IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 May 2024, accepted 25 June 2024, date of publication 5 July 2024, date of current version 15 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3423790

RESEARCH ARTICLE

Semi-Supervised Image Captioning by Adversarially Propagating Labeled Data

DONG-JIN KIM[1], (Member, IEEE), TAE-HYUN OH[2,3,4], (Member, IEEE), JINSOO CHOI[5], AND IN SO KWEON[5], (Member, IEEE)

[1]Department of Data Science, Hanyang University, Seoul 04763, South Korea
[2]Department of Electrical Engineering, POSTECH, Pohang 37673, South Korea
[3]Graduate School of AI, POSTECH, Pohang 37673, South Korea
[4]Institute for Convergence Research and Education in Advanced Technology, Yonsei University, Seoul 03722, Republic of Korea
[5]Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea

Corresponding author: Dong-Jin Kim (djdkim@hanyang.ac.kr)

This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) Grant funded by the Korean Government [Ministry of Science and ICT (MSIT)] through the Artificial Intelligence Graduate School Program, Hanyang University, under Grant RS-2020-II201373; and in part by the National Research Foundation of Korea (NRF) Grant funded by Korea Government (MSIT) under Grant RS-2023-00245661. The work of Tae-Hyun Oh was supported in part by IITP Grant funded by Korea Government (MSIT) (Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities) under Grant RS-2022-II220124.

ABSTRACT We present a novel data-efficient *semi-supervised* framework to improve the generalization of image captioning models. Constructing a large-scale labeled image captioning dataset is expensive in terms of labor, time, and cost. In contrast to manually annotating all the training samples, separately collecting uni-modal datasets is immensely easier, *e.g.*, a large-scale image dataset and a sentence dataset. We leverage such massive *unpaired* image and caption data upon standard paired data by learning to associate them. To this end, our novel semi-supervised learning method assigns pseudo-labels to unpaired images and captions in an adversarial learning fashion, where the joint distribution of image and caption is learned. This approach shows noticeable performance improvement even in challenging scenarios, including out-of-task data and web-crawled data. We also show that our proposed method is theoretically well-motivated and has a favorable global optimal property. Our extensive and comprehensive empirical results on captioning datasets, followed by a comprehensive analysis of the scarcely-paired COCO dataset, demonstrate the consistent effectiveness of our method compared to competing ones.

INDEX TERMS Image captioning, unpaired captioning, semi-supervised learning, generative adversarial networks.

I. INTRODUCTION

Image captioning is to generate a natural language description of a given image. It is highly useful for image understanding in that 1) it extracts the essence of an image into a self-descriptive form, and 2) the output format is an interpretable natural language, which is free-form and easy to manipulate so that it can be beneficial to user interactable applications such as language-based image retrieval [1], video summarization [2], navigation [3], and vehicle control [4]. Image captioning is not limited to a few pre-defined classes, allowing descriptive analysis of general images.

Recent works on image captioning have shown remarkable progress [5], [6], [7]. However, most works are trained only with supervised learning where it would be hard to transfer a model to a target domain with significant domain shift [8]. One way to improve the image captioning model's generalizability would be to add more supervised data, which is difficult in practice. Specifically, the MS COCO caption dataset is constructed with 120,000 number of images that

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenbao Liu.

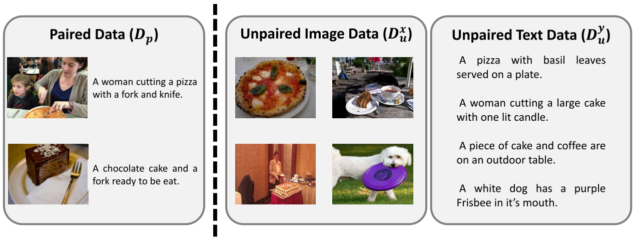

Paired Data ($\mathcal{D}_p$)
A woman cutting a pizza with a fork and knife.
A chocolate cake and a fork ready to be eat.
Unpaired Image Data ($\mathcal{D}_u^x$)
Unpaired Text Data ($\mathcal{D}_u^y$)
A pizza with basil leaves served on a plate.
A woman cutting a large cake with one lit candle.
A piece of cake and coffee are on an outdoor table.
A white dog has a purple Frisbee in it's mouth.

FIGURE 1. **We utilize "unpaired" image-caption data upon "paired" data. We denote paired data as $\mathcal{D}_p$ and unpaired image and caption datasets as $\mathcal{D}_u^x$ and $\mathcal{D}_u^y$ respectively.**

were asked annotators to provide five plausible sentences for each image, which is an expensive task in terms of labor, time, and cost. Moreover, if the target task is a higher-level task involving multiple captions and bounding boxes per image, annotating the dataset becomes even more challenging. For example, for the relational captioning task [9], dense and combinatorially associated captions and a pair of two bounding boxes are used as a label, and the data for this task has much higher complexity than that of the standard image captioning task. Constructing such human-labeled datasets is an immensely laborious and time-consuming task, so building new datasets according to the different needs of target themes or application scenarios would be impractical. Therefore, our goal is to more data-efficiently improve image captioning performance.

In this paper, we introduce a novel method to utilize *unpaired* image and caption data from the web upon traditional elaborately labeled paired data to effectively improve image captioning neural networks. The motivation of our method is that images can be easily obtained from the web, and captions can be easily augmented and synthesized by replacing or adding other words given a base sentence according to parts of speech as done in [10]. Also, once a sufficient number of captions are given, we can easily crawl *corresponding but noisy* images through Google or Flickr image databases [11] to build a large image corpus. Thereby, it is easy to construct large-scale *unpaired* image and caption datasets, which require no or minimal human effort.

As the input image and output caption datasets are unpaired in our problem, the conventional supervised learning approaches can no longer be directly used. To make unpaired data paired, we propose to assign pseudo-labels to the unpaired samples. The pseudo-label is used as a *learned* supervision label. To develop the mechanism of pseudo-labeling, we are motivated and leverage the generative capability of generative adversarial networks (GAN) [12], for searching pseudo-labels from unpaired data. In other words, we propose to utilize the discriminator model trained with an adversarial training method to find the pseudo-labels for unpaired samples. As the discriminator is trained to distinguish between real and fake image-caption *pairs*, we can exploit the discriminator to retrieve pseudo-labels and improve the captioner training. Our assumption is that if the decision boundary of the discriminator is tight enough, we can use the discriminator to retrieve a proper pseudo-label when the unpaired datasets are sufficiently large.

This work is the extension of Kim et al. [13]. In this work, we further improve our method with a simple yet significantly effective concept transfer technique and analyze our framework by extensively evaluating our method in diverse and challenging scenarios: more challenging image captioning baselines [6], [14], additional caption domain of MS COCO and Flickr [15]. In addition to empirical results, we also show the theoretical justification of our design of the proposed learning method with respect to a global optimum.

In short, our main contributions are summarized as follows. (1) We propose a novel image captioning framework to train the model with the unpaired image and caption data upon traditional paired data. (2) To better exploit the knowledge in the unpaired data, we propose a novel semi-supervised learning approach by utilizing the GAN discriminator. In particular, for the scenarios when the number of paired data is scarce, we additionally propose a simple yet effective teacher-student based concept transfer method to leverage an external high-level knowledge to help bridge between unpaired image and caption data in different domains from the paired data. (3) Beyond the naïve image-level captioning task, we extend our method to the relational captioning task in order to demonstrate that our framework can be easily applied to region-based captioning datasets as well with a simple modification. (4) We link our practical realization of the proposed learning method and theoretical algorithmic behaviors. (5) We show the effectiveness of our method through extensive experiments compared with recent methods in comprehensive experimental setups. Our model trained by our learning method with 1% of paired data plausibly performs well in a qualitative sense.

II. RELATED WORK

The main goal of our work is to address unpaired image-caption data to improve the generalizability of image captioning. Therefore, we introduce image captioning works and the works on how to handle unpaired data.

A. GENERALIZABILITY IN IMAGE CAPTIONING

Since the introduction of the MS COCO dataset [16], image captioning has been extensively studied in computer vision and language community [5], [6], [7], [14], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27] by virtue of the advancement of deep neural networks [28]. As neural network architectures become more advanced, *e.g.*, Transformer [6], [21], [22], [29], they require a much larger dataset scale for generalization [30] as the image captioning models tend to show limited generalizability.

Traditionally, utilizing unpaired image-caption data requires additional information to associate images and captions. Gu et al. [31] introduce additional modal information, Chinese captions, and use it as a strong pivot language for language pivoting [32]. Feng et al. [33] propose an unpaired captioning framework that trains a model without either image or sentence labels via learning a visual concept detector with external data, the OpenImage dataset [34]. Laina et al. [35] and Guo et al. [36] propose improved training methods given the same visual concept detector as Feng et al. trained with the OpenImage dataset. We later show that our method can be easily extended with a similar visual concept learning to enhance the performance. Gu et al. [37] utilize scene graph to bridge between unpaired image and caption data. Chen et al. [8] approach image captioning as a domain adaptation by utilizing the large-scale paired MS COCO caption dataset as the source domain and adapting on separate unpaired image or caption datasets as the target domain. Zhu et al. [38], [39] utilize a pre-trained large-scale vision and language foundation model, CLIP [40], as a bridge between image and caption data. Kim et al. [41] propose a multi-task learning method to use an action recognition dataset without caption labels to improve video captioning performance. Liu et al. [42] use self-retrieval rewards for captioning to facilitate training a model with partially labeled data, where the self-retrieval module retrieves corresponding images with the captions generated from the model. As a separate line of work, there are novel object captioning methods [43], [44], [45] that additionally exploit unpaired image and caption data to mine a description of a novel word.

Most of aforementioned works including [31], [33], [35], [36], [37], [41], [43], [46], [47], and [48] exploit large auxiliary *supervised* datasets such as class labels or scene graph. To the best of our knowledge, we are the first to bridge between unpaired images and caption data in the image captioning task without any auxiliary information but by leveraging semi-supervised image-caption data only. Although Chen et al. [8] do not use auxiliary information as well, it requires large-scale paired source data, of which the data regime is different from ours. Liu et al. [42] is also this case, where they use the fully paired MS COCO caption dataset with an additional large unlabeled image set. Our method can deal with those regimes as well as paired data where the scale can be a minimum of 1% of the COCO dataset.

B. MULTI-MODALITY IN UNPAIRED DATA HANDLING

With the advance on generative modeling, *e.g.*, GAN [12], multi-modal translation recently emerged as a popular field. Among various modalities, image-to-image translation between two different and unpaired domains has been actively explored. To tackle this problem, the cycle-consistency constraint between unpaired data is exploited in CycleGAN [49] and DiscoGAN [50], and it is further improved in UNIT [51].

In this paper, image captioning can be considered as a multi-modal translation. Our work has a similar motivation to the unpaired image-to-image translation [49], unsupervised machine translation [52], and machine translation with monolingual data [53]. Interestingly, we show that the cycle-consistency does not work on our problem setup due to a significant modality gap. Instead, our results suggest that the traditional label propagation based semi-supervised framework [54] is more effective for our task.

C. SEMI-SUPERVISED LEARNING

In general, the goal of semi-supervised learning (SSL) is to improve the model performance by training with unlabeled data under a transductive assumption [55]. Recent deep learning based SSL methods can be divided into four main categories: (1) pseudo-label generation [56], (2) consistency regularization [57], [58], (3) combination of pseudo-labeling with consistency regularization [59], [60], [61], [62], and (4) generative model based methods [63], [64]. Our method is motivated by the generative model based semi-supervised learning [63]. While the prior work is mostly limited to dealing with simple image classification, our work extends the regime to image and caption modalities.

III. PROPOSED METHOD

In this section, we first introduce the image caption learning pipeline and describe how to leverage the unpaired dataset. Then, we describe our adversarial learning method using a GAN model that encourages matching the distribution of latent features of images and captions. The GAN model is used for assigning pseudo-labels, which allows challenging semi-supervised learning with both labeled and unlabeled data. Moreover, we analyze the theoretical properties of our proposed framework. Lastly, we extend our method to the relational captioning scenario.

A. ADVERSARIAL SEMI-SUPERVISED TRAINING

Let us denote a dataset with N_p image-caption pairs as $\mathcal{D}_p = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_p}$. A typical image captioning task is defined as follows: given an image x_i, the model generates a caption y_i that best describes the image. Traditionally, a captioning model is trained on a large paired dataset $(\mathbf{x}, y) \in \mathcal{D}_p$, *e.g.*, the MS COCO dataset, by minimizing the negative log likelihood against the ground truth caption as follows:

$$\sum_{(\mathbf{x},y)\in\mathcal{D}_p} L_{\mathsf{CE}}(y, \hat{y}(\mathbf{x})), \tag{1}$$

where L_{CE} denotes the cross-entropy loss, and $\hat{y}(\mathbf{x})$ denotes output of the model. Motivated by early neural machine translation literature [65], captioning frameworks have been typically implemented as an encoder-decoder architecture [7], *i.e.* CNN-RNN. The CNN encoder $F(\mathbf{x})$ outputs a latent feature vector $\mathbf{z}^x$ from a given input image $\mathbf{x}$, followed by the Language decoder $H(\mathbf{z}^x)$ (*e.g.*, RNN or Transformers), as depicted in Fig. 2, to generate a caption y from $\mathbf{z}^x$ in a natural language form, *i.e.* $\hat{y}(\mathbf{x}) = p(y|\mathbf{x}; F, H) = H \circ F(\mathbf{x})$.

1) LEARNING WITH UNPAIRED DATA

Our problem deals with unpaired data, where the image and caption sets $\mathcal{D}_u^x = \{\mathbf{x}_i\}_{i=0}^{N_x}$ and $\mathcal{D}_u^y = \{y_i\}_{i=0}^{N_y}$ are not paired. Given the unpaired datasets, due to missing annotations, the

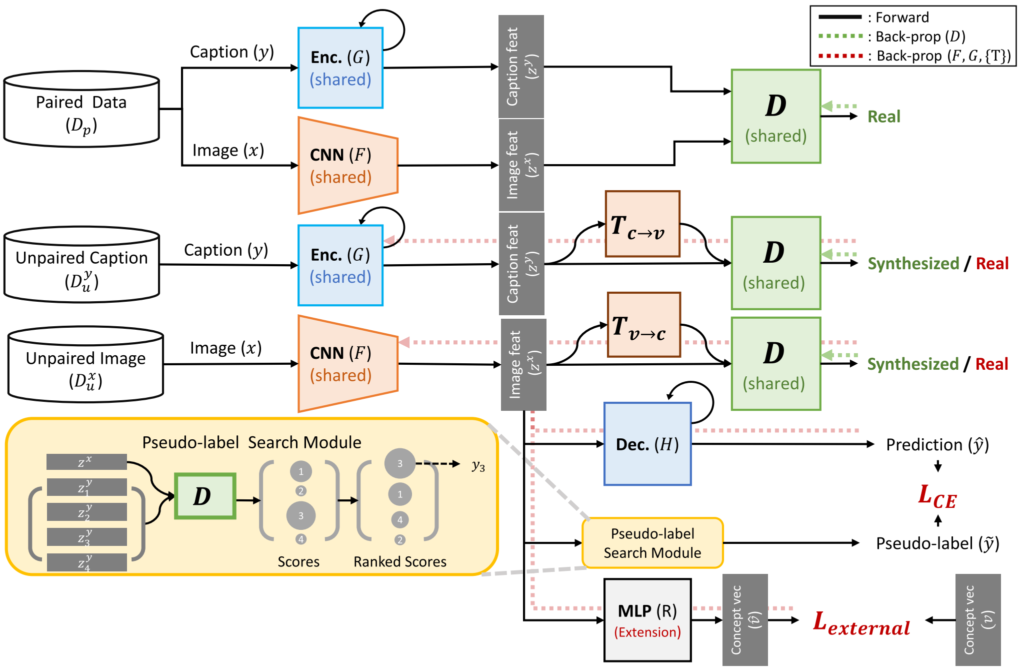


FIGURE 2. Illustration of the proposed method. Dotted arrows denote the path of the gradients via back-propagation. Given any image and caption pair, CNN and LSTM/Transformer encoders encode input image and caption into the respective feature spaces. A discriminator (*D*) is trained to discriminate whether the given feature pairs are real or fake, while the encoders are trained to fool the discriminator. The learned discriminator is also used to assign the most likely pseudo-labels to unpaired samples through the pseudo-label search module. We additionally introduce an auxiliary multi-layer perceptron to learn external knowledge via concept transfer.

loss in Eq. (1) cannot be directly computed. Motivated by the semi-supervised framework [66], we artificially generate *pseudo-labels* for respective unpaired datasets so that the supervision loss in Eq. (1) can be leveraged with unpaired data.

Specifically, we retrieve the best matched caption $\tilde{y}_i$ in $\mathcal{D}_u^y$ given a query image $\mathbf{x}_i$, assign it as a pseudo-label, and vice versa ($\tilde{\mathbf{x}}_i$ for y_i). We express the pseudo-labeling as a function for simplicity, *i.e.* $\tilde{y}_i = \tilde{y}(\mathbf{x}_i)$. To retrieve a semantically meaningful match, we need a measure to assess proper matches. We use a discriminator network to determine real or fake pairs in a similar way to GANs, which will be described in later sections. With the retrieved pseudo-labels, we can now compute Eq. (1) with unpaired data as:

$$\min_{F,H} \lambda_x \sum_{\mathbf{x}\in\mathcal{D}_u^x} L_{\text{CE}}(\tilde{y}(\mathbf{x}), \hat{y}(\mathbf{x})) + \lambda_y \sum_{y\in\mathcal{D}_u^y} L_{\text{CE}}(y, \hat{y}(\tilde{\mathbf{x}}(y))), \quad (2)$$

where $\lambda_{\{\cdot\}}$ denote the balance parameters.

2) DISCRIMINATOR LEARNING BY UNPAIRED FEATURE MATCHING

We train via a criterion to find a semantically meaningful match so that pseudo-labels for each modality are effectively retrieved. To this end, we train a discriminator and then use the discriminator for pseudo-labeling to train the image captioning model on both paired and unpaired datasets.

We introduce a caption encoder, $G(y)$, which embeds the caption y into a feature $\mathbf{z}^y$. This can be implemented with an LSTM or Transformer [67], and we take the output of the last time step as the caption representation $\mathbf{z}^y$. Likewise, given an image $\mathbf{x}$, we obtain $\mathbf{z}^x$ by the image encoder $F(\mathbf{x})$. Now, we have a comparable feature space of $\mathbf{z}^x$ and $\mathbf{z}^y$, of which the number of dimensions are set to be the same. We train the discriminator to distinguish whether the pair $(\mathbf{z}^x, \mathbf{z}^y)$ comes from true paired data $(\mathbf{x}, y) \in \mathcal{D}_p$, *i.e.* the pair belongs to the real distribution $p(\mathbf{x}, y)$ or not.

To train the discriminator, we could use random data of $\mathbf{x}$ and y independently sampled from respective unpaired datasets, but we found that it is detrimental to performance due to uninformative pairs in training. Instead, we conditionally synthesize $\mathbf{z}^x$ or $\mathbf{z}^y$, to form a synthesized pair that appears to be as realistic as possible. We use the feature transformer networks $\tilde{\mathbf{z}}^y = T_{v\to c}(\mathbf{z}^x)$ and $\tilde{\mathbf{z}}^x = T_{c\to v}(\mathbf{z}^y)$, where $v\to c$ denotes the mapping from visual data to caption data and vice versa, and $\tilde{\mathbf{z}}^{(\cdot)}$ denotes the conditionally synthesized feature. $\{T\}$ are implemented with a multi-layer-perceptron with four fully-connected (FC) layers with the ReLU nonlinearity.

The discriminator $D(\cdot, \cdot)$ learns to distinguish features if they are real or not. At the same time, the other associated networks F, G, $T_{\{\cdot\}}$ are learned to fool the discriminator by matching the distribution of paired and unpaired data. We formulate this adversarial training as follows:

$$\begin{aligned}&\min_{F,G,\{T\}} \max_{D} \tilde{U}(F, G, \{T\}, D)\\ &= \min_{F,G,\{T\}} \max_{D} U(F, G, \{T\}, D) + \mathop{\mathbb{E}}_{\substack{(\mathbf{z}^x,\mathbf{z}^y)\\ \sim(F,G)\circ\mathcal{D}_p}} [L_{reg}(\mathbf{z}^x, \mathbf{z}^y, \{T\})],\end{aligned} \quad (3)$$

where

$$
\begin{aligned}
&U(F, G, \{T\}, D) \\
&= \underset{\substack{(\mathbf{z}^x,\mathbf{z}^y)\\ \sim(F,G)\circ\mathcal{D}_p}}{\mathbb{E}} [\log(D(\mathbf{z}^x, \mathbf{z}^y))] \\
&\quad + \frac{1}{2}\Bigg(\underset{\mathbf{x}\sim p(\mathbf{x})}{\mathbb{E}} [\log(1 - D(F(\mathbf{x}), T_{v\to c}(F(\mathbf{x}))))] \\
&\qquad + \underset{y\sim p(y)}{\mathbb{E}} [\log(1 - D(T_{c\to v}(G(y)), G(y)))]\Bigg), \qquad (4)
\end{aligned}
$$

$L_{reg}(\mathbf{z}^x, \mathbf{z}^y, \{T\}) = \lambda_{reg}(\| \underset{v\to c}{T}(\mathbf{z}^x) - \mathbf{z}^y\|_F^2 + \|\mathbf{z}^x - \underset{c\to v}{T}(\mathbf{z}^y)\|_F^2)$. Note that the first log term in Eq. (4) is not used for updating any learnable parameters related to $F, G, \{T\}$, but only used for updating D. The overall architecture related to this formulation is illustrated in Fig. 2.

Through alternating training of the discriminator (D) and generators ($F, G, \{T\}$), the latent feature distribution of paired and unpaired data should be close to each other, *i.e.*, $p(\mathbf{z}^x, \mathbf{z}^y) \approx p_{v\to c}(\mathbf{z}^x, \mathbf{z}^y) \approx p_{c\to v}(\mathbf{z}^x, \mathbf{z}^y)$, where $p_{v\to c}(\mathbf{z}^x, \mathbf{z}^y) = p(\mathbf{z}^x)p_{v\to c}(\mathbf{z}^y|\mathbf{z}^x)$, $p_{c\to v}(\mathbf{z}^x, \mathbf{z}^y) = p(\mathbf{z}^y)p_{c\to v}(\mathbf{z}^x|\mathbf{z}^y)$, and $p_{v\to c}(\mathbf{z}^y|\mathbf{z}^x)$ and $p_{c\to v}(\mathbf{z}^x|\mathbf{z}^y)$ are modeled with $T_{v\to c}$ and $T_{c\to v}$, respectively. It implies that, as the generator is trained, the decision boundary of the discriminator tightens; hence, we can use the D to retrieve a proper pseudo-label if the unpaired datasets are sufficiently large such that semantically meaningful matches exist between the different modality datasets.

3) PSEUDO-LABELING

Given an image $\mathbf{x} \in \mathcal{D}_u^x$, we retrieve a caption in the unpaired dataset, $\tilde{y} \in \mathcal{D}_u^y$, with the highest score obtained by the discriminator, *i.e.* the most likely caption to be paired with the given image as

$$\tilde{y}_i = \tilde{y}(\mathbf{x}_i) = \underset{y\in\mathcal{D}_u^y}{\operatorname{argmax}}\ D\left(F(\mathbf{x}_i), G(y)\right), \qquad (5)$$

vice versa for unpaired captions:

$$\tilde{\mathbf{x}}_i = \tilde{\mathbf{x}}(y_i) = \underset{\mathbf{x}\in\mathcal{D}_u^x}{\operatorname{argmax}}\ D\left(F(\mathbf{x}), G(y_i)\right). \qquad (6)$$

By this retrieval process over all the unpaired datasets, we now have fully paired data; *i.e.* image-caption pairs $\{(\mathbf{x}_i, y_i)\}$ from the paired data and the pairs with pseudo-labels $\{(\mathbf{x}_j, \tilde{y}_j)\}$ and $\{(\tilde{\mathbf{x}}_k, y_k)\}$ from the unpaired data. However, these pseudo-labels are likely to be noisy or biased, thus treating them equally with the paired ones would not be desirable [68], [69]. Motivated by learning with noisy labels [70], [71], we re-weigh the data pairs by defining a confidence score for each of the assigned pseudo-labels. In order to obtain the confidence score, we propose to use the output score from the discriminator as the confidence score, *i.e.* $\alpha_i^x = \hat{D}(\mathbf{x}_i, \tilde{y}_i)$ and $\alpha_i^y = \hat{D}(\tilde{\mathbf{x}}_i, y_i)$, where we denote $\hat{D}(\mathbf{x}, y) = D(F(\mathbf{x}), G(y))$, and $\alpha \in [0, 1]$ due to the sigmoid function at the final layer. We utilize the confidence scores to assign weights to the unpaired samples. The final weighted loss $\min_{F,H} \mathcal{L}_{cap}(F, H)$ is defined as follows:

$$
\begin{aligned}
\min_{F,H} &\sum_{(\mathbf{x},y)\in\mathcal{D}_p} L_{\mathsf{CE}}(y, \hat{y}(\mathbf{x})) + \lambda_x \sum_{\mathbf{x}\in\mathcal{D}_u^x} \alpha^x_{(\tilde{y}(\mathbf{x}),\mathbf{x})} L_{\mathsf{CE}}(\tilde{y}(\mathbf{x}), \hat{y}(\mathbf{x})) \\
&+ \lambda_y \sum_{y\in\mathcal{D}_u^y} \alpha^y_{(y,\tilde{\mathbf{x}}(y))} L_{\mathsf{CE}}(y, \hat{y}(\tilde{\mathbf{x}}(y))). \qquad (7)
\end{aligned}
$$

4) LEVERAGING EXTERNAL KNOWLEDGE VIA CONCEPT TRANSFER

Although our semi-supervised learning method works properly to associate unpaired image and caption data despite scarce paired data, the smaller the paired data size is, the more difficult it becomes to associate unpaired samples from *unseen* domains. This is because the small paired data lacks the information to capture any snippet of image or text, *i.e.* concept of each data. Therefore, as an extension, we propose to borrow a pre-trained knowledge to effectively associate unpaired samples by capturing concepts regardless of domain, which is crucial for semi-supervised learning, especially when paired data is scarce.

As an external source of knowledge, we propose to use concept embeddings obtained from an off-the-shelf and pre-trained scene understanding model that provides a high-level scene understanding. We extract a set of dense vectors[1] from an image by using the pre-trained model. By averaging the vectors of the image, we obtain a single vector $\mathbf{v} = \text{Concept}(\mathbf{x})$ that represents an image, which we call "concept vector."

To borrow knowledge from an external pre-trained model *regardless of its network architecture*, we utilize this concept vector in a way of the knowledge distillation [72], where our image encoder $F(\cdot)$ learns the knowledge encoded in the vector. To make the encoder deal with this auxiliary task, we add an auxiliary concept regression branch $R(\cdot)$ to the penultimate layer. The auxiliary branch is implemented by a multi-layer perceptron to create a vector $\hat{\mathbf{v}} = R \circ F(\mathbf{x})$ that mimics the concept vector provided by the high-level scene understanding model. Then, the image captioning model is trained by adding the additional concept regression loss $\mathcal{L}_{external}$ as follows:

$$\mathcal{L}_{external}(F) = \underset{\mathbf{x}\sim p(\mathbf{x})}{\mathbb{E}} \|R \circ F(\mathbf{x}) - \text{Concept}(\mathbf{x})\|_F^2, \qquad (8)$$

which is described in Fig. 2. Thereby, the knowledge from the external model could be effectively transferred to the image captioning model. This simple approach significantly improves the performance of an image captioning model when the number of paired data is scarce, as shown in Sections IV-C and IV-E.

To produce the concept vector, we use the pre-trained relational captioning model [73]. We generate abundant relational caption proposals from an image by using the model, and each caption is mapped to an embedding by

[1] The dense vectors can be any dense representation, *e.g.*, a pixel-wise feature map, feature vectors corresponding to region proposals, *etc.*

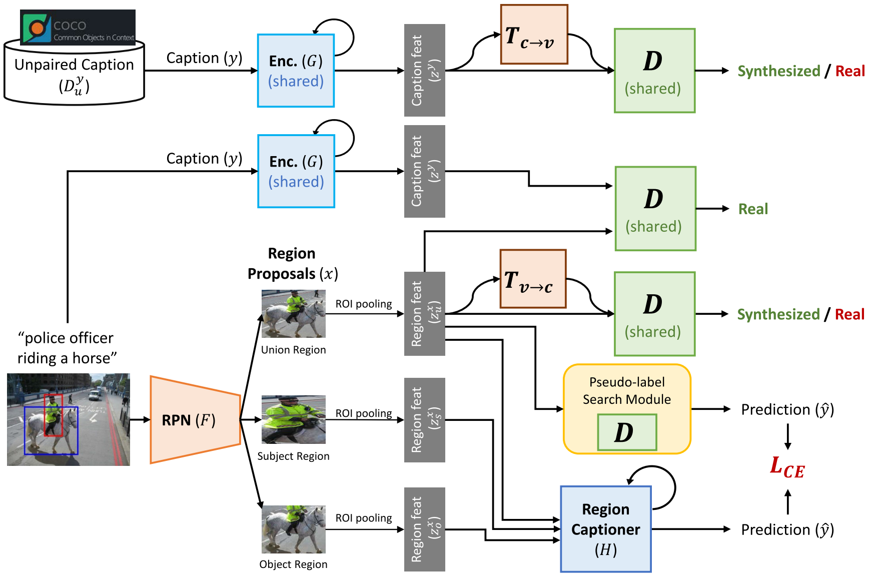


FIGURE 3. Illustration of the proposed semi-supervised *region-based* image captioning structure. In addition to the paired region-based image captioning data $\mathcal{D}_p$, we leverage an external image captioning dataset as an unpaired caption dataset $\mathcal{D}_u^y$, and the instances having no caption label as an unpaired image dataset $\mathcal{D}_u^x$. (is it possible to add unpaired dataset like figure2?)

utilizing Glove word vector [74]. Then, in order to represent the global *image-level* concept, we average out all the vectors obtained from the image to form a concept vector $\mathbf{v}$ of the image. The concept vector encodes the semantic concept of the scene.

The total loss function for training our model is as follows:

$$\min_{F,G,H,\{T\}} \max_{D} \mathcal{L}_{cap}(F,H) + \lambda_1 \tilde{U}(F,G,\{T\},D) + \lambda_3 \mathcal{L}_{external}(F), \qquad (9)$$

where $\mathcal{L}_{cap}$ denotes the captioning loss defined in Eq. (7), $\tilde{U}$ the loss for adversarial training defined in Eq. (3), $\mathcal{L}_{external}$ the concept regression loss defined in Eq. (8), and $\lambda_1 = \lambda_2 = \lambda_3 = 0.1$.

B. THEORETICAL ANALYSIS

In this section, we analyze our minimax-style learning framework and its favorable guarantees, which include a global equilibrium exists in our learning framework and it is also achievable. These analyses show that our design of the system and loss functions are well-grounded to pursue our objective of the multi-modal distribution match. To reach this conclusion, we first show the following Lemma 1.

Lemma 1: For any fixed generators F, G, and $\{T\}$, the optimal discriminator D of the minimax game defined by the objective function $U(F,G,\{T\},D)$ in Eq. (4) is

$$D^*(\mathbf{z}^x, \mathbf{z}^y) = \frac{p(\mathbf{z}^x, \mathbf{z}^y)}{p(\mathbf{z}^x, \mathbf{z}^y) + p_{1/2}(\mathbf{z}^x, \mathbf{z}^y)}, \qquad (10)$$

where $p_{1/2}(\mathbf{z}^x, \mathbf{z}^y) = \frac{(p_{v\to c}(\mathbf{z}^x,\mathbf{z}^y) + p_{c\to v}(\mathbf{z}^x,\mathbf{z}^y))}{2}$ is a mixture distribution.

This shows that the optimal discriminator D^* is at the balance between the true data distribution and the mixture distribution defined by F, G, and $\{T\}$. Given the fixed $D^*(\mathbf{z}^x, \mathbf{z}^y)$, we can reformulate the minimax game with the function $U(F,G,\{T\},D)$ as minimizing the sub-problem $V(F,G,\{T\}) = \max_D U$ over F, G and $\{T\}$. Then, we have the following lemma.

Lemma 2: Given $D = D^*(\mathbf{z}^x, \mathbf{z}^y)$, the global minimum of $V(F,G,\{T\})$ is achieved if and only if $p(\mathbf{z}^x, \mathbf{z}^y) = p_{1/2}(\mathbf{z}^x, \mathbf{z}^y)$, and the optimum value is $-\log 4$.

Furthermore, the marginal distributions $p(\mathbf{z}^x)$ and $p(\mathbf{z}^y)$ can be captured by the learned marginal distributions, *i.e.* $p(\mathbf{z}^y) = \underset{c\to v}{p}(\mathbf{z}^y) = \underset{v\to c}{p}(\mathbf{z}^y)$ and $p(\mathbf{z}^x) = \underset{c\to v}{p}(\mathbf{z}^x) = \underset{v\to c}{p}(\mathbf{z}^x)$.

The standard adversarial training in GAN [12] uses a similar way with Lemma 2 and shows that a generator perfectly replicates the data generating process if the optimal discriminator can be found. However, Lemma 2 shows only up to the fact that our model can at least replicate data marginal distributions and a mixture of $\{T\}$ can replicate the joint data distribution. In the next step, we show that we can actually find a global equilibrium point $p(\mathbf{z}^x, \mathbf{z}^y) = p_{v\to c}(\mathbf{z}^x, \mathbf{z}^y) = p_{c\to v}(\mathbf{z}^x, \mathbf{z}^y)$ that mimics the data generating (transformation) process in both directions as follows.

Theorem 1: Given an augmented objective function defined as:

$$U(F,G,\{T\},D) + \mathrm{KL}\left[p(\mathbf{z}^x|\mathbf{z}^y)||p_{c\to v}(\mathbf{z}^x|\mathbf{z}^y)\right] + \mathrm{KL}\left[p(\mathbf{z}^y|\mathbf{z}^x)||p_{v\to c}(\mathbf{z}^y|\mathbf{z}^x)\right]. \qquad (11)$$

The equilibrium of Eq. (11) is achieved if and only if $p(\mathbf{z}^x, \mathbf{z}^y) = p_{v\to c}(\mathbf{z}^x, \mathbf{z}^y) = p_{c\to v}(\mathbf{z}^x, \mathbf{z}^y)$.

Lemma 2 and Theorem 1 show that, without the additional regularization, the learned distribution is only matched up to marginal distributions and the true data distribution may be achieved with the non-unique mix of two distributions,

TABLE 1. Data source of each experiment setup. The numbers in the parentheses indicate the number of samples.

Experiments	Paired data	Unpaired images	Unpaired captions
Partially Labeled COCO (Sec. IV-B)	MS COCO caption (113k)	Unlabeled-COCO (123k)	–
Web-Crawled (Sec. IV-C)	0.5–1% of MS COCO caption (0.56k to 1.13k)	MS COCO image (112k)	Shutterstock (2.2M)
Relational Caption (Sec. IV-D)	Labeled regions of Rel.Cap.	Unlabeled regions of Rel.Cap.	MS COCO caption (113k × 5)
Scarcely-paired COCO (Sec. IV-E)	1% of MS COCO caption (1.3k)	99% of MS COCO caption (112k)	99% of MS COCO caption (112k × 5)

$p_{1/2}(\mathbf{z}^x, \mathbf{z}^y)$. With the additional regularization, Theorem 1 shows that the true distribution can be matched with the favorable unique global equilibrium guarantee. Finally, by Theorem 1, we can ensure that F, G, and $\{T\}$ will converge to the true distribution if F, G, and $\{T\}$ have enough capacity and each model has been trained to achieve the optimum.

Unfortunately, directly minimizing the KL divergence terms in Eq. (11) is infeasible in practice. In Eq. (3), we use the simple alternative of L_{reg} as a practical solution, which can be regarded as a Monte Carlo approximation of distribution matching and is proportional to those matching. Note that despite departing from the theoretical guarantees, the noticeable performance improvement in our empirical study suggests that our method is indeed a reasonable realization of the theory.

C. EXTENSION TO REGION-BASED CAPTIONING

Our semi-supervised learning method can be extended to other advanced visual captioning tasks. In this work, we extend our approach to region-based image captioning tasks, which require localizing object instances in the scenes [9], [75]. We especially focus on the relational captioning task [73], where the task is to caption the interactions of object instances in the visual scene, which can be regarded as a generalization of the instance-wise captioning [75].

The pipeline of the relational captioning [73] work is as follows. Given an input image, B number of object proposals from the region proposal network (RPN) [76] are obtained to localize each object instance. To take interactions between objects into account, the combination layer [73] produces the subject-object region pairs of the object proposals by assigning each instance into either subject or object role, *i.e.* $B\times(B-1)$ subject-object region pairs. Given a region pair, we obtain a triplet of features consisting of the subject ($\mathbf{z}^x_s$), object ($\mathbf{z}^x_o$), and the union of their regions ($\mathbf{z}^x_u$), as illustrated in Fig. 3. In this task, our semi-supervised method (illustrated in Fig. 2) is applied to captions in the dataset (denoted as y) and the union region features (denoted as $\mathbf{z}^x_u$) in addition to the supervised loss with the target task data. Thereby, the learned model predicts a large number of relational captions describing each pair of objects in the input image.

As the region-based caption labels in the existing datasets [9], [75] are mostly in the form of subject-predicate-object triplet, most descriptive phrases in general can be thought of as following a similar form. Therefore, we postulate that it would be helpful to leverage more natural human-labeled language (caption) datasets as unpaired caption information $\mathcal{D}^y_u$. Also, in these region-based tasks, we can leverage the instances having no caption label (*i.e.* negative sample) as an unpaired image dataset $\mathcal{D}^x_u$ as well for further regularization.

IV. EXPERIMENTS

In this section, we describe the experimental setups and competing methods and demonstrate the performance of our semi-supervised captioning with both quantitative and qualitative results.

A. EXPERIMENTAL SETUPS

We utilize the MS COCO caption dataset [16] (we will refer to MS COCO for simplicity) as our target dataset, which contains 123k images with 5 caption labels per image. To validate our model, we follow *Karpathy* splits [1], which have been broadly used in various image captioning works. The Karpathy splits contain 113k training, 5k validation, and 5k test images in total. In our experiment, to simulate the scenario that both paired and unpaired data exist, we use four different setups: 1) partially labeled COCO [42], 2) web-crawled data [33], 3) relational captioning data [9], and 4) scarcely-paired COCO setup we proposed. The data source of each experiment is described in Table 1.

We set the channel size to be 1024 for the hidden layers of LSTMs, 512 for the attention layer, and 1024 for the word embeddings. We use a mini-batch of size 100 and the Adam optimizer for training with the hyper-parameters $lr{=}5e^{-4}$, $b_1{=}0.9$, $b_2{=}0.999$. We set λ_x and λ_y to be equal to 0.1.

For evaluation, we use the following metrics conventionally used in image captioning: BLEU [77], ROUGE-L [78], SPICE [79], METEOR [80], and CIDEr [81]. All the evaluation is done on the MS COCO caption test set.

B. EVALUATION ON PARTIALLY LABELED COCO

For the *partially labeled* COCO experiment, we follow Liu et al. [42] and use the whole MS COCO caption data (paired) and add the *"Unlabeled-COCO"* split. The Unlabeled-COCO split includes unpaired images from the official MS COCO dataset [16], which involves 123k images without any caption label (no additional unpaired caption is used). Note that the MS COCO caption dataset and the Unlabeled-COCO split do not have overlapped data. In this setup, a separate *unpaired* caption data $\mathcal{D}^y_u$ does not exist. To compute the cross-entropy loss, we apply the pseudo-label assignment to the Unlabeled-COCO images. We use captions from the paired COCO data $\mathcal{D}_p$ as pseudo-label candidates.

We compare on different advanced backbone architectures equipped with attention mechanism [5], [7], [23], self-attention approach [14], and the recent Transformer based architecture [6], [21], which were originally developed for fully-supervised methods. We use the same data setup as the above, but we replace CNN (F) and LSTM (H) in our

TABLE 2. Evaluation of our method with different backbone architectures as an add-on module. All models are reproduced and trained with the fully paired MS COCO caption data and the cross entropy loss. Our training method, with adding the Unlabeled-COCO images, is applied to each method in a semi-supervised way, which shows consistent improvement in all the metrics.

	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE-L	SPICE	METEOR	CIDEr
NIC [7]	72.5	55.1	40.4	29.4	52.7	18.2	25.0	95.7
NIC **+Ours**	**74.4**	**57.5**	**43.0**	**32.1**	**54.3**	**19.5**	**25.8**	**101.0**
Att2in [23]	74.5	58.1	44.0	33.1	54.6	19.4	26.2	104.3
Att2in **+Ours**	**76.2**	**59.4**	**46.1**	**35.4**	**55.4**	**20.0**	**26.8**	**105.3**
Up-Down [5]	76.7	60.7	47.1	36.4	56.6	20.6	27.6	113.6
Up-Down **+Ours**	**77.5**	**64.2**	**47.8**	**37.1**	**57.5**	**21.2**	**28.0**	**117.9**
AoANet [14]	76.5	61.5	47.1	35.8	56.6	20.8	27.7	116.0
AoANet **+Ours**	**77.3**	**62.2**	**48.5**	**37.8**	**57.5**	**21.3**	**28.1**	**119.2**
$\mathcal{M}^2$ Transformer [6]	77.3	61.3	47.8	37.2	57.6	21.2	28.1	120.8
$\mathcal{M}^2$ Transformer **+Ours**	**78.8**	**62.7**	**49.0**	**38.7**	**58.3**	**22.0**	**28.7**	**126.2**
GRIT [21]	81.9	66.8	52.9	40.9	59.5	23.5	29.6	135.5
GRIT **+Ours**	**82.5**	**67.4**	**53.1**	**41.3**	**60.2**	**24.2**	**30.2**	**138.2**

TABLE 3. Performance comparison with web-crawled data. On top of unpaired image and caption data, our method is trained with 0.5 – 1% of paired data, while Feng et al. and Guo et al. use 36M additional images of the OpenImage dataset. We also show the relative performance improvements before and after applying the concept transfer on our method and add the comparison with a baseline "Concept," which only uses paired data with the concept transfer for reference. Applying the concept transfer significantly improves the image captioning performance, especially when the number of paired samples is scarce.

	BLEU4	ROUGE-L	SPICE	METEOR	CIDEr
Paired only (0.5% paired)	4.4	33.7	3.6	10.8	8.6
Ours (0.5%)	5.4	34.6	4.2	12.0	10.5
Concept (0.5%)	11.7	40.7	7.0	14.2	27.0
Ours + Concept (0.5%)	**12.2** (×2.3)	**41.6** (×1.2)	**7.6** (×1.8)	**14.8** (×1.2)	**30.5** (×2.9)
Paired only (0.7% paired)	3.5	36.1	3.7	11.4	8.9
Ours (0.7%)	8.5	39.0	5.2	13.6	20.2
Concept (0.7%)	13.3	42.0	8.5	15.4	33.6
Ours + Concept (0.7%)	**14.3** (×1.7)	**42.8** (×1.1)	**9.2** (×1.8)	**16.3** (×1.2)	**38.9** (×1.9)
Paired only (0.8% paired)	8.8	39.1	5.9	13.2	21.9
Ours (0.8%)	12.2	41.6	7.6	15.1	29.0
Concept (0.8%)	14.8	43.2	9.4	16.4	39.5
Ours + Concept (0.8%)	**15.5** (×1.3)	**44.0** (×1.1)	**10.2** (×1.3)	**16.9** (×1.1)	**42.0** (×1.4)
Paired only (1% paired)	13.4	41.9	8.3	15.9	36.0
Ours (1%)	15.2	43.3	9.4	16.9	39.7
Concept (1%)	16.2	44.0	10.1	17.2	44.5
Ours + Concept (1%)	**17.4** (×1.1)	**45.0** (×1.04)	**10.9** (×1.2)	**17.9** (×1.1)	**47.7** (×1.2)
Feng *et al*. [33]	5.6	28.7	8.1	12.4	28.6
Guo *et al*. [36]	6.4	31.3	9.1	13.0	29.0
Zhu *et al*. [38]	5.9	28.0	7.6	12.0	26.9
Zhu *et al*. [39]	10.0	35.8	11.5	16.2	45.8

framework with the image encoder and the caption decoder of their image captioning models. Then, these models are trained by our learning method as it is without the concept transfer method, which consists of alternating between the discriminator update and pseudo-labeling. Table 2 shows that training with the additional Unlabeled-COCO data via *our* training scheme consistently improves all the baselines in all the metrics.

C. EVALUATION ON WEB-CRAWLED DATA SETUP

To simulate a more realistic scenario involving crawled data from the web, we use the setup suggested by Feng et al. [33]. They collect a sentence corpus by crawling the image descriptions from Shutterstock[2] as unpaired caption data $\mathcal{D}_u^y$, whereby 2.2M sentences are collected. For unpaired image data $\mathcal{D}_u^x$, they use only the images from the MS COCO data, while the captions are not used for training. For training our method, we leverage from 0.5% to 1% of the paired MS COCO caption data as our paired dataset $\mathcal{D}_p$, *i.e.* very scarce data with a few hundreds or a thousand. This is an extremely challenging scenario as the paired and unpaired datasets are disjoint with different domains. In other words, there is no guarantee that all unpaired samples have their exact matches in the counterpart dataset. The results are shown in Table 3 including the comparison with Feng et al., Guo et al. [36], Zhu et al. [38], and Zhu et al. [39]. Note that all of Feng et al., Guo et al., and Zhu et al. exploit external large-scale data, *i.e.* 36M images of the OpenImages dataset. Up to 0.7% of paired-only data (793 pairs), the baseline shows lower scores in terms of BLEU4 and METEOR than Feng et al., while Ours shows comparable or favorable performance in BLEU4, ROUGE-L, and METEOR against Feng et al., Guo et al., and Zhu et al.. Ours starts to have significantly higher scores in all the metrics from 1% of paired data (1,133 pairs), even without external knowledge.

Moreover, additionally applying the concept transfer with additional loss in Eq. (8) by exploiting relational captions [9] (denoted as Ours + Concept) shows significant performance improvement, especially when the number of paired samples is scarce. Note that although applying the concept transfer to the Paired only baseline also shows noticeable performance improvement, combining both Ours and the concept transfer consistently shows the best performance in all settings. With 0.5% paired data, compared to our model without the concept transfer (Ours), our final model (Ours + Concept) shows nearly 2 times performance improvement on average; in particular, almost 3 times in terms of the CIDEr metric. Moreover, compared with a recent CLIP-based pseudo-labeling approach, Zhu et al. [39], our discriminator-based method outperforms the CLIP-based approach.

D. EVALUATION ON RELATIONAL CAPTIONING TASK

We apply our semi-supervised learning method to a dense relational object region based image captioning task, *i.e.*

[2] https://www.shutterstock.com

TABLE 4. **Evaluation of the relational dense captioning result with the Relational Captioning dataset [9]. We annotate the extended MTTSNet (MTTSNet + Relational module) by Kim et al. [73] with †. The extended MTTSNet trained with the proposed framework shows improvement over the one without the proposed framework by a noticeable margin.**

	mAP (%)	Img-Lv. Recall	METEOR
Direct Union	–	17.32	11.02
MTTSNet [9]	0.88	34.27	18.73
MTTSNet† [73]	1.12	45.96	18.44
MTTSNet† + Ours	**1.19**	**47.25**	**19.03**

TABLE 5. **Comparisons of the holistic image captioning on the Relational Captioning dataset [9].**

	Recall	METEOR	#Caption	Caption/Box
Relational Cap. (Union)	38.88	18.22	85.84	9.18
Relational Cap. (MTTSNet [9])	46.78	21.87	89.32	9.36
Relational Cap. (MTTSNet† [73])	56.52	22.03	80.95	9.24
Relational Cap. (MTTSNet† + **Ours**)	**61.40**	**23.88**	89.46	9.65

relational captioning [9]. For evaluation, we use the Relational Captioning dataset [9] consisting of 85,200 images with 75,456 / 4,871 / 4,873 splits for train / validation / test sets, respectively. We regard the whole paired Relational Captioning dataset as our paired data $\mathcal{D}_p$, and we utilize the captions from the MS COCO caption dataset as the unpaired caption dataset $\mathcal{D}_u^y$. In particular, we define the visual features in the training batch ($\mathbf{z}^x$) as the region features from individual object regions. As the relational captioning is a region based task, we utilize the negative regions with no captions label as the unpaired image dataset $\mathcal{D}_u^x$. We apply our method to the extended version of MTTSNet (MTTSNet + Relational embedding module annotated with †) by Kim et al. [73] and compare it with the other strong baselines.

We follow the evaluation protocols suggested by Kim et al. [9]. The relational dense captioning performance on the Relational Captioning dataset is shown in Tables 4 and 5. In addition, the relational dense captioning performance on the VRD dataset [82] is shown in Table 6. The extended MTTSNet trained with our proposed method shows an improvement by a noticeable margin over the MTTSNet counterpart in all the metrics and all the tables.

We also show the caption based image region-pair retrieval results in Fig. 4 as an application. As the Relational Captioning dataset might have limited generalizability, MTTSNet without the proposed framework (denoted as w/o Unpaired) shows several incorrect retrieval results, whereas the extended MTTSNet trained with our framework (denoted as w/ Unpaired) correctly retrieves image region-pairs. Note that, even if the MTTSNet without our framework retrieves correct images, the semantic reasoning in the region-pairs is incorrect when we do not leverage external knowledge. We also show the quantitative results of the retrieval in Table 7. Similar to the other experiment, the extended MTTSNet with our framework shows favorable image retrieval performance in all the metrics, which demonstrates our method is beneficial to the application level as well.

TABLE 6. **Evaluation on the relational dense captioning task with the VRD dataset [82]. The extended MTTSNet trained with our method shows the best performance among the baselines and the competing methods.**

	mAP (%)	Img-Lv. Recall	METEOR
Direct Union	–	54.51	25.53
MTTSNet [9]	2.18	71.44	35.47
MTTSNet† [73]	2.21	73.36	35.65
MTTSNet† + **Ours**	**2.58**	**73.47**	**35.97**
Language Prior [82]	2.13	46.60	28.12
Shuffle-Then-Assemble [83]	2.20	69.98	29.50

TABLE 7. **Caption-based image retrieval results on the Relational Captioning dataset [9]. Our framework improves the performance of the MTTSNet, even in the image retrieval application, across all the metrics.**

	R@1	R@5	R@10	Med
Image Cap. (Full Image RNN) [1]	9	27	36	14
Dense Cap. (DenseCap) [75]	25	48	61	6
Dense Cap. (TLSTM) [84]	27	52	67	5
MTTSNet [9]	29	60	73	4
MTTSNet† [73]	32	64	79	**3**
MTTSNet† + **Ours**	**34**	**66**	**82**	**3**
Random chance	0.1	0.5	1.0	-

E. ANALYSIS ON SCARCELY-PAIRED COCO

In order to understand the algorithmic characteristic of our method, we also provide an extensive and comprehensive analysis of our scarcely-paired COCO dataset. For the *scarcely-paired* COCO setup, we remove the pairing information of the MS COCO caption dataset, while leaving a small fraction of pairs unaltered. We randomly select only 1% of the total data as the paired training data $\mathcal{D}_p$, and remove the pairing information of the rest to obtain unpaired data $\mathcal{D}_u$. This dataset allows us to evaluate the proposed framework by assessing whether small paired data can lead to learning plausible pseudo-label assignment and what performance can be achieved compared to the fully supervised case. We follow the same setting with Vinyals et al. [7], if not mentioned. The performance evaluated on the MS COCO caption test set is reported.

In Table 8, we compare our method with several baselines: *Paired Only*; we train our model only on the small fraction (1%) of the paired data, *CycleGAN*; we train our model with the cycle-consistency loss [49]. Additionally, we train variants of our model denoted as *Ours* (ver1, ver2, and final). *Ours ver1* is the base model trained with our GAN model (Eq. (3)) that distinguishes real or fake image-caption pairs. Even without pseudo-labeling, GAN training unpaired image and caption data already helps better train the encoder networks in an unsupervised way, which improves the image captioning performance. As one could expect, semi-supervising with unpaired samples from MS COCO data is more helpful in improving the performance than with unpaired web-crawled samples in Table 3. *Ours ver2* adds training with pseudo-labeled unpaired data using Eq. (7) to *Ours ver1*, while setting the confidence scores $\alpha^x = \alpha^y = 1$ for all training samples. *Ours (final)* add the noise handling

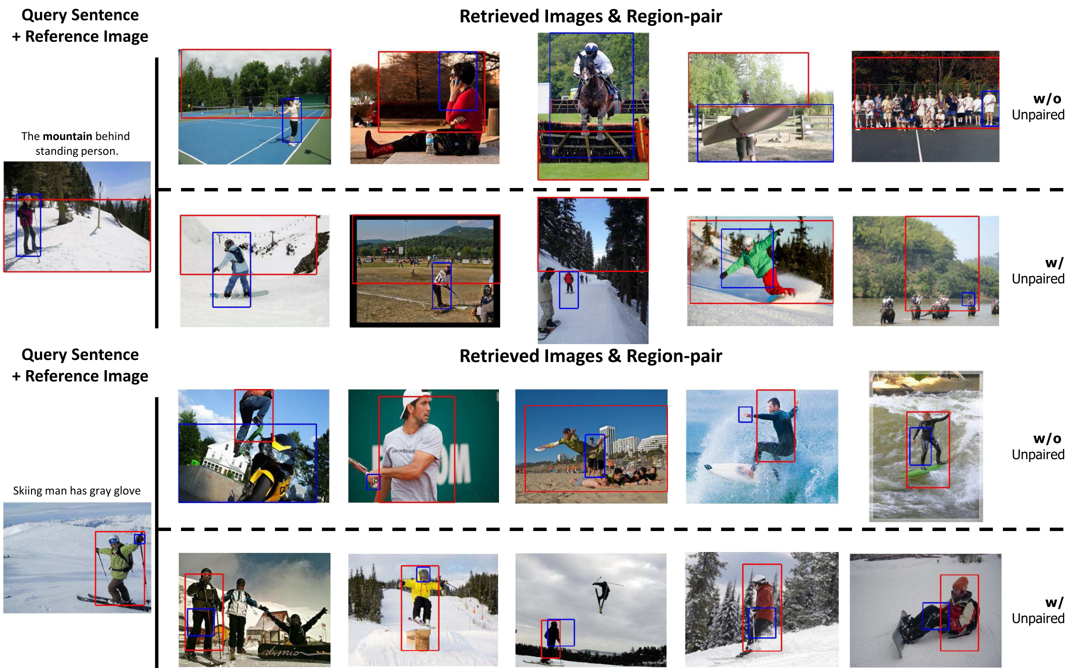

Query Sentence
+ Reference Image
Retrieved Images & Region-pair
The **mountain** behind standing person.
w/o Unpaired
w/ Unpaired
Query Sentence
+ Reference Image
Retrieved Images & Region-pair
Skiing man has gray glove
w/o Unpaired
w/ Unpaired

FIGURE 4. Qualitative results of the caption-based image retrieval on the Relational Captioning dataset [9]. The results are obtained by the relational captioning methods, which improve the caption-based image retrieval in multiple aspects. MTTSNet without the proposed framework (w/o Unpaired) shows a few incorrect retrieval results, whereas the extended MTTSNet trained with our framework (w/ Unpaired) correctly retrieves image region-pairs.

TABLE 8. Captioning performance comparison on the MS COCO caption test set. The "Paired only" baseline is trained only with 1% of paired data from our *scarcely-paired COCO* dataset. We denote the ablation study as: (A) the usage of the proposed GAN that distinguishes real or fake image-caption pairs, (B) pseudo-labeling, and (C) noise handling by sample re-weighting. In addition, we extend Ours (final) by adding the concept transfer and compare with a baseline "Paired only + Concept" for reference. We also compare with Gu et al. [31], Feng et al. [33], Lania et al. [35], Gu et al. [37], and Chen et al. [85] which are trained with unpaired datasets.

	(A)	(B)	(C)	BLEU1	BLEU4	METEOR	CIDEr
Fully paired (100%)				72.5	29.4	25.0	95.7
Paired only (1%)				58.1	13.4	15.9	36.0
Zhu *et al.* [49]				58.7	14.1	16.2	37.7
Ours ver1	✓			60.1	15.7	16.5	45.3
Ours ver2	✓	✓		61.3	17.1	19.1	51.7
Ours	✓	✓	✓	**63.0**	**18.7**	**20.7**	**55.2**
Gu *et al.* [31]				46.2	5.4	13.2	17.7
Feng *et al.* [33]				58.9	18.6	17.9	54.9
Lania *et al.* [35]				–	19.3	20.2	61.8
Gu *et al.* [37]				67.1	21.5	20.9	69.5
Chen *et al.* [85]				64.5	22.5	20.0	62.4
Zhu *et al.* [38]				-	21.5	20.1	65.7
Ours + Concept (final)	✓	✓	✓	**67.5**	**23.0**	**23.1**	**70.1**
Paired only + **Concept**				61.1	17.3	17.8	48.0

technique to *Ours ver2*, which is done by re-weighting each sample in the loss (Eq. (7)) with the confidence scores α^x and α^y. We present the accuracy of the fully supervised (*Fully paired*) model using 100% of the MS COCO caption training data for reference.

As shown in Table 8, in a scarce data regime, utilizing the unpaired data improves the captioning performance in terms of all metrics by noticeable margins. Also, our models show favorable performance compared to the CycleGAN model in all the metrics. Our final model with the pseudo-labels and the noise handling achieves the best performance in all metrics among the baselines. In addition, applying our concept transfer by utilizing relational captions [9] as an external knowledge (Ours+Concept) further improves the image captioning performance with noticeable margins. Note that the CIDEr score of our final model with the concept transfer is almost 2 times that of the Paired only baseline. Also, applying our concept transfer on the Paired only baseline shows lower improvement than that of Ours, indicating that the concept transfer is helpful when combined with our semi-supervised learning framework.

We also compare the recent unpaired image captioning methods [31], [33], [35], [37], [38], [85] in Table 8. In Gu et al. [31], the AIC-ICC image-to-Chinese dataset [86] is used as unpaired images $\mathcal{D}_u^x$ and the captions from the MS COCO caption dataset are used as unpaired captions $\mathcal{D}_u^y$. Note that our dataset setup is unfavorable to our method in that Gu et al. [31] use a far larger amount of additional labeled data (10M Chinese-English parallel sentences of the AIC-MT dataset [86]), Feng et al. and Laina et al. [35] use 36M samples of the additional OpenImages dataset, and Gu et al. [37] use scene graphs from the Visual Genome dataset [87] (108k). In contrast, our model only uses a small amount of paired samples (1k) and 122k unpaired data. Despite far lower reliance on paired data, our final model shows favorable performance against the recent unpaired image captioners in all the metrics.

V. CONCLUSION

We introduce a method to train an image captioning model with a large-scale unpaired image and caption data upon typical paired data. Our framework achieves favorable performance compared to various methods and setups. Unpaired captions and images are the data that can be easily collected from the web. It can also facilitate application-specific captioning models, *e.g.*, sign language recognition [88], [89], or visual question answering models [90], [91], [92] where labeled data is scarce. Furthermore, our semi-supervised learning method can be applied to various active learning scenarios [93], [94], [95], [96]. One of the potential directions to further improve our method may exploit the analogy between our method and GAN. Such directions might include research on the stability of the discriminator training or the study on the hyper-parameter sensitivity of the GAN model. These would be crucial ingredients to stimulate creative follow-up research.

REFERENCES

[1] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3128–3137.

[2] J. Choi, T.-H. Oh, and I. S. Kweon, "Contextually customized video summaries via natural language," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1718–1726.

[3] X. Wang, W. Xiong, H. Wang, and W. Y. Wang, "Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 37–53.

[4] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 563–578.

[5] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6077–6086.

[6] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10575–10584.

[7] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.

[8] T.-H. Chen, Y.-H. Liao, C.-Y. Chuang, W.-T. Hsu, J. Fu, and M. Sun, "Show, adapt and tell: Adversarial training of cross-domain image captioner," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 521–530.

[9] D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon, "Dense relational captioning: Triple-stream networks for relationship-based captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6271–6280.

[10] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 1–9.

[11] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "YFCC100M: The new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 1–9.

[13] D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon, "Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 1–12.

[14] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4633–4642.

[15] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," in *Proc. Assoc. Comput. Linguistics (ACL)*, 2014, pp. 1–12.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Zürich, Switzerland: Springer, 2014, pp. 740–755.

[17] X. Hu, Z. Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, and L. Wang, "Scaling up vision-language pretraining for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17959–17968.

[18] J. Ji, Z. Du, and X. Zhang, "Divergent-convergent attention for image captioning," *Pattern Recognit.*, vol. 115, Jul. 2021, Art. no. 107928.

[19] C.-W. Kuo and Z. Kira, "Beyond a pre-trained object detector: Cross-modal textual and visual context for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17948–17958.

[20] Y. Ma, J. Ji, X. Sun, Y. Zhou, and R. Ji, "Towards local visual modeling for image captioning," *Pattern Recognit.*, vol. 138, Jun. 2023, Art. no. 109420.

[21] V.-Q. Nguyen, M. Suganuma, and T. Okatani, "GRIT: Faster and better image captioning transformer using dual visual features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 167–184.

[22] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10968–10977.

[23] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1179–1195.

[24] J. Wang, W. Wang, L. Wang, Z. Wang, D. D. Feng, and T. Tan, "Learning visual relationship and context-aware attention for image captioning," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107075.

[25] X. Xiao, L. Wang, K. Ding, S. Xiang, and C. Pan, "Dense semantic embedding network for image captioning," *Pattern Recognit.*, vol. 90, pp. 285–296, Jun. 2019.

[26] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 1–10.

[27] Z. Yang, P. Wang, T. Chu, and J. Yang, "Human-centric image captioning," *Pattern Recognit.*, vol. 126, Jun. 2022, Art. no. 108545.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1–9.

[29] Z. Fang, J. Wang, X. Hu, L. Liang, Z. Gan, L. Wang, Y. Yang, and Z. Liu, "Injecting semantic concepts into end-to-end image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17988–17998.

[30] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[31] J. Gu, S. Joty, J. Cai, and G. Wang, "Unpaired image captioning by language pivoting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 503–519.

[32] M. Utiyama and H. Isahara, "A comparison of pivot methods for phrase-based statistical machine translation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol. (NAACL-HLT)*, 2007, pp. 1–8.

[33] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4120–4129.

[34] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, and A. Veit, "Openimages: A public dataset for large-scale multi-label and multi-class image classification," vol. 2, p. 3, 2017. [Online]. Available: https://github.com/openimages

[35] I. Laina, C. Rupprecht, and N. Navab, "Towards unsupervised image captioning with shared multimodal embeddings," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7413–7423.

[36] D. Guo, Y. Wang, P. Song, and M. Wang, "Recurrent relational memory network for unsupervised image captioning," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2020, pp. 1–8.

[37] J. Gu, S. Joty, J. Cai, H. Zhao, X. Yang, and G. Wang, "Unpaired image captioning via scene graph alignments," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10322–10331.

[38] P. Zhu, X. Wang, Y. Luo, Z. Sun, W.-S. Zheng, Y. Wang, and C. Chen, "Unpaired image captioning by image-level weakly-supervised visual concept recognition," 2022, *arXiv:2203.03195*.

[39] P. Zhu, X. Wang, L. Zhu, Z. Sun, W.-S. Zheng, Y. Wang, and C. Chen, "Prompt-based learning for unpaired image captioning," *IEEE Trans. Multimedia*, vol. 26, pp. 379–393, 2023.

[40] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[41] D.-J. Kim, J. Choi, T.-H. Oh, Y. Yoon, and I. S. Kweon, "Disjoint multi-task learning between heterogeneous human-centric tasks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1699–1708.

[42] X. Liu, H. Li, J. Shao, D. Chen, and X. Wang, "Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 338–354.

[43] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, "Deep compositional captioning: Describing novel object categories without paired training data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1–10.

[44] S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko, "Captioning images with diverse objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1170–1178.

[45] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7219–7228.

[46] H. Ben, Y. Pan, Y. Li, T. Yao, R. Hong, M. Wang, and T. Mei, "Unpaired image captioning with semantic-constrained self-learning," *IEEE Trans. Multimedia*, vol. 24, pp. 904–916, 2022.

[47] A. Jain, P. R. Samala, P. Jyothi, D. Mittal, and M. K. Singh, "Perturb, predict & paraphrase: Semi-supervised learning using noisy student for image captioning," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2021, pp. 1–7.

[48] Z. Meng, D. Yang, X. Cao, A. Shah, and S.-N. Lim, "Object-centric unsupervised image captioning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 219–235.

[49] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[50] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1857–1865.

[51] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 1–9.

[52] M. Artetxe, G. Labaka, E. Agirre, and K. Cho, "Unsupervised neural machine translation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–12.

[53] Z. Zhang, S. Liu, M. Li, M. Zhou, and E. Chen, "Joint training for neural machine translation models with monolingual data," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 1–8.

[54] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2004, pp. 1–8.

[55] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning," *IEEE Trans. Neural Netw.*, vol. 20, no. 3, p. 542, Oct. 2009.

[56] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Workshop Challenges Represent. Learn. (ICML)*, 2013, pp. 1–6.

[57] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, Aug. 2019.

[58] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 1–10.

[59] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "MixMatch: A holistic approach to semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2019, pp. 1–11.

[60] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "RemixMatch: Semi-supervised learning with distribution matching and augmentation anchoring," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–13.

[61] K. Sohn, D. Berthelot, C. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 1–13.

[62] C. Kuo, C. Ma, J. Huang, and Z. Kira, "FeatMatch: Feature-based augmentation for semi-supervised learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 479–495.

[63] L. Chongxuan, T. Xu, J. Zhu, and B. Zhang, "Triple generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 1–11.

[64] Z. Gan, L. Chen, W. Wang, Y. Pu, Y. Zhang, H. Liu, C. Li, and L. Carin, "Triangle generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 1–10.

[65] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1–15.

[66] W. Shi, Y. Gong, C. Ding, Z. MaXiaoyu Tao, and N. Zheng, "Transductive semi-supervised deep learning using min-max features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 299–315.

[67] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[68] D.-J. Kim, T.-W. Ke, and X. Y. Stella, "Local pseudo-attributes for long-tailed recognition," *Pattern Recognit. Lett.*, vol. 172, pp. 51–57, Aug. 2023.

[69] Y. Oh, D.-J. Kim, and I. S. Kweon, "DASO: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9776–9786.

[70] K.-H. Lee, X. He, L. Zhang, and L. Yang, "CleanNet: Transfer learning for scalable image classifier training with label noise," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5447–5456.

[71] Y. Wang, W. Liu, X. Ma, J. Bailey, H. Zha, L. Song, and S.-T. Xia, "Iterative learning with open-set noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8688–8696.

[72] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[73] D.-J. Kim, T.-H. Oh, J. Choi, and I. S. Kweon, "Dense relational image captioning via multi-task triple-stream networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7348–7362, Nov. 2022.

[74] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1–12.

[75] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully convolutional localization networks for dense captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4565–4574.

[76] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 1–9.

[77] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. Assoc. Comput. Linguistics (ACL)*, 2002, pp. 1–8.

[78] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 1–8.

[79] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, 2016, pp. 382–398.

[80] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. 9th Workshop Stat. Mach. Transl.*, 2014, pp. 1–5.

[81] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.

[82] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, 2016, pp. 852–869.

[83] X. Yang, H. Zhang, and J. Cai, "Shuffle-then-assemble: Learning object-agnostic visual relationship features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 36–52.

[84] L. Yang, K. Tang, J. Yang, and L.-J. Li, "Dense captioning with joint inference and visual context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1978–1987.

[85] X. Chen, M. Jiang, and Q. Zhao, "Self-distillation for few-shot image captioning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 545–555.

[86] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu, Y. Wang, and Y. Wang, "AI challenger: A large-scale dataset for going deeper in image understanding," 2017, *arXiv:1711.06475*.

[87] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, May 2017.

[88] Y. Jang, Y. Oh, J. W. Cho, D.-J. Kim, J. S. Chung, and I. S. Kweon, "Signing outside the studio: Benchmarking background robustness for continuous sign language recognition," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2022, pp. 1–23.

[89] Y. Jang, Y. Oh, J. W. Cho, M. Kim, D.-J. Kim, I. S. Kweon, and J. Son Chung, "Self-sufficient framework for continuous sign language recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[90] J. W. Cho, D.-J. Kim, Y. Jung, and I. S. Kweon, "Counterfactual mix-up for visual question answering," *IEEE Access*, vol. 11, pp. 95201–95212, 2023.

[91] J. W. Cho, D. M. Argaw, Y. Oh, D.-J. Kim, and I. S. Kweon, "Empirical study on using adapters for debiased visual question answering," *Comput. Vis. Image Understand.*, vol. 237, Dec. 2023, Art. no. 103842.

[92] J. W. Cho, D.-J. Kim, H. Ryu, and I. S. Kweon, "Generative bias for robust visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11681–11690.

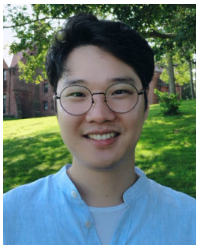
[93] J. W. Cho, D.-J. Kim, Y. Jung, and I. S. Kweon, "MCDAL: Maximum classifier discrepancy for active learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 8753–8763, 2022.

[94] D.-J. Kim, J. W. Cho, J. Choi, Y. Jung, and I. S. Kweon, "Single-modal entropy based active learning for visual question answering," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2021, pp. 1–15.

[95] D.-J. Kim, Z. Miao, Y. Guo, and S. X. Yu, "Modeling semantic correlation and hierarchy for real-world wildlife recognition," *IEEE Signal Process. Lett.*, vol. 30, pp. 259–263, 2023.

[96] I. Shin, D.-J. Kim, J. W. Cho, S. Woo, K. Park, and I. S. Kweon, "LabOR: Labeling only if required for domain adaptive semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8568–8578.

DONG-JIN KIM (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2015, 2017, and 2021, respectively. He was a Postdoctoral Researcher in EECS with UC Berkeley, in 2022. He is currently an Assistant Professor with Hanyang University. He was a Research Intern with the Visual Computing Group, Microsoft Research Asia (MSRA). His research interest includes data issues in computer vision, especially in high-level computer vision problems. He was awarded the Silver Prize from the Samsung Humantech Paper Award and the Qualcomm Innovation Award.



TAE-HYUN OH (Member, IEEE) received the B.E. degree (Hons.) in computer engineering as a Valedictorian from Kwang-Woon University, South Korea, in 2010, and the M.S. and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 2012 and 2017, respectively. He was the Research Director of OpenLab, POSCO-RIST, South Korea, jointly affiliated with POSTECH, South Korea. Before joining POSTECH, he was a Postdoctoral Associate with MIT CSAIL, Cambridge, MA, USA, and was with Facebook AI Research, Cambridge. He was a Research Intern with Microsoft Research, in 2014 and 2016. He is currently an Associate Professor of electrical engineering (with adjunct positions with the Graduate School of AI and the Department of Convergence IT) with POSTECH. He was a recipient of the Microsoft Research Asia Fellowship, the Samsung HumanTech Thesis Gold Award, the Qualcomm Innovation Award, and the Top Research Achievement Award from KAIST. He has served as an Area Chair for ICCV 2023, NeurIPS 2023–2024, CVPR 2024, ICLR 2024, and ACCV 2024, and as a Senior Program Chair for AAAI 2022. He is also an Associate Editor of IJCV, TVCJ, and ICRA 2023–2024. He was also selected as an Outstanding Reviewer for CVPR 2020 and ICLR 2022.

JINSOO CHOI received the B.S., M.S., and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), in 2013, 2015, and 2020, respectively. His research interests include deep learning, computer vision, and computer graphics, with an emphasis on video enhancement and processing. He received the Grand Prize from the Electronic Times Paper Award hosted by the Ministry of Science and ICT, Republic of Korea; the Silver Prize from Samsung Electro-Mechanics Paper Award; the Silver Prize from the Samsung Humantech Paper Award; the Qualcomm Innovation Award; and recognition as top research achievements and top 1% research achievements from KAIST Annual and Biannual Research and Development Reports.

IN SO KWEON (Member, IEEE) received the B.S. and M.S. degrees in mechanical design and production engineering from Seoul National University, Seoul, South Korea, in 1981 and 1983, respectively, and the Ph.D. degree in robotics from the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, in 1990. He was with the Toshiba Research and Development Center, Japan. He joined the Department of Automation and Design Engineering, Korea Advanced Institute of Science and Technology, Seoul, in 1992, where he is currently a Professor with the Department of Electrical Engineering. His research interests include camera and 3D sensor fusion, color modeling and analysis, visual tracking, and visual SLAM. He is a member of KROS. He was a recipient of the Best Student Paper Runner-Up Award from the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009). He was the Program Co-Chair of the Asian Conference on Computer Vision (ACCV 2007) and was the General Chair for ACCV 2012. He is also on the editorial board of the *International Journal of Computer Vision*.