# Exploiting Cross-Modal Prediction and Relation Consistency for Semisupervised Image Captioning

Yang Yang, Hongchen Wei, Hengshu Zhu, *Senior Member, IEEE*, Dianhai Yu, Hui Xiong, *Fellow, IEEE*, and Jian Yang, *Member, IEEE*

***Abstract*—The task of image captioning aims to generate captions directly from images via the automatically learned cross-modal generator. To build a well-performing generator, existing approaches usually need a large number of described images (i.e., supervised image-sentence pairs), requiring a huge effects on manual labeling. However, in real-world applications, a more general scenario is that we only have limited amount of described images and a large number of undescribed images. Therefore, a resulting challenge is how to effectively combine the undescribed images into the learning of cross-modal generator (i.e., *semisupervised image captioning*). To solve this problem, we propose a novel image captioning method by exploiting the cross-modal prediction and relation consistency (CPRC), which aims to utilize the raw image input to constrain the generated sentence in the semantic space. In detail, considering that the heterogeneous gap between modalities always leads to the supervision difficulty while using the global embedding directly, CPRC turns to transform both the raw image and corresponding generated sentence into the shared semantic space, and measure the generated sentence from two aspects: 1) *prediction consistency:* CPRC utilizes the prediction of raw image as soft label to distill useful supervision for the generated sentence, rather than employing the traditional pseudo labeling and 2) *relation consistency:* CPRC develops a novel relation consistency between augmented images and corresponding generated sentences to retain the important relational knowledge. In result, CPRC supervises the generated sentence from both the informativeness and representativeness perspectives, and can reasonably use the undescribed images to learn a more effective generator under the semisupervised scenario. The experiments show that our method outperforms state-of-the-art comparison methods on the MS-COCO "Karpathy" offline test split under complex nonparallel scenarios, for example, CPRC achieves at least 6% improvements on the CIDEr-D score.**

***Index Terms*—Cross-modal learning, image captioning, relation consistency, semisupervised learning.**

## I. Introduction

IN REAL-WORLD applications, object can always be represented by multiple source information, that is, multiple modalities [1], [2]. For example, the news always contains image and text information, the video can be divided into image, audio, and text modalities. Along this line, the study of cross-modal learning has emerged for bridging the connections among different modalities, so as to better perform downstream tasks, in which the image captioning is one of the important research directions. Specifically, image captioning aims to automatically generate natural language descriptions for images, and has emerged as a prominent research problem in both academia and industry [3]–[6]. For example, we can automatically broadcast road conditions by learning visual images to assist driving, and can also help visually impaired users to read more conveniently. In fact, the challenge of image captioning is to learn the generator between two heterogeneous modalities (i.e., the image and text modalities), which needs to recognize salient objects in an image using computer vision techniques and generate coherent descriptions using natural language processing.

To solve this problem, researchers first explore the neural encoder–decoder models [3], [7], which are composed of a CNN encoder and an LSTM (or transformer) decoder. In detail, these methods first encode the image into a set of feature vectors using a CNN based model, with each segmentation captures semantic information about an image region, then sequentially decode these feature vectors to words via an LSTM-based or transformer-based network. Furthermore, Xu *et al.* [4], Lu *et al.* [8], and Huang *et al.* [9] adopted the single or hierarchical attention mechanism that enables the model to focus on particular image regions during decoding process. To mitigate the incorrect or repetitive content,

Manuscript received 9 May 2021; revised 20 October 2021 and 2 January 2022; accepted 26 February 2022. Date of publication 27 July 2022; date of current version 17 January 2024. This work was supported in part by NSFC under Grant 62006118 and Grant 61906092; in part by the Natural Science Foundation of Jiangsu Province of China under Grant BK20200460 and Grant BK20190441; in part by the Jiangsu Shuangchuang (Mass Innovation and Entrepreneurship) Talent Program; in part by the Young Elite Scientists Sponsorship Program by CAST; and in part by the CCF-Baidu Open Fund under Grant CCF-BAIDU OF2020011. This article was recommended by Associate Editor Y. Xia. *(Corresponding author: Yang Yang.)*

Yang Yang is with the School of Computer Science and Engineering, Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Laboratory of Image and Video Understanding for Social Security, Nanjing University of Science and Technology, Nanjing 210094, China, also with the Key Laboratory of Computer Network and Information Integration, Ministry of Education, Southeast University, Nanjing 211189, China, and also with the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China (e-mail: yyang@njust.edu.cn).

Hongchen Wei and Jian Yang are with PCA Lab, Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Laboratory of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: weihc@njust.edu.cn; csjyang@njust.edu.cn).

Hengshu Zhu and Dianhai Yu are with Baidu Inc., Beijing 100000, China (e-mail: zhuhengshu@baidu.com; yudianhai@baidu.com).

Hui Xiong is with the Management Science and Information Systems Department, Rutgers Business School, Rutgers University, Newark, NJ 07102 USA (e-mail: hxiong@rutgers.edu).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TCYB.2022.3156367.

Digital Object Identifier 10.1109/TCYB.2022.3156367

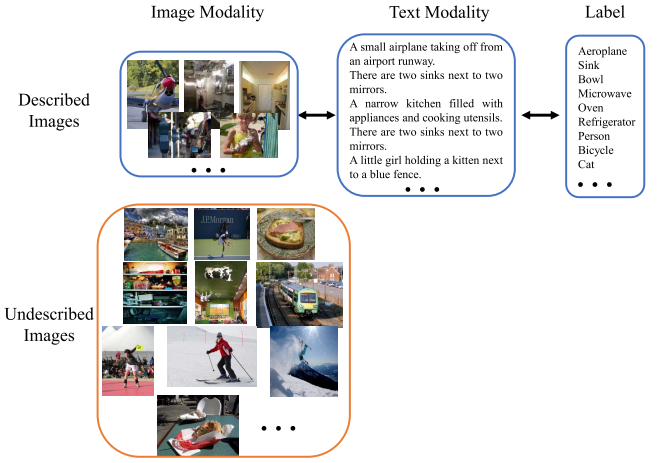



Fig. 1. Semisupervised image–sentence pairs, which include limited described images and a huge number of undescribed images. It is notable that we have two types of supervision: text and label ground truths. Described images have all the supervisions, whereas the undescribed images do not have any kind of supervision information here.

several researches consider to edit inputs independently from the problem of generating inputs [5], [10]. However, note that all these methods require full image–sentence pairs in advance, that is, all the images need to be described manually, which is hard to accomplish in real-world applications. Fig. 1 indicates a more general scenario: a limited amount of described images with corresponding label ground truths, and a large number of undescribed images. Therefore, a resulting challenge is the "*Semisupervised Image Captioning*," which aims to conduct the captioning task by reasonably utilizing the huge number of undescribed images.

The key difficulty of semisupervised image captioning is to design the pseudo supervision for the generated sentences. Actually, there have been some preliminary attempts recently. For example, [11], [12] propose unsupervised captioning methods, which combine the adversarial learning [13] with traditional encoder–decoder models to evaluate the quality of generated sentences. In detail, based on the traditional encoder–decoder models, these approaches employ adversarial training to generate sentences indistinguishable from the sentences within auxiliary corpus. In order to ensure that the generated captions contain the visual concepts, they particularly distill the knowledge provided by a visual concept detector into the image captioning model. However, the domain discriminator and visual concept distiller do not fundamentally evaluate the matching degree and structural rationality of the generated sentence, so the captioning performance is poor. As for semisupervised image captioning, a straightforward way is utilizing the undescribed images together with their machine-generated sentences directly [14], [15] as the pseudo image–sentence pair to train the model. However, limited amount of parallel data can hardly establish a proper initial generator to generate precisely pseudo descriptions, which may have negative affection to the training of mapping function.

To circumvent these issues, we attempt to utilize the raw image as pseudo supervision. However, heterogeneous gap between modalities always leads the supervision difficulty if we directly constrain the consistency between global embedding of image and sentence. Thereby, we switch to use the broader and more effective semantic prediction information, rather than directly utilize the embedding, and introduce a novel approach, dubbed semisupervised image captioning by exploiting the cross-modal prediction and relation consistency (CPRC). In detail, there are two common approaches in traditional semisupervised learning: 1) *pseudo labeling:* it minimizes the entropy of unlabeled data using predictions and 2) *consistency regularization:* it transforms the unlabeled raw images using data augmentation techniques, then constrains the consistency of transformed instances' outputs. Different form these two techniques, we design CPRC by comprehensively considering the informativeness and representativeness: 1) *prediction consistency:* we utilize the soft label of image to distill effective supervision for generated sentence and 2) *relation consistency:* we work on constraining the generated sentences and the augmented image inputs to have similar relational distributions. Consequently, CPRC can effectively qualify the generated sentences from perspectives of both the prediction confidence and distribution alignment, thereby to learn a more effective mapping function. Note that CPRC can be implemented with any current captioning model, and we adopt several typical approaches for verification [16], [17]. The source code is available at https://github.com/njustkmg/CPRC.

In summary, the contributions in this article can be summarized as follows.

1) We propose a novel semisupervised image captioning framework for processing undescribed images, which is universal for any captioning model.
2) We design the CPRC to measure the undescribed images, which maps the raw image and corresponding generated sentence into the shared semantic space, and supervises the generated sentence by distilling the soft label from image prediction and constraining the cross-modal relation consistency.
3) In experiments, our approach improves the performance under the semisupervised scenario, which validates that knowledge hidden in the content and relation is effective for enhancing the generator.

## II. Related Work

### A. Image Captioning

Image captioning approaches can be roughly divided into three categories as follows.

1) Template-based methods, which generate slotted captioning templates manually, and then utilize the detected keywords to fill the templates [18], but their expressive power is limited because of the need for designing templates manually.
2) Encoder–decoder-based methods, which are inspired by the neural machine translation [19]. For example, Vinyals *et al.* [20] proposed an end-to-end framework with a CNN encoding the image to feature vector and an LSTM decoding to caption; Huang *et al.* [9] added an attention-on-attention module after both the LSTM

and the attention mechanism, which can measure the relevance between attention result and query.

3) Editing-based methods, which consider editing inputs independent from generating inputs. For example, Hashimoto *et al.* [10] learned a retrieval model that embeds the input in a task-dependent way for code generation; Sammani and Elsayed [5] introduced a framework that learns to modify existing captions from a given framework by modeling the residual information.

However, all these methods need huge amount of supervised image–sentence pairs for training, whereas the scenario with a large amount of undescribed images is more general in real applications. To handle the undescribed images, several attempts propose unsupervised image captioning approaches. Feng *et al.* [11] distilled the knowledge in visual concept detector into the captioning model to recognize the visual concepts, and adopted sentence corpus to teach the captioning model; Gu *et al.* [12] developed an unsupervised feature alignment method with adversarial learning that maps the scene graph features from the image to sentence modality. Nevertheless, these methods mainly depend on employing the domain discriminator for learning plausible sentences, an thus are difficult for generating matched sentences. On the other hand, considering the semisupervised image captioning, Mithun *et al.* [14] and Huang *et al.* [15] proposed to extract regional semantics from un-annotated images as additional weak supervision to learn visual-semantic embeddings. However, the generated pseudo sentences are always unqualified to train the generator in real experiments.

### B. Semisupervised Learning

Recently, deep networks achieve strong performance by supervised learning, which requires a large number of labeled data. However, it comes at a significant cost when labeling by human labor, especially by domain experts. To this end, semisupervised learning, which permits harnessing the large amounts of unlabeled data in combination with typically smaller sets of labeled data, attracts more and more attention. Existing semisupervised learning mainly considers two aspects as follows.

1) *Self-Training [21]:* The generality of self-training is to use a model's predictions to obtain artificial labels for unlabeled data. A specific variant is the pseudo labeling, which converts the model predictions of unlabeled data to hard labels for calculating the cross entropy. Besides, pseudo labeling is often used along with a confidence thresholding that retains sufficiently confident unlabeled instances. In result, pseudo labeling results in entropy minimization, has been used as a component for many semisupervised algorithms, and been validated to produce better results [22].
2) *Consistency Regularization [23]:* Early extensions include exponential moving average of model parameters [24] or previous model checkpoints [25]. Recently, data augmentation, which integrates these techniques into the self-training framework, has shown better results [26], [27]. A mainstream technology is to produce random perturbations with data augmentation [28], and then enforce consistency between the augmentations. For example, Xie *et al.* [26] proposed unsupervised data augmentation with distribution alignment and augmentation anchoring, which encourages each output close to the weakly augmented version of the same input; Berthelot *et al.* [27] used a weakly augmented example to generate an artificial label and enforce consistency against strongly augmented example. Furthermore, Sohn *et al.* [29] combined the pseudo labeling and consistency regularization into a unified framework, which generates pseudo labels using the model's predictions on weakly augmented unlabeled images, and constrains the prediction consistency between weakly augmented and strongly augmented versions. Note that the targets in previous semisupervised methods are uniform and simple, that is, the label ground truths. However, cross-modal semisupervised learning is more complicated, that is, each image has the corresponding sentence and label ground truth. It is more difficult to build cross-modal generator than single modal classifier with limited supervised data, thereby it may causes noise accumulation if we directly employ the traditional semisupervised technique for the generated sentences.

The remainder of this article is organized as follows. Section III presents the proposed method, including the model, solution, and extension. Section IV shows the experimental results under different semisupervised settings. Section V concludes this article.

## III. Proposed Method

### A. Notations

Without any loss of generality, we define the semisupervised image–sentence set as: $\mathcal{D} = \{\{\mathbf{v}_i, \mathbf{w}_i, \mathbf{y}_i\}_{i=1}^{N_l}, \{\mathbf{v}_j\}_{j=1}^{N_u}\}$, where $\mathbf{v}_i \in \mathcal{R}^{d_v}$ denotes the $i$th image instance, $\mathbf{w}_i \in \mathcal{R}^{d_w}$ represents the aligned sentence instance, $\mathbf{y}_i \in \mathcal{R}^C$ denotes the instance label, $\mathbf{y}_{i,k} = 1$ if the $i$th instance belongs to the $k$th label, otherwise is 0. $\mathbf{v}_j$ is the $j$th undescribed image. $N_l$ and $N_u$ ($N_l \ll N_u$) are the numbers of described and undescribed instances, respectively.

*Definition 1 (Semisupervised Image Captioning):* Given limited parallel image–sentence pairs $\{\mathbf{v}_i, \mathbf{w}_i, \mathbf{y}_i\}_{i=1}^{N_l}$ and a huge number of undescribed images $\{\mathbf{v}_j\}_{j=1}^{N_u}$, we aim to construct a generator $G$ for image captioning by reliably utilizing the undescribed images.

### B. Framework

It is notable that CPRC focuses on employing the undescribed images, and is a general semisupervised framework. Thereby the image–sentence generator, that is, $G : \mathbf{v} \rightarrow \mathbf{w}$, can be represented as any state-of-the-art captioning model. In this article, considering the effectiveness and reproducibility, we adopt the attention model, that is, AoANet [9], for $G$ as the base model. As shown in Fig. 2, AoANet applies an attention on attention (AoA) module to the image encoder and the caption decoder. AoA first generates an "information

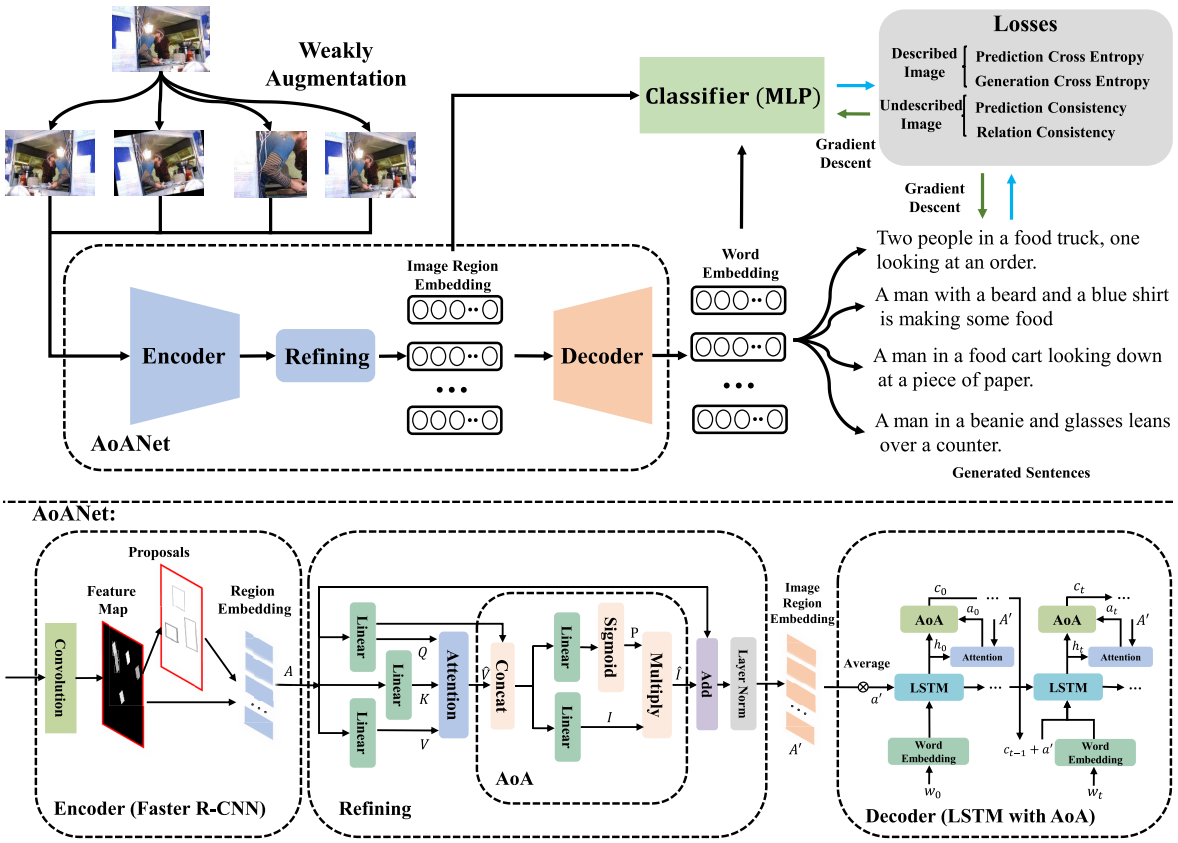



Fig. 2. Diagram of the proposed CPRC. For example, three weakly augmented images and the raw image are fed into the generator, which can generate four corresponding sentences. Without any loss of generality, we adopt the AoANet [9] as the generator. Then, the embeddings of image inputs and generated sentences are fed into the shared classifier to obtain the predictions. The model is trained by considering two objectives: 1) *supervised loss* includes the generation and prediction losses for described images. In detail, generation loss measures the quality of generated sentence sequence, and prediction loss considers the multilabel prediction of generated sentence and 2) *unsupervised loss* includes the prediction consistency and relation consistency for undescribed images. In detail, prediction consistency utilizes the image's prediction as pseudo labels for corresponding generated sentence, and relation consistency constrains the generated sentences' distribution with image inputs' distribution.

vector" (i.e., $I$) and an "attention gate" (i.e., $P$) with two linear transformations. The information vector is derived from the current context (i.e., the query $Q$) and the attention result (i.e., $\hat{V}$) via a linear transformation, and stores the newly obtained information from the attention result together with the information from the current context. The attention gate is also derived from the query and the attention result via another linear transformation with sigmoid activation followed. Subsequently, AoA adds another attention by applying the attention gate to the information vector using element-wise multiplication and finally obtains the "attended information" (i.e., $\hat{I}$). Details can refer to [9]. Note that the CPRC is a general framework, which can apply to any existing captioning model. AoANet only trains with the generation loss, which has not made effective use of category labels. Thereby, AoANet has no label predictor, and can not employ the undescribed images.

The learning process of CPRC is shown in Fig. 2. Specifically, CPRC first samples a minibatch of images from the dataset $\mathcal{D}$ (including described and undescribed images), and adopts the data augmentation techniques for each undescribed image (i.e., each image has $K$ variants). Then we can acquire the generated sentences for both augmented images and the raw image using the $G$, and compute the predictions for image inputs and generated sentences using the shared prediction classifier $f$. The model is trained through two main objects:

1) *supervised loss*, which is designed for described images, that is, supervised image–sentence pairs. In detail, supervised loss considers both the label and sentence predictions, including: a) *generation loss*, which employs the cross-entropy loss or reinforcement learning-based reward [16] of generated sentence and ground truth sentence and b) *prediction loss*, which calculates the multilabel loss between image/sentence's prediction and label ground truth.
2) *unsupervised loss*, which is designed for undescribed images. In detail, unsupervised loss considers both the informativeness and representativeness: a) *prediction consistency*, which uses the image's prediction as pseudo label to distill effective information for generated sentence, so as to measure the instance's informativeness and b) *relation consistency*, which adopts the relational structure of the augmented images as the supervision distribution for generated sentences, so as to measure the instance's representativeness. Therefore, in addition

to the traditional loss for described images, we constrain the sentences generated from undescribed images by comprehensively using the raw image inputs. The details are described as follows.

### C. Supervised Loss

*1) Generation Loss:* Given an image $\mathbf{v}$, the decoder generates a sequence of sentence $\hat{\mathbf{w}} = \{w_1, w_2, \ldots, w_T\}$ describing the image, $T$ is the length of sentence. Then, we can minimize the cross-entropy loss (i.e., $\ell_{\mathrm{XE}}$) or maximize a reinforcement learning-based reward [16] (i.e., $\ell_{\mathrm{RL}}$), according to ground truth caption $\mathbf{w}$

$$\begin{aligned}\ell_{\mathrm{XE}} &= -\sum_{t=1}^{T} \log p(\mathbf{w}_t|\mathbf{w}_{1:t-1}) \\ \ell_{\mathrm{RL}} &= -\mathbb{E}_{\mathbf{w}_{1:T}}\, p[r(\mathbf{w}_{1:T})]\end{aligned} \tag{1}$$

where $\mathbf{w}_{1:T}$ denotes the target ground truth sequence, $p(\cdot)$ is the prediction probability. The reward $r(\cdot)$ is a sentence-level metric for the sampled sentence and the ground truth, which always uses the score of some metric (e.g., CIDEr-D [30]). In detail, as introduced in [16], captioning approaches traditionally train the models using the cross-entropy loss. On the other hand, to directly optimize NLP metrics and address the exposure bias issue, [16] casts the generative models in the Reinforcement Learning terminology as [31]. In detail, traditional decoder (i.e., LSTM) can be viewed as an "agent" that interacts with the "environment" (i.e., words and image features). The parameters of the network define a policy, that results in an "action" (i.e., the prediction of the next word). After each action, the agent updates its internal "state" (i.e., parameters of the LSTM, attention weights etc). Upon generating the end-of-sequence (EOS) token, the agent observes a "reward" that is, for example, the CIDEr-D score of the generated sentence.

*2) Prediction Loss:* On the other hand, we can measure the generation with classification task using label ground truth $\mathbf{y}$. We extract the embeddings of image input and generated sentence from the representation output layer. Considering that the image and corresponding sentence share the same semantic representations, the embeddings of image input and generated sentence can be further put into the shared classifier $f$ for predicting. Thereby, the forward prediction process can be represented as

$$\mathbf{p}^v = f(E_e(\mathbf{v})), \quad \mathbf{p}^w = f(D_e(E_e(\mathbf{v})))$$

where $\mathbf{p}^v$ and $\mathbf{p}^w$ are normalized prediction distribution of image input and generated sentence. $f(\cdot)$ denotes the shared classification model for text and image modalities. Without any loss of generality, we utilize a network with three fully connected layers here. $E_e$ denotes the encoder, $D_e$ represents the decoder. $E_e(\mathbf{v}), D_e(E_e(\mathbf{v})) \in \mathcal{R}^d$ represents the embeddings of image input and generated sentence. Note that $E_e(\mathbf{v})$ and $D_e(E_e(\mathbf{v}))$ are the final embeddings of image/text region embedding with *mean*$(\cdot)$ operator. The commonly used image captioning dataset (i.e., the COCO dataset) is a multilabel dataset, that is, different from the multiclass dataset that each instance only has one ground truth, each instance has multiple labels. Therefore, we utilize the binary cross-entropy loss (BCELoss) here

$$\begin{aligned}\ell_p &= \sum_{m\in\{v,w\}} H\left(\mathbf{p}^m, \mathbf{y}^m\right) \\ H\left(\mathbf{p}^m, \mathbf{y}^m\right) &= -\sum_j \left(y_j^m \log p_j^m + \left(1 - y_j^m\right) \log\left(1 - p_j^m\right)\right)\end{aligned} \tag{2}$$

where $H(\cdot)$ denotes the BCELoss for multilabel prediction, and the model's predictions are encouraged to be low-entropy on supervised data.

### D. Unsupervised Loss

*1) Prediction Consistency:* First, we introduce the augmentation technique for transforming the images. Existing methods usually leverage two kinds of augmentations: 1) weak augmentation is a standard flip-and-shift strategy, which does not significantly change the content of the input and 2) strong augmentation always refers to the AutoAugment [32] and its variant, which uses reinforcement learning to find an augmentation strategy comprising transformations from the Python Imaging Library.[1] Considering that "strong" augmented (i.e., heavily augmented) instances are almost certainly outside the data distribution, which leads to the low quality of generated sentence, we leverage the "weak" augmentation instead. In result, each image can be expanded to $K+1$ variants, that is, $\Psi(\mathbf{v}) = \{\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_K\}$, $\mathbf{v}_0$ denotes the raw input.

Then, we input the augmented image set to the image–sentence generator $G$, and extract the embeddings of generated sentences from the representation output layer. The embeddings are further put into the shared classifier for prediction. Thereby, the prediction process can be represented as

$$\mathbf{p}_k^w = f(D_e(E_e(\mathbf{v}_k))), \quad k \in \{0, 1, \ldots, K\}. \tag{3}$$

Similarly, we can acquire the prediction of image inputs: $\mathbf{p}_k^v = f(E_e(\mathbf{v}_k)), \quad k \in \{0, 1, \ldots, K\}$. Considering that the commonly used image captioning datasets are multilabel datasets, traditional pseudo labeling that leverages "hard" labels (i.e., the arg max of model's output) is inappropriate, because it is difficult to determine the number of hard labels for each instance. As a consequence, we directly utilize the prediction of image for knowledge distillation [33]

$$\begin{aligned}\ell_{pc} &= \sum_{k\in\{0,1,\ldots,K\}} H\left(\mathbf{p}_k^v, \mathbf{p}_k^w\right) \\ H\left(\mathbf{p}_k^v, \mathbf{p}_k^w\right) &= -\sum_j \left(p_{k_j}^v \log p_{k_j}^w + \left(1 - p_{k_j}^v\right) \log\left(1 - p_{k_j}^w\right)\right)\end{aligned} \tag{4}$$

where $H(\cdot)$ denotes the BCELoss.

*2) Relation Consistency:* Inspired by the linguistic structuralism [34] that relations can better present the knowledge than individual example, the primary information actually lies in the structure of the data space. Therefore, we define a new relation consistency loss $\ell_{rc}$ using a metric learning-based constraint, which calculates the KL divergence of the similarity vectors between the image inputs and generated sentences. The relation consistency aims to ensure the structural knowledge

[1] https://www.pythonware.com/products/pil/

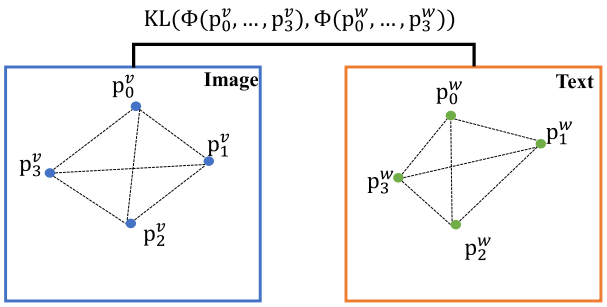



Fig. 3. Relation consistency. The blue and orange rectangles represent the image domain and text domain, respectively. Any point inside the rectangles represents a specific instance in that domain. Relation Consistency: for example, given a tuple of image instances $\{\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$, relation consistency loss requires that the generated sentences, $\{\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3\}$, should share the similar relation structure with the raw inputs.

using mutual relations of data examples in the raw inputs. Specifically, each image input can be denoted as a bag of $K+1$ instances, that is, $\Psi(\mathbf{v})$, while the corresponding generated sentences can also be represented as a bag of instances, that is, $G(\Psi(\mathbf{v}))$. With the shared classifier, the image and sentence prediction can be formulated as

$$\mathbf{p}_k^v = f(E_e(\mathbf{v}_k)), \quad k \in \{0, 1, \ldots, K\}$$
$$\mathbf{p}_k^w = f(D_e(E_e(\mathbf{v}_k))), \quad k \in \{0, 1, \ldots, K\}.$$

With the predictions of image inputs and generated sentences, the objective of relation consistency can be formulated as

$$\ell_{rc} = \mathrm{KL}\big(\Phi\big(\mathbf{p}_0^v, \mathbf{p}_1^v, \ldots, \mathbf{p}_K^v\big), \Phi\big(\mathbf{p}_0^w, \mathbf{p}_1^w, \ldots, \mathbf{p}_K^w\big)\big). \quad (5)$$

$\mathrm{KL}(a, b) = a\log(a/b)$ is the KL divergence that penalizes difference between the similarity distributions of image inputs and generated sentences. $\Phi$ is a relation function, which measures a relation energy of the given tuple. In detail, $\Phi$ aims to measure the similarities formed by the examples in semantic prediction space

$$\Phi\big(\mathbf{p}_0^v, \mathbf{p}_1^v, \ldots, \mathbf{p}_K^v\big) = \big[q_{mn}^v\big] \quad m, n \in [0, \ldots, K]$$
$$\Phi\big(\mathbf{p}_0^w, \mathbf{p}_1^w, \ldots, \mathbf{p}_K^w\big) = \big[q_{mn}^w\big] \quad m, n \in [0, \ldots, K]$$
$$q_{mn}^v = \frac{\exp\big(d_{mn}^v\big)}{\sum \exp\big(d_{\cdot}^v\big)}$$
$$q_{mn}^w = \frac{\exp\big(d_{mn}^w\big)}{\sum \exp\big(d_{\cdot}^w\big)} \quad (6)$$

where $d_{mn}^v = -\mathrm{Dist}(\mathbf{p}_m^v, \mathbf{p}_n^v)$, $d_{mn}^w = -\mathrm{Dist}(\mathbf{p}_m^w, \mathbf{p}_n^w)$ measures the distance between $(\mathbf{p}_m^v, \mathbf{p}_n^v)$, $(\mathbf{p}_m^w, \mathbf{p}_n^w)$, respectively, $\mathrm{Dist}(\mathbf{p}_m^v, \mathbf{p}_n^v) = \|\mathbf{p}_m^v - \mathbf{p}_n^v\|_2$ and $\mathrm{Dist}(\mathbf{p}_m^w, \mathbf{p}_n^w) = \|\mathbf{p}_m^w - \mathbf{p}_n^w\|_2$. $q_{mn}^v$ and $q_{mn}^w$ denote the relative instance-wise similarity. Finally, we pull the $[q_{mn}^v]$ and $[q_{mn}^w]$ into vector form. In result, the relation consistency loss can deliver the relationship of examples by penalizing structure differences. Since the structure has higher order properties than single output, it can transfer knowledge more effectively, and is more suitable for consistency measure.

### *E. Overall Function*

In summary, with the limited amount of parallel image–sentence pairs and large amount of undescribed images, we

**Algorithm 1** Code of CPRC

**Input**:
Data: $\mathcal{D} = \{\{\mathbf{v}_i, \mathbf{w}_i, \mathbf{y}_i\}_{i=1}^{N_l}, \{\mathbf{v}_j\}_{j=1}^{N_u}\}$
Parameters: $\lambda_1$, $\lambda_2$, $\tau$, epoch number $T$
**Output**:
Image captioning mapping function: $G$

```
Initialize the G and f randomly;
for t = 1:T do
    for sample mini-batch B_{t,k} from D do
        Calculate supervised loss ℓ_s with described images according to Equation 1 or Equation 2;
        Calculate prediction consistency ℓ_pc with undescribed images according to Equation 4;
        Calculate relation consistency ℓ_rc with undescribed images according to Equation 5;
        Calculate overall loss L with ℓ_s, ℓ_pc and ℓ_rc according to Equation 8;
        Update model parameters of G, f using SGD;
    end for
end for
```

define the total loss by combining (1), (2), (4), and (5)

$$L = \sum_{i=1}^{N_l} \ell_s(\mathbf{v}_i, \mathbf{w}_i, \mathbf{y}_i) + \sum_{j=1}^{N_u} \big(\lambda_1 \ell_{pc}(\mathbf{v}_j) + \lambda_2 \ell_{rc}(\mathbf{v}_j)\big)$$
$$\ell_s(\mathbf{v}_i, \mathbf{w}_i, \mathbf{y}_i) = \ell_c(\mathbf{v}_i, \mathbf{w}_i) + \ell_p(\mathbf{v}_i, \mathbf{w}_i, \mathbf{y}_i) \quad (7)$$

where $\ell_c$ denotes the captioning loss, which can be adopted as $\ell_{\mathrm{XE}}$ or $\ell_{\mathrm{RL}}$ in (1). Note that $\ell_c$ and $\ell_p$ are with same order of magnitude, so we do not add hyperparameter here. $\lambda_1$ and $\lambda_2$ are scale values that control the weights of different losses. In $\ell_s$, we use labeled images and sentences to jointly train the shared classifier $f$, which increases the amount of training data, as well as adjusts the classifier to better suit subsequent prediction of augmented images and generated sentences. Furthermore, considering that the pseudo labels $\mathbf{p}^v$ and $\mathbf{p}^w$ may exist noises, we can also adopt a confidence threshold that retains confident instances. Equation (7) can be reformulated as

$$L = \sum_{i=1}^{N_l} \ell_s(\mathbf{v}_i, \mathbf{w}_i, \mathbf{y}_i) + \sum_{j=1}^{N_u} \mathbf{1}\Big(\max\Big(\mathbf{p}_{j_0}^v\Big) \geq \tau\Big) \big\{\lambda_1 \ell_{pc}(\mathbf{v}_j) + \lambda_2 \ell_{rc}(\mathbf{v}_j)\big\}$$
$$\ell_s(\mathbf{v}_i, \mathbf{w}_i, \mathbf{y}_i) = \ell_c(\mathbf{v}_i, \mathbf{w}_i) + \ell_p(\mathbf{v}_i, \mathbf{w}_i, \mathbf{y}_i) \quad (8)$$

where $\mathbf{p}_{j_0}^v$ denotes the prediction probability of the $j$th raw image input, and $\tau$ is a scalar hyperparameter denoting the threshold above which we retain the generated sentences. $\mathbf{1}(\cdot)$ denotes the indicator function. Details are shown in Algorithm 1. In each epoch, we randomly sample minibatch data containing described and undescribed images, then utilize the data in the minibatch to calculate both the supervised and unsupervised losses. Finally, we calculate the gradient using overall loss for back propagation. Loop the entire process until reaching the defined epochs.

## IV. Experiments

### A. Datasets

We adopt the popular MS-COCO dataset [35] as mostly former related methods [9], [16], [17], [36], [37]. The MS-COCO dataset contains 123 287 images (82 783 training images and 40 504 validation images), each labeled with five captions. The popular test sets are divided into two categories: 1) online evaluation and 2) offline evaluation. Considering that all methods are evaluated under semisupervised scenario, online evaluation cannot be used, so we only use offline evaluation. The offline Karpathy data split [38] contains 5000 images for validation, 5000 images for testing, and the rest for training. To construct the semisupervised scenario, we randomly selected examples with artificially set proportions as supervised data from the training set, and the rest are unsupervised data.

### B. Implementation Details

The target of CPRC is to train the generator $G$. In detail, we employ the AoANet [9] structure for $G$ as base model. Meanwhile, we adopt fully connected networks for $f$ with three fully connected layers (with 1024 dimension for the hidden layers). The dimension of original image vectors is 2048 and we project them to a new space with the dimension of 1024 following [9]. The $K = 3$, that is, each image has three augmentations using random occlusion technique. As for the training process, we train AoANet for 40 epochs with a minibatch size of 16, and the ADAM [39] optimizer is used with a learning rate initialized by $10^{-4}$ and annealed by 0.8 every three epochs. The parameters $\lambda_1$ and $\lambda_2$ are tuned in $\{0.01, 0.1, 1, 10\}$, and $\tau = 0.1$. The entire network is trained on an Nvidia TITAN X GPU.

### C. Baselines and Evaluation Protocol

The comparison models fall into three categories: 1) *state-of-the-art supervised captioning methods:* SCST [16], AoANet [9], AAT [36], ORT [37], GIC [17], Anchor [40] and RSTNet [41]. Note that these methods can only utilize the supervised image–sentence pairs; 2) *unsupervised captioning methods:* Graph-align [12] and UIC [11]. These approaches utilize the independent image set and corpus set for training; and 3) *state-of-the-art semisupervised method:* A3VSE [15].

Moreover, we conduct extra ablation studies to evaluate each term in our proposed CPRC: 1) AoANet+P, we combine the label prediction consistency with the original AoANet generation loss as multitask loss (only using the supervised data); 2) AoANet+C, we combine the relation consistency loss with the original AoANet generation loss as multitask loss (only using the supervised data); 3) PL, we replace the prediction consistency with pseudo labeling as traditional semisupervised methods; 4) AC, we replace the relation consistency with augmentation consistency as traditional semisupervised methods; 5) Embedding+, we replace the relation consistency loss with embedding consistency loss, which minimizes the difference between the embedding of image inputs and generated sentences; 6) Semantic+, we replace the relation consistency loss with prediction consistency loss, which minimizes the difference between the predictions of image inputs and generated sentences; 7) Strong+, we replace the weak augmentation with strong augmentation for CPRC; 8) w/o Prediction, CPRC only retains the relation consistency loss in (8); 9) w/o Relation, CPRC only retains the prediction consistency in (8); and 10) w/o $\tau$, CPRC removes the confidence threshold in (7). For evaluation, we use different metrics, including BLEU [42], METEOR [43], ROUGE-L [9], CIDEr-D [30], and SPICE [44], to evaluate the proposed method and comparison methods. All the metrics are computed with the publicly released code.[2] In fact, the CIDEr-D and SPICE metric is more suitable for the image captioning task [30], [44]. One of the problems with using metrics, such as BLEU, ROUGE-L, CIDEr-D, and METEOR is that these metrics are primarily sensitive to $n$-gram overlap. However, $n$-gram overlap is neither necessary nor sufficient for two sentences to convey the same meaning [45].

### D. Qualitative Analysis

Table I presents the quantitative comparison results with state-of-the-art methods (i.e., 1% supervised data and 99% unsupervised in the training set), it is notable that supervised captioning methods can only develop the mapping functions with supervised data, and leave out the unsupervised data. For fairness, all the models are first trained under cross-entropy loss and then optimized for CIDEr-D score as [9]. "−" represents the results have not given in the raw paper. The results reveal the following.

1) Unsupervised approach, that is, UIC, achieve the worst performance on all metrics under different losses. This phenomenon verifies that the generated sentence may mismatch the image with a high probability when only considering the domain discriminator. Graph-align performs better than supervised approaches, but worse than A3VSE on most metrics, because it ignores to measure specific matching.
2) Semisupervised method, that is, A3VSE, has little effect on improving the captioning performance, for example, cross-entropy loss/CIDEr-D score optimization only improves 0.4/2.0 and 0.2/0.1 on CIDEr-D and SPICE scores compared to AoANet, because it is more difficult to ensure the quality of generated sentences.
3) CPRC achieves the highest scores among all compared methods in terms of all metrics, on both the cross-entropy loss and CIDEr-D score optimization stage, except ROUGE-L on cross-entropy loss. For example, CPRC achieves a state-of-the-art performance of 77.9/78.8 (CIDEr-D score) and 16.2/16.8 (SPICE score) under two losses (cross entropy and CIDEr-D score), that acquires 8.4/8.1 and 1.9/1.5 improvements compared to RSTNet (i.e., SOTA supervised model), 8.4 and 1.8 improvements compared to Graph-align (i.e., SOTA unsupervised model), and 8.3/6.4 and 1.7/1.5 improvements compared to A3VSE (i.e., SOTA semisupervised model). The phenomena indicates that, with limited amount of supervised data, existing methods cannot construct a well mapping function, whereas CPRC can reliably utilize the undescribed image to enhance the model.

[2] https://github.com/tylin/coco-caption

TABLE I
PERFORMANCE OF COMPARISON METHODS ON MS-COCO "KARPATHY" TEST SPLIT, WHERE B@N, M, R, C, AND S ARE SHORT FOR BLEU@N, METEOR, ROUGE-L, CIDER-D, AND SPICE SCORES

| Methods | 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 | 56.8 | 38.6 | 25.4 | 16.3 | 16.0 | 42.4 | 38.9 | 9.3 | 59.4 | 39.5 | 25.3 | 16.3 | 17.0 | 42.9 | 43.7 | 9.9 |
| AoANet | 67.9 | 49.8 | 34.7 | 23.2 | 20.9 | 49.2 | 69.2 | 14.3 | 66.8 | 48.6 | 34.1 | 23.6 | 21.8 | 48.7 | 70.4 | 15.2 |
| AAT | 63.2 | 45.8 | 31.7 | 21.3 | 19.0 | 47.6 | 58.0 | 12.4 | 66.7 | 48.1 | 33.3 | 22.7 | 20.4 | 47.8 | 63.5 | 13.2 |
| ORT | 63.6 | 45.8 | 31.7 | 21.4 | 19.4 | 46.9 | 61.1 | 12.6 | 65.3 | 46.5 | 31.9 | 21.3 | 20.3 | 47.2 | 62.0 | 13.3 |
| GIC | 63.0 | 46.8 | 33.2 | 20.0 | 19.2 | 50.3 | 50.5 | 12.3 | 64.7 | 46.9 | 32.0 | 20.7 | 19.0 | 47.8 | 55.7 | 12.5 |
| Anchor | 56.5 | 38.1 | 24.9 | 16.5 | 16.2 | 42.2 | 40.5 | 10.5 | 57.2 | 38.9 | 25.5 | 17.3 | 16.7 | 42.9 | 45.6 | 10.7 |
| RSTNet | 68.1 | 50.2 | 34.8 | 23.4 | 20.9 | 49.6 | 69.5 | 14.3 | 67.3 | 48.9 | 34.5 | 23.6 | 22.0 | 48.8 | 70.7 | 15.3 |
| Graph-align | - | - | - | - | - | - | - | - | 67.1 | 47.8 | 32.3 | 21.5 | 20.9 | 47.2 | 69.5 | 15.0 |
| UIC | - | - | - | - | - | - | - | - | 41.0 | 22.5 | 11.2 | 5.6 | 12.4 | 28.7 | 28.6 | 8.1 |
| A3VSE | 68.0 | 50.0 | 34.9 | 23.3 | 20.8 | 49.3 | 69.6 | 14.5 | 67.6 | 49.6 | 35.2 | 24.5 | 22.1 | 49.3 | 72.4 | 15.3 |
| AoANet+P | 67.4 | 49.7 | 35.2 | 24.3 | 22.3 | 49.1 | 71.7 | 14.9 | 67.2 | 49.5 | 35.9 | 24.4 | 21.6 | 50.1 | 74.2 | 15.7 |
| AoANet+C | 67.1 | 49.4 | 35.2 | 24.5 | 22.7 | 49.5 | 71.5 | 14.9 | 67.8 | 49.4 | 35.5 | 24.7 | 22.0 | 50.0 | 73.9 | 15.6 |
| PL | 67.8 | 49.6 | 35.2 | 24.2 | 22.0 | 50.4 | 74.7 | 15.6 | 67.9 | 50.0 | 35.6 | 24.3 | 22.2 | 49.7 | 76.6 | 16.1 |
| AC | 67.8 | 48.8 | 34.6 | 23.7 | 21.9 | 49.1 | 69.7 | 14.5 | 67.9 | 50.0 | 25.3 | 24.1 | 22.1 | 49.7 | 73.0 | 15.5 |
| Embedding+ | 65.1 | 46.4 | 31.9 | 21.5 | 20.7 | 47.6 | 65.1 | 14.1 | 65.6 | 47.1 | 32.3 | 22.6 | 20.8 | 47.8 | 69.1 | 14.5 |
| Semantic+ | 68.3 | 49.9 | 34.9 | 23.8 | 21.5 | 49.9 | 70.3 | 14.7 | 69.3 | 50.8 | 35.5 | 24.1 | 21.6 | 50.0 | 72.7 | 14.9 |
| Strong+ | 68.4 | 50.8 | 35.4 | 24.8 | 22.5 | **50.6** | 77.8 | 16.2 | 69.5 | 51.5 | 36.7 | 25.5 | 23.3 | 50.6 | 78.6 | 16.7 |
| w/o Prediction | 68.3 | 49.6 | 35.3 | 24.4 | 22.2 | 49.6 | 70.5 | 15.0 | 68.2 | 50.4 | 35.8 | 24.8 | 22.5 | 50.1 | 73.6 | 15.6 |
| w/o Relation | 68.1 | 50.0 | 35.5 | 24.8 | 22.4 | 50.5 | 75.2 | 15.8 | 68.3 | 50.5 | 35.8 | 24.9 | 22.7 | 50.4 | 76.9 | 16.3 |
| w/o $\tau$ | 66.9 | 49.8 | 34.5 | 24.2 | 21.5 | 49.5 | 76.2 | 15.4 | 68.5 | 50.8 | 36.2 | 25.0 | 22.5 | 49.8 | 77.5 | 16.2 |
| CPRC | **68.8** | **51.1** | **35.5** | **24.9** | **22.8** | 50.4 | **77.9** | **16.2** | **69.9** | **51.8** | **36.7** | **25.5** | **23.4** | **50.7** | **78.8** | **16.8** |

4) CPRC performs better than w/o $\tau$ on all metrics, which indicates the effectiveness of threshold confidence.

### *E. Ablation Study*

To quantify the impact of the proposed CPRC modules, we compare CPRC against other variant models under various settings. The bottom half of Table I presents the results.

1) AoANet+P and AoANet+C achieve better performance than AoANet, which indicates that the prediction loss and relation consistency loss can improve the generator learning, because the labels can provide extra semantic information; meanwhile, AoANet+P performs better than AoANet+C on most metric, which indicates that prediction loss is more significant than relation consistency.
2) PL and AC perform worse than the w/o Prediction and w/o Relation, which verifies that traditional semisupervised techniques considering pseudo labeling are not as good as cross-modal semisupervised techniques considering raw image as pseudo supervision.
3) Embedding+ performs worse than the Semantic+, which reveals that embeddings are more difficult to compare than predictions since image and text have heterogeneous representations.
4) Strong+ performs worse than CPRC, which validates that the strong augmentation may impacts the generated sentence, and further affects the prediction.
5) Both the w/o Prediction and w/o Relation can improve the captioning performance on most criteria, especially on the important criteria, that is, CIDEr-D and SPICE. The results indicate that both the prediction and relation consistencies can provide effective supervision to ensure the quality of generated sentences.

TABLE II
PERFORMANCE OF CPRC WITH DIFFERENT CAPTION MODEL ON MS-COCO KARPATHY TEST SPLIT, WHERE B@N, M, R, C, AND S ARE SHORT FOR BLEU@N, METEOR, ROUGE-L, CIDER-D, AND SPICE SCORES

| Methods | Cross Entropy Loss | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | B@1 | B@2 | B@3 | B@4 | M | R | C | S |
| SCST | 56.8 | 38.6 | 25.4 | 16.3 | 16.0 | 42.4 | 38.9 | 9.3 |
| GIC | 63.0 | 46.8 | 33.2 | 20.0 | 19.2 | 50.3 | 50.5 | 12.3 |
| SCST+CPRC | **63.5** | **45.9** | **31.7** | **21.6** | **19.4** | **45.8** | **48.1** | **10.2** |
| GIC+CPRC | **66.8** | **47.5** | **34.5** | **21.4** | **19.2** | **50.8** | **57.7** | **13.4** |
| Methods | CIDEr-D Score Optimization | | | | | | | |
| | B@1 | B@2 | B@3 | B@4 | M | R | C | S |
| SCST | 59.4 | 39.5 | 25.3 | 16.3 | 17.0 | 42.9 | 43.7 | 9.9 |
| GIC | 64.7 | 46.9 | 32.0 | 20.7 | 19.0 | 47.8 | 55.7 | 12.5 |
| SCST+CPRC | **66.5** | **48.0** | **33.7** | **22.7** | **20.4** | **47.9** | **48.7** | **10.7** |
| GIC+CPRC | **66.9** | **47.9** | **34.8** | **21.8** | **19.8** | **48.2** | **58.9** | **13.6** |

6) The effect of w/o Relation is more obvious, which shows that prediction loss can further improve the scores by comprehensively considering the semantic information.
7) CPRC achieves the best scores on most metrics, which indicates that it is better to combine the content and relation information.

### *F. CPRC With Different Captioning Model*

To explore the generality of CPRC, we conduct more experiments by incorporating CPRC with different supervised captioning approaches, that is, SCST (encoder–decoder-based model) and GIC (attention-based model). Note that we have not adopted the editing-based method considering the reproducibility, the results are recorded in Table II. We find that all

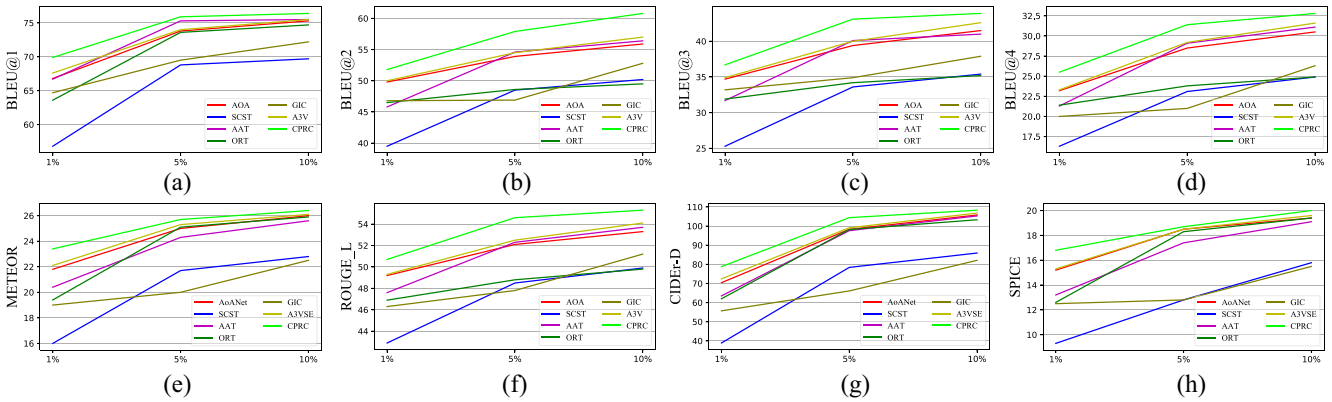



Fig. 4. Relationship between captioning performance with different ratio of supervised data. (a) BLEU@1. (b) BLEU@2. (c) BLEU@3. (d) BLEU@4. (e) METEOR. (f) ROUGE-L. (g) CIDEr-D. (h) SPICE.

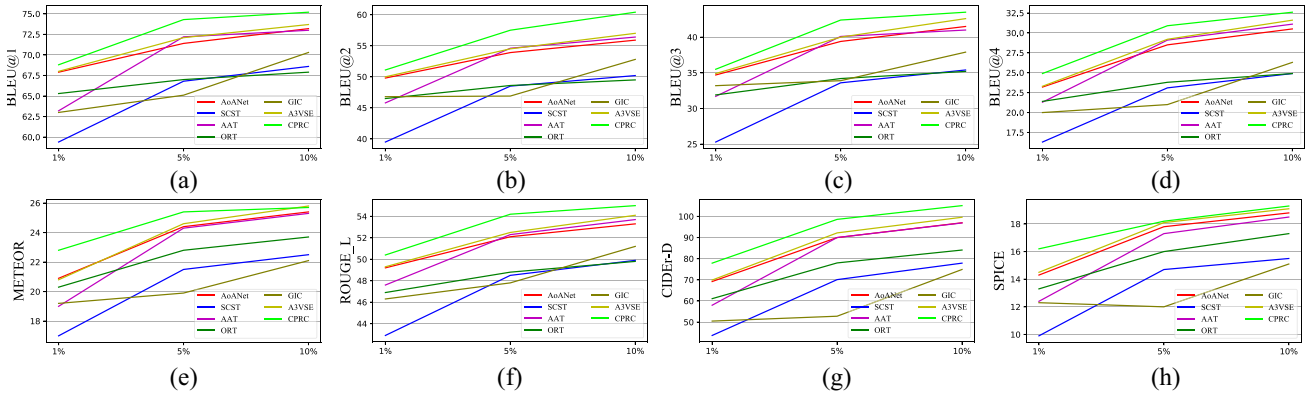



Fig. 5. Relationship between caption performance with different ratio of supervised data (cross-entropy loss). (a) BLEU@1. (b) BLEU@2. (c) BLEU@3. (d) BLEU@4. (e) METEOR. (f) ROUGE-L. (g) CIDEr-D. (h) SPICE.

methods, that is, SCST, GIC, and AoANet (results can refer to the Table I), have improved the performance after combing the CPRC framework. This phenomenon validates that CPRC can well combine the undescribed images for existing supervised captioning models.

### G. Influence of the Supervised and Unsupervised Images

To explore the influence of supervised data, we tune the ratio of supervised data, and the results are recorded in Figs. 4 and 5 with different metrics. Here, we find that, with the percentage of supervised data increases, the performance of CPRC improves faster than other state-of-the-art methods. This phenomenon indicates that CPRC can reasonably utilize the undescribed images to improve the learning of generator. Furthermore, we validate the influence of unsupervised data, that is, we fix the supervised ratio to 1%, and tune the ratio of unsupervised data in $\{10\%, 40\%, 70\%, 100\%\}$, the results are recorded in Fig. 6. Note that one of the problems by using metrics, such as BLEU, ROUGE-L, CIDEr-D, and METEOR to evaluate captions, is that these metrics are primarily sensitive to $n$-gram overlap [9], [44]. Therefore, we only give the results of CIDEr-D and SPICE here (refer to the supplementary material for more details). We find that with the percentage of unsupervised data increases, the performance of CPRC also improves. This phenomenon indicates that CPRC can make full use of undescribed images for positive training.

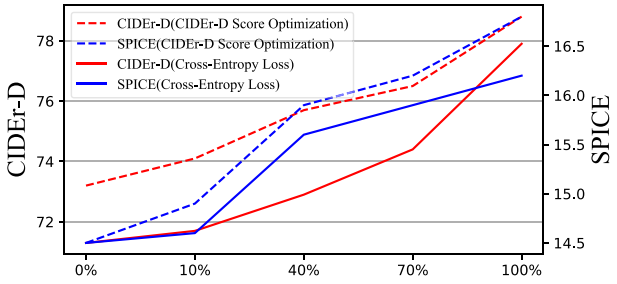



Fig. 6. Relationship between captioning performance with different ratio of unsupervised data (CIDEr-D Score Optimization).

### H. Influence of the Augmentation Number

To explore the influence of augmentation number, that is, $K$, we conduct more experiments. In detail, we tune $K$ in $\{1, 2, 3, 4, 5\}$ and recorded the results in Table III. The results reveal that the CPRC achieves the best performance with $K = 3$, for the reason that additional inconsistent noises between images and sentences may be introduced with the number of augmentations increases.

### I. Influence of the Confidence Threshold

To explore the influence of confidence threshold, that is, $\tau$, we conduct more experiments. In detail, we tune the $\tau$ in $\{0, 0.1, 0.4, 0.7\}$ and recorded the results in Table IV. The

**CPRC:** a man doing a trick on a skateboard on a rail.
**AoANet:** a man on a skateboard on a skateboard.
**A3VSE:** a snowboarder is riding a trick on a skateboard.
**GT1:** The skateboarder is putting on a show using the picnic table as his stage.
**GT2:** A skateboarder pulling tricks on top of a picnic table.
**GT3:** A man riding on a skateboard on top of a table.
**GT4:** A skate boarder doing a trick on a picnic table.
**GT5:** A person is riding a skateboard on a picnic table with a crowd watching.

**CPRC:** a bathroom with a white sink and a toilet in it.
**AoANet:** a bathroom with a sink and a sink.
**A3VSE:** a bathroom with a toilet sink and sink.
**GT1:** A bathroom with a toilet, sink, bath tub, and a small cabinet.
**GT2:** a bath room with a bath tub a toilet and a sink.
**GT3:** A bathroom with lighting and a toilet and other things.
**GT4:** A large whirlpool tub in the corner of a tiled bathroom.
**GT5:** A white bathroom with all of the necessities.

**CPRC:** a dog is standing next to a brown.
**AoANet:** a close up of a dog in a chair.
**A3VSE:** a dog is standing on top of a boat.
**GT1:** A dog sticking its head into a small pet bed.
**GT2:** a little dog checking out a pet bed.
**GT3:** A small dog biting a small dog bed.
**GT4:** A small dog sniffs the inside of its bed.
**GT5:** A dog small dog sniffing a dog bed.

**CPRC:** a pizza sitting on top of a wooden cutting board.
**AoANet:** a pizza with cheese on a wooden board.
**A3VSE:** a pizza sitting on top of a cutting board.
**GT1:** A colorful pizza sits on a wooden cutting board.
**GT2:** An uncooked meat and vegetable pizza sitting on a wooden board.
**GT3:** A pizza with different toppings mushrooms, peppers, onions peperoni.
**GT4:** A pizza on a cutting board has colorful vegetables and pepperoni.
**GT5:** A pizza with multiple toppings sits on a wooden board.

**CPRC:** a person is jumping in the air with a snowboard.
**AoANet:** a person riding a snowboard on a snowboard.
**A3VSE:** a man jumping in the air on a skateboard.
**GT1:** A man flying through the air while riding a snowboard.
**GT2:** A man on a snowboard in the air above a pile of snow.
**GT3:** A person is high in the air on their skateboard.
**GT4:** Snowboarder performs airborne stunt after doing a jump.
**GT5:** A person up in the air on a snowboard.

**CPRC:** a group of horses standing in a field.
**AoANet:** two horses are standing in the snow on the snow.
**A3VSE:** two horses standing in the snow.
**GT1:** Two brown horses standing next to a white horse on a snow covered field.
**GT2:** A group of horses grazing in the snow.
**GT3:** horses eating grass through the snow in a field.
**GT4:** Three horses are grazing in the snow covered pasture.
**GT5:** Three horses grazing in the snow with trees in the background.

Fig. 7. Examples of captions generated by CPRC and baseline models as well as the corresponding ground truths.

Raw

Rotation 45 degrees

Turn around

Upside down

**CPRC (raw):** a sheep standing next to a baby sheep in a field.
**CPRC (Rotation 45 degrees):** a sheep standing next to a baby sheep.
**CPRC (Turn around):** two sheep standing in a grassy field.
**CPRC (Upside down):** a cow sheep standing next to a herd of sheep.
**GT1:** Adult sheep with young white lamb in farm setting.
**GT2:** A large lamb stands next to a baby lamb.
**GT3:** An adult sheep standing next to a baby sheep on a farm.
**GT4:** A sheep and lamb in a grassy field next to fence.
**GT5:** A lamb standing next to its mother in a field.

Raw

Rotation 45 degrees

Turn around

Upside down

**CPRC (raw):** a bowl of vegetables sitting next to a table.
**CPRC (Rotation 45 degrees):** a bowl of vegetables and vegetables on a table.
**CPRC (Turn around):** a plate of fruits and vegetables.
**CPRC (Upside down):** a bowl of fruit and bananas on a table.
**GT1:** A wooden table topped with green juice and with fruits and vegetables.
**GT2:** A smoothie is pictured next to several fruits and vegetables.
**GT3:** a brown table two apples a banana a jar and some other green vegetables.
**GT4:** There are fruits and vegetables set next to a green smoothie.
**GT5:** A cup of green drink and some vegetables.

Fig. 8. (Best viewed in color) Examples of captions generated by augmented images.

results reveal that the performance of CPRC increases first, and then decreases with the increasing of $\tau$. The reason is that fewer undescribed images are used with the increasing of $\tau$, thereby the generator training has not fully explored the information in the unsupervised data.

### J. Visualization and Analysis

Fig. 7 shows a few examples with captions generated by our CPRC and two baselines, A3VSE and AoANet, as well as the human-annotated ground truths. From these examples, we find that the generated captions of baseline models lack language logic and lose accuracy for the image objects, while CPRC can generate accurate captions in high quality.

Fig. 8 shows an example of augmented images and corresponding generated captions. We find that the generated captions basically have similar semantic information, which can help the prediction and relation consistencies for the undescribed images.

### K. Influence of Label Prediction

To explore the effect of prediction loss, we conduct more experiments and exhibit several cases. Fig. 9 shows a few examples with captions generated by our CPRC and two baselines, A3VSE and AoANet, as well as the human-annotated ground truths. From these examples, we find that the label prediction helps the generator to understand the image from these words marked in red within the sentence generated by CPRC, for example, in Fig. 9(a), the content of the image is complicated and the bird is not obvious, which causes the inconsistency of sentences generated by AoANet and A3VSE with the ground truths. But CPRC can generate a good description of "bird" and "umbrella" by combining label prediction information.

**CPRC:** two birds are flying under a red umbrella
**AoANet:** a computer that is on the top of a pole
**A3VSE:** there is a picture of a computer on a computer
**GT1:** A bird is just barely visible in the sky around a red umbrella.
**GT2:** A view of someone in the sky parachuting to the ground.
**GT3:** there is a kite that is being flown high in the sky
**GT4:** close up of a paper with type on it and a blue sky
**GT5:** A man is floating with a parachute high in the sky

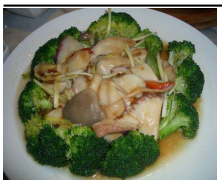
**CPRC:** a teddy bear sitting on the beach with a surfboard
**AoANet:** a group of teddy bears standing next to a beach
**A3VSE:** a group of teddy bears sitting on a beach
**GT1:** A giant sonic the hedgehog standing on a beach with a surfboard.
**GT2:** A cartoon character and North Shore surfboard are in the sands of this beach.
**GT3:** A large figure and a surf board in the sand.
**GT4:** A large Sonic statue on a beach next to a surf board.
**GT5:** A cartoon character standing on the beach by a surf boards standing in the sand.

**CPRC:** a plate of cake and apple on a plate
**AoANet:** a white plate with a piece of cake on it
**A3VSE:** a piece of cake that is on a plate
**GT1:** A desert with icing and a sliced apple beside it.
**GT2:** A piece of cake with a fork and one and a half apples on the plate.
**GT3:** An fresh apple and a half on a plate with a slice of apple bread dribbled with cream
**GT4:** A dessert with cake and apples is on a plate with a fork.
**GT5:** The cut apples give a hint about what the cake tastes like.

**CPRC:** a couple of men flying a frisbee on a beach
**AoANet:** a man is laying on the beach with a man on the beach
**A3VSE:** a man is laying on a beach with a kite
**GT1:** The men are both diving to catch the frisbee
**GT2:** Two people on the beach are flying to catch a Frisbee.
**GT3:** Two people are reaching for a toy on a beach.
**GT4:** some men are jumping around at a beach
**GT5:** Two people jump toward each other to catch a frisbee

**CPRC:** a group of people standing in a building with a clock
**AoANet:** a group of people standing next to a large building
**A3VSE:** a group of people standing outside of a building
**GT1:** A bunch of people sit in an open court yard
**GT2:** a small group of people standing around a ball patio
**GT3:** A group of people walking around a parking lot.
**GT4:** A group of people in front of a white building
**GT5:** Many people on a courtyard under a clock.',

**CPRC:** a knife is laying on top of a cutting board
**AoANet:** a pair of scissors on a wooden table
**A3VSE:** a pair of scissors sitting on a wooden table
**GT1:** A knife sticking out of the side of a block of cheese.
**GT2:** A knife is cutting into a large yam.
**GT3:** A white cutting board with a serrated steak knife cut into and resting on a vegetable.
**GT4:** A knife and cheese sit on a cutting board.
**GT5:** A block of cheese on a cutting board with a knife in it



**CPRC:** a plate of food with broccoli and broccoli
**AoANet:** a plate of vegetables in a table
**A3VSE:** a plate of pasta with vegetables
**GT1:** A filled plate of Chinese food with broccoli.
**GT2:** a broccoli dish on a plate on a table
**GT3:** Some sort of dish on a plate with pasta, mushrooms and broccoli
**GT4:** There is broccoli in the food on the plate.
**GT5:** Asian vegetable stir fry dish with wreath of broccoli and assorted mushroom varieties.

**CPRC:** a woman standing in a kitchen with a oven
**AoANet:** a woman in a woman in a kitchen
**A3VSE:** a woman standing in a kitchen with a sink and a kitchen
**GT1:** A tall woman is standing in a small kitchen.
**GT2:** A girl is standing in a kitchen with a mug in her hands..
**GT3:** Woman in knitted jump pants and yellow sleeveless top, in kitchen scene with matching yellow tone area.
**GT4:** a woman standing in a kitchen near a refrigerator and a stove
**GT5:** A woman with pigtails is standing in a kitchen.

Fig. 9. Examples of captions generated by CPRC and baseline models as well as the corresponding ground truths (GT1–GT5 are the five given ground truth sentences).

TABLE III
PERFORMANCE OF CPRC WITH DIFFERENT AUGMENTATION NUMBER ON MS-COCO KARPATHY TEST SPLIT, WHERE B@N, M, R, C, AND S ARE SHORT FOR BLEU@N, METEOR, ROUGE-L, CIDER-D, AND SPICE SCORES

| Methods | Cross Entropy Loss | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | B@1 | B@2 | B@3 | B@4 | M | R | C | S |
| K=1 | 67.5 | 48.9 | 34.6 | 22.5 | 21.1 | 48.4 | 74.7 | 15.5 |
| K=2 | 67.8 | 49.5 | 34.9 | 23.4 | 21.7 | 49.5 | 75.9 | 15.8 |
| K=3 | **68.8** | **51.1** | **35.5** | **24.9** | **22.8** | **50.4** | **77.9** | **16.2** |
| K=4 | 67.9 | 49.8 | 34.8 | 24.2 | 22.2 | 50.1 | 76.8 | 16.0 |
| K=5 | 67.6 | 49.7 | 34.5 | 23.8 | 22.0 | 49.8 | 76.2 | 16.0 |
| Methods | CIDEr-D Score Optimization | | | | | | | |
| | B@1 | B@2 | B@3 | B@4 | M | R | C | S |
| K=1 | 68.0 | 50.1 | 35.7 | 24.8 | 22.0 | 49.5 | 77.1 | 16.1 |
| K=2 | 68.3 | 50.5 | 35.9 | 25.3 | 22.1 | 49.7 | 77.7 | 16.5 |
| K=3 | **69.9** | **51.8** | **36.7** | **25.5** | **23.4** | **50.7** | **78.8** | **16.8** |
| K=4 | 68.7 | 51.4 | 36.5 | 25.2 | 22.8 | 49.7 | 77.4 | 16.3 |
| K=5 | 68.3 | 50.8 | 35.9 | 25.1 | 22.7 | 49.4 | 77.3 | 16.2 |

TABLE IV
PERFORMANCE OF CPRC WITH DIFFERENT $\tau$ ON MS-COCO KARPATHY TEST SPLIT, WHERE B@N, M, R, C, AND S ARE SHORT FOR BLEU@N, METEOR, ROUGE-L, CIDER-D, AND SPICE SCORES

| Methods | Cross Entropy Loss | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | B@1 | B@2 | B@3 | B@4 | M | R | C | S |
| $\tau = 0$ | 66.9 | 49.8 | 34.5 | 24.2 | 21.5 | 49.5 | 76.2 | 15.4 |
| $\tau = 0.1$ | **68.8** | **51.1** | **35.5** | **24.9** | **22.8** | **50.4** | **77.9** | **16.2** |
| $\tau = 0.4$ | 66.4 | 49.5 | 34.3 | 24.0 | 21.1 | 48.8 | 75.8 | 15.2 |
| $\tau = 0.7$ | 64.2 | 48.1 | 33.4 | 22.9 | 20.4 | 46.5 | 73.3 | 15.0 |
| Methods | CIDEr-D Score Optimization | | | | | | | |
| | B@1 | B@2 | B@3 | B@4 | M | R | C | S |
| $\tau = 0$ | 68.5 | 50.8 | 36.2 | 25.0 | 22.5 | 49.8 | 77.5 | 16.2 |
| $\tau = 0.1$ | **69.9** | **51.8** | **36.7** | **25.5** | **23.4** | **50.7** | **78.8** | **16.8** |
| $\tau = 0.4$ | 68.4 | 50.2 | 36.1 | 24.8 | 22.1 | 49.5 | 77.1 | 16.1 |
| $\tau = 0.7$ | 64.8 | 48.6 | 34.2 | 23.5 | 20.8 | 47.3 | 73.7 | 15.1 |

## L. Discussion

In this article, we concentrate the semisupervised image captioning task, in which the key challenge is the reasonable using of undescribed images to improve the performance of generator. Compared to current semisupervised approaches that only consider the pseudo descriptions, we design a novel captioning model by constraining the undescribed images from both the prediction and relation consistencies. In experiments, we comprehensively compare with state-of-the-art captioning methods to validate the effectiveness of proposed CPRC on popular datasets. Furthermore, we conduct adequate ablation studies,

which verify that: 1) the designed prediction and relation consistencies can positively promote the use of undescribed images; 2) CPRC can be effectively applied to any supervised models; 3) CPRC is highly interpretable for parameter selection; and 4) CPRC has a good visualization effect, and the significance of each module is reflected.

## V. Conclusion

Since traditional image captioning methods usually work on supervised multimodal data, in this article, we investigated how to use undescribed images for semisupervised image captioning. Specifically, our method can take CPRC into consideration. CPRC employs prediction distillation for the predictions of sentences generated from undescribed images, and develops a novel relation consistency between augmented images and generated sentences to retain the important relational knowledge. As demonstrated by the experiments, CPRC outperforms state-of-the-art methods in various complex semisupervised scenarios.

## References


[1] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[2] E. S. Debie *et al.*, "Multimodal fusion for objective assessment of cognitive workload: A review," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1542–1555, Mar. 2021.

[3] A. Karpathy and F. Li, "Deep visual-semantic alignments for generating image descriptions," in *Proc. CVPR*, 2015, pp. 3128–3137.

[4] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, 2015, pp. 2048–2057.

[5] F. Sammani and M. Elsayed, "Look and modify: Modification networks for image captioning," in *Proc. BMVC*, 2019, p. 75.

[6] Y. Bin, Y. Yang, F. Shen, N. Xie, H. T. Shen, and X. Li, "Describing video with attention-based bidirectional LSTM," *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2631–2641, Jul. 2019.

[7] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. Salakhutdinov, "Review networks for caption generation," in *Proc. NeurIPS*, 2016, pp. 2361–2369.

[8] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. CVPR*, 2017, pp. 3242–3250.

[9] L. Huang, W. Wang, J. Chen, and X. Wei, "Attention on attention for image captioning," in *Proc. ICCV*, 2019, pp. 4633–4642.

[10] T. B. Hashimoto, K. Guu, Y. Oren, and P. Liang, "A retrieve-and-edit framework for predicting structured outputs," in *Proc. NeurIPS*, 2018, pp. 10073–10083.

[11] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised image captioning," in *Proc. CVPR*, Long Beach, CA, USA, 2019, pp. 4125–4134.

[12] J. Gu, S. R. Joty, J. Cai, H. Zhao, X. Yang, and G. Wang, "Unpaired image captioning via scene graph alignments," in *Proc. ICCV*, 2019, pp. 10322–10331.

[13] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NeurIPS*, Montreal, QC, Canada, 2014, pp. 2672–2680.

[14] N. C. Mithun, R. Panda, E. E. Papalexakis, and A. K. Roy-Chowdhury, "Webly supervised joint embedding for cross-modal image-text retrieval," in *Proc. ACMMM*, 2018, pp. 1856–1864.

[15] P. Huang, G. Kang, W. Liu, X. Chang, and A. G. Hauptmann, "Annotation efficient cross-modal retrieval with adversarial attentive alignment," in *Proc. ACMMM*, 2019, pp. 1758–1767.

[16] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. CVPR*, 2017, pp. 1179–1195.

[17] Y. Zhou, M. Wang, D. Liu, Z. Hu, and H. Zhang, "More grounded image captioning by distilling image-text matching model," in *Proc. CVPR*, 2020, pp. 4776–4785.

[18] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S. C. Zhu, "I2T: Image parsing to text description," *Proc. IEEE*, vol. 98, no. 8, pp. 1485–1508, Aug. 2010.

[19] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. EMNLP*, 2014, pp. 1724–1734.

[20] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. CVPR*, 2015, pp. 3156–3164.

[21] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. NeurIPS*, 2004, pp. 529–536.

[22] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *Proc. IJCNN*, 2020, pp. 1–8.

[23] P. Bachman, O. Alsharif, and D. Precup, "Learning with pseudo-ensembles," in *Proc. NeurIPS*, 2014, pp. 3365–3373.

[24] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. NeurIPS*, Long Beach, CA, USA, 2017, pp. 1195–1204.

[25] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Proc. ICLR*, Toulon, France, 2017, pp. 1–13.

[26] Q. Xie, Z. Dai, E. H. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," in *Proc. NeurIPS*, 2020, pp. 1–13.

[27] D. Berthelot *et al.*, "ReMixMatch: Semi-supervised learning with distribution matching and augmentation anchoring," in *Proc. ICLR*, 2020, pp. 1–13.

[28] G. French, M. Mackiewicz, and M. H. Fisher, "Self-ensembling for visual domain adaptation," in *Proc. ICLR*, 2018, pp. 1–20.

[29] K. Sohn *et al.*, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," 2020, *arXiv:2001.07685*.

[30] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. CVPR*, 2015, pp. 4566–4575.

[31] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," in *Proc. ICLR*, 2016, pp. 1–16.

[32] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation policies from data," 2018, *arXiv:1805.09501*.

[33] Y. Lin, C. Wang, C. Chang, and H. Sun, "An efficient framework for counting pedestrians crossing a line using low-cost devices: The benefits of distilling the knowledge in a neural network," *Multim. Tools Appl.*, vol. 80, no. 3, pp. 4037–4051, 2021.

[34] P. Matthews, *A Short History of Structural Linguistics*. Cambridge, U.K.: Cambridge Univ. Press, 2001.

[35] T. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.

[36] L. Huang, W. Wang, Y. Xia, and J. Chen, "Adaptively aligned image captioning via adaptive attention time," in *Proc. NeurIPS*, 2019, pp. 8940–8949.

[37] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in *Proc. NeurIPS*, 2019, pp. 11135–11145.

[38] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 664–676, Apr. 2017.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.

[40] G. Xu, S. Niu, M. Tan, Y. Luo, Q. Du, and Q. Wu, "Towards accurate text-based image captioning with content diversity exploration," in *Proc. CVPR*, 2021, pp. 12637–12646.

[41] X. Zhang *et al.*, "RSTNet: Captioning with adaptive attention on visual and non-visual words," in *Proc. CVPR*, 2021, pp. 15465–15474.

[42] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. ACL*, 2002, pp. 311–318.

[43] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. IEEMMT*, 2005, pp. 65–72.

[44] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in *Proc. ECCV*, 2016, pp. 382–398.

[45] J. Giménez and L. Màrquez, "Linguistic features for automatic evaluation of heterogenous MT systems," in *Proc. ACL Workshop*, 2007, pp. 256–264.

**Yang Yang** received the Ph.D. degree in computer science from Nanjing University, Nanjing, China, in 2019.

He is currently a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing. He has published over 20 papers in leading international journal/conferences. His research interests lie primarily in machine learning and data mining, including heterogeneous learning, model reuse, and incremental mining.

**Hongchen Wei** is currently pursuing the M.Sc. degree with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China.

His research interests lie primarily in machine learning and data mining, including cross-modal learning.

**Hengshu Zhu** (Senior Member, IEEE) received the B.E. and Ph.D. degrees in computer science from the University of Science and Technology of China, Hefei, China, in 2009 and 2014, respectively.

He is currently a Principal Data Scientist and an Architect with Baidu Inc., Beijing, China. He has published prolifically in refereed journals and conference proceedings, including the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON MOBILE COMPUTING, ACM TRANSACTIONS ON INFORMATION SYSTEMS, ACM TRANSACTIONS ON KNOWLEDGE DISCOVERY FROM DATA, ACM SIGKDD, ACM SIGIR, WWW, IJCAI, and AAAI. His general area of research is data mining and machine learning, with a focus on developing advanced data analysis techniques for innovative business applications.

Dr. Zhu was the recipient of the Distinguished Dissertation Award of CAS in 2016, the Distinguished Dissertation Award of CAAI in 2016, the Special Prize of President Scholarship for Postgraduate Students of CAS in 2014, the Best Student Paper Award of KSEM-2011, WAIM-2013, and CCDM-2014, and the Best Paper Nomination of ICDM-2014. He has served regularly on the organization and program committees of numerous conferences, including as a Program Co-Chair for the KDD Cup-2019 Regular ML Track, and a Founding Co-Chair for the first International Workshop on Organizational Behavior and Talent Analytics and the International Workshop on Talent and Management Computing, in conjunction with ACM SIGKDD. He is the Senior Member of ACM and CCF.

**Dianhai Yu** received a bachelor's degree from Jilin University, Changchun, China, in 2005, and the master's degree from Peking University, Beijing, China, in 2008.

He is currently a Chief Machine Learning Architect with Baidu Inc., Beijing, China, and is in charge of the opensource deep learning platform PaddlePaddle. He joined Baidu in 2008 after graduating from Peking University. He has published over ten papers in the field of Artificial Intelligence. His research interests lie primarily in machine learning and natural language processing, including machine learning systems, scalable distributed deep learning, semantic computing, and human–machine dialogue systems.

Dr. Yu received the CCF Outstanding Engineer Award in 2019. He is the Senior Member of CCF.

**Hui Xiong** (Fellow, IEEE) received the Ph.D. degree from the University of Minnesota, Minneapolis, MN, USA, in 2005.

He is currently a Full Professor with the Rutgers University, Newark, NJ, USA.

Prof. Xiong received from the Rutgers University, the 2018 Ram Charan Management Practice Award as the Grand Prix winner from the Harvard Business Review, the RBS Dean's Research Professorship in 2016, the Rutgers University Board of Trustees Research Fellowship for Scholarly Excellence in 2009, the ICDM Best Research Paper Award in 2011, and the IEEE ICDM Outstanding Service Award in 2017. He is a Co-Editor-in-Chief of the *Encyclopedia of GIS*, an Associate Editor of IEEE TRANSACTIONS ON BIG DATA, ACM TRANSACTIONS ON KNOWLEDGE DISCOVERY FROM DATA, and ACM TRANSACTIONS ON MANAGEMENT INFORMATION SYSTEMS. He has served regularly on the organization and program committees of numerous conferences, including as a Program Co-Chair for the Industrial and Government Track for the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, a Program Co-Chair for the IEEE 2013 International Conference on Data Mining (ICDM), a General Co-Chair for the IEEE 2015 ICDM, and a Program Co-Chair for the Research Track for the 2018 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. He is an ACM Distinguished Scientist.

**Jian Yang** (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology (NUST), Nanjing, China, in 2002.

In 2003, he was a Postdoctoral Researcher with the University of Zaragoza, Zaragoza, Spain. From 2004 to 2006, he was a Postdoctoral Fellow with the Biometrics Centre, The Hong Kong Polytechnic University, Hong Kong. From 2006 to 2007, he was a Postdoctoral Fellow with the Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA. He is currently a Chang-Jiang Professor with the School of Computer Science and Engineering, NUST. He has authored more than 200 scientific papers in pattern recognition and computer vision. His papers have been cited more than 6000 times in the Web of Science and 15 000 times in the Scholar Google. His current research interests include pattern recognition, computer vision, and machine learning.

Dr. Yang is currently an Associate Editor of *Pattern Recognition*, *Pattern Recognition Letters*, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and *Neurocomputing*. He is a Fellow of IAPR.