer Vision Foundation / IEEE, 2020. 7

[31] Andrés Mafla, Rafael Sampaio de Rezende, Lluís Gómez, Diane Larlus, and Dimosthenis Karatzas. Stacmr: Scene-text aware cross-modal retrieval. In *WACV*, pages 2219–2229. IEEE, 2021. 2, 3, 4, 6, 7

[32] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *CVPR*, pages 9826–9836. Computer Vision Foundation / IEEE, 2021. 7

[33] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR (Workshop Poster)*, 2013. 2

[34] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *NIPS*, volume 34, 2021. 5

[35] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, pages 1143–1151, 2011. 6

[36] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *CoRR*, abs/2001.07966, 2020. 1, 2, 6

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 2, 3

[38] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 6

[39] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: A dataset for image captioning with reading comprehension. In *ECCV (2)*, volume 12347 of *Lecture Notes in Computer Science*, pages 742–758. Springer, 2020. 3

[40] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *CVPR*, pages 8317–8326. Computer Vision Foundation / IEEE, 2019. 3

[41] 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*. OpenReview.net, 2020. 2, 7

[42] Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In *EMNLP/IJCNLP (1)*, pages 5099–5110. Association for Computational Linguistics, 2019. 2

[43] Hao Wang, Xiang Bai, Mingkun Yang, Shenggao Zhu, Jing Wang, and Wenyu Liu. Scene text retrieval via joint text detection and similarity learning. In *CVPR*, pages 4558–4567. Computer Vision Foundation / IEEE, 2021. 3

[44] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. Probing inter-modality: Visual parsing with self-attention for vision-language pre-training. *CoRR*, abs/2106.13488, 2021. 1, 2, 3, 7

[45] Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florêncio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. TAP: text-aware pre-training for text-vqa and text-caption. In *CVPR*, pages 8751–8761. Computer Vision Foundation / IEEE, 2021. 3

[46] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78, 2014. 6

[47] 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. AAAI Press, 2021. 1, 2, 7

[48] Gangyan Zeng, Yuan Zhang, Yu Zhou, and Xiaomeng Yang. Beyond OCR + VQA: involving OCR into the flow for robust and accurate textvqa. In *ACM Multimedia*, pages 376–385. ACM, 2021. 3

[49] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, pages 5579–5588. Computer Vision Foundation / IEEE, 2021. 1, 2

[50] Qi Zhu, Chenyu Gao, Peng Wang, and Qi Wu. Simple is not easy: A simple strong baseline for textvqa and textcaps. In *AAAI*, pages 3608–3615. AAAI Press, 2021. 3