# Crossing the Gap: Domain Generalization for Image Captioning

Yuchen Ren[12], Zhendong Mao[13] *, Shancheng Fang[1], Yan Lu[2], Tong He[2],
Hao Du[1], Yongdong Zhang[13] and Wanli Ouyang[2]
[1]University of Science and Technology of China, Hefei, China
[2]Shanghai Artificial Intelligence Laboratory, Shanghai, China
[3]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, China

## Abstract

*Existing image captioning methods are under the assumption that the training and testing data are from the same domain or that the data from the target domain (i.e., the domain that testing data lie in) are accessible. However, this assumption is invalid in real-world applications where the data from the target domain is inaccessible. In this paper, we introduce a new setting called Domain Generalization for Image Captioning (DGIC), where the data from the target domain is unseen in the learning process. We first construct a benchmark dataset for DGIC, which helps us to investigate models' domain generalization (DG) ability on unseen domains. With the support of the new benchmark, we further propose a new framework called language-guided semantic metric learning (LSML) for the DGIC setting. Experiments on multiple datasets demonstrate the challenge of the task and the effectiveness of our newly proposed benchmark and LSML framework.*

## 1. Introduction

Image captioning (IC) builds a bridge between vision and language, and it aims at understanding images [20, 28, 36, 37, 51, 69] and generating correct natural language descriptions. Many novel methods [3, 13, 26, 27, 45, 55, 58] have made impressive progress under a domain-specific setting; namely, they assume the training and testing data are from the same domain. However, this assumption may not hold in real-world applications. To relax the reliance of different domains, many methods [10, 19, 24] are recently proposed. However, these approaches also have a strong assumption that the data from the target domain are available. In real-world applications, this assumption will also be invalid. For example, in the medical report generation task, the data from the target domain related to patient privacy is hard to obtain. As a result, it is important to design an image captioning approach with domain generalization ability to different unseen domains.

*Zhendong Mao is the corresponding author.

| Task | Training data | Test data | $\mathcal{Y}_S = \mathcal{Y}_T$ | Target access |
|---|---|---|---|---|
| PivotIC | $\mathcal{D}^{\text{src-p}}$, $\mathcal{D}^{\text{p-tar}}$ | $\mathcal{D}^{\text{src-tar}}$ | ✓ | ✓ |
| NOIC | $\mathcal{D}^{\text{src}}$, $\mathcal{D}^{\text{tar}}$ | $\mathcal{D}^{\text{tar}}$ | × | ✓ |
| DAIC | $\mathcal{D}^{\text{src}}$, $\mathcal{D}^{\text{tar}}$ | $\mathcal{D}^{\text{tar}}$ | × | ✓ |
| DGIC | $\mathcal{D}^{\text{src}}$ | $\mathcal{D}^{\text{tar}}$ | × | × |

Table 1. Comparison of domain-related tasks for image captioning. $\mathcal{Y}_{S/T}$: distribution of source/target label space. $\mathcal{D}^{\text{src/tar/p}}$ : source/target/pivot domain. Pivot image captioning (PivotIC), novel object image captioning (NOIC), and domain adaptive image captioning (DAIC) are all assumed to be able to obtain the data of the target domain. Domain generalizable image captioning (DGIC) does not require the target domain data.

To this end, we propose a new benchmark setting called **D**omain **G**eneralization for **I**mage **C**aptioning (DGIC) with multi-source domain and cross-dataset setting in this work. Specifically, we employ existing popular datasets from five domains: common domain sourced from MSCOCO [35], assistive domain sourced from Vizwiz [21], social domain sourced from Flickr30k [62], avian domain sourced from CUB-200 [44, 57], and floral domain sourced from Oxford-102 [44, 57]. To explore the DGIC, we divide these domains into two parts: multiple source domains for training and a target domain for testing, mimicking the unseen domain scenario and mining underlying patterns from multiple datasets. The difference between our DGIC setting and other image captioning settings is summarized in Tab. 1.

With the help of this benchmark, we analyze the existing methods for unseen domains and observe the following limitations: (1) The model can generate fluent captions but cannot ensure semantic correctness when meeting an unseen domain without target data (Fig. 1a). In other words, the generated captions are prone to overfit domain-specific bias and only learn the domain-specific features. We argue that this is because the existing image captioning models are trained with maximum likelihood estimation, which will cause the model lacks discriminative semantic information between different instances [5, 25]. So it is difficult to distinguish the relationship between unseen domain data

and learned data. Therefore, it is desirable to introduce semantic information in the learning process. (2) The existing domain generalization (DG) method designed for other tasks [6, 7, 31, 32, 39, 56] cannot be well applied directly to image captioning because the image and label of most DG tasks are simple. So these methods only use coarse-grained information in the learning process. But for the DGIC task, the image and label are often complex and contain rich textual information (Fig. 1b). Therefore, simply applying the existing DG method to the image captioning task may not work well, and it is beneficial to utilize the rich contextual information in images and labels when performing feature alignment of different domains.

To tackle the aforementioned two challenges, we propose a new framework called language-guided semantic metric learning (LSML) for the DGIC task. To solve the first issue (i.e., lack of semantic information), we introduce both inter-domain and intra-domain metric learning to help captioning models considers the semantic relationship among different instances in the learning process. Specifically, we leverage contrastive learning to pull semantically similar visual features closer and push the irrelevant visual features far away from each other, which makes features more discriminative, allowing the model to learn domain-independent features that are more easily generalized to unseen domains. To solve the second issue (i.e., contextual information utilization), we propose a visual word guidance and sentence guidance strategy in the learning process. Specifically, we use the visual word and sentence similarity to sample discriminative triplets, allowing the model to capture fine-grained contextual information. As a consequence, our LSML framework aims to achieve promising performance under the DGIC setting.

In a nutshell, our contributions are summarized below: (1) We make the first attempt to conduct the task of domain generalization for image captioning (DGIC), which is used to explore the generalization ability of existing models. To achieve this, we construct a benchmark from existing datasets for this task. (2) We propose a new framework called language-guided semantic metric learning (LSML), which uses both inter- and intra-domain metric learning to help the model better learn discriminative semantic information among different instances. We also introduce a language guidance strategy in the learning process to utilize the rich contextual information in the image and labels during the learning process. (3) Extensive experiments demonstrate that our language-guided semantic metric learning framework outperforms previous state-of-the-art methods by a large margin under the DGIC setting.

## 2. Related Work

**Domain-related image captioning:** Many novel methods [3, 13, 26, 45, 55, 58, 66] have made impressive progress

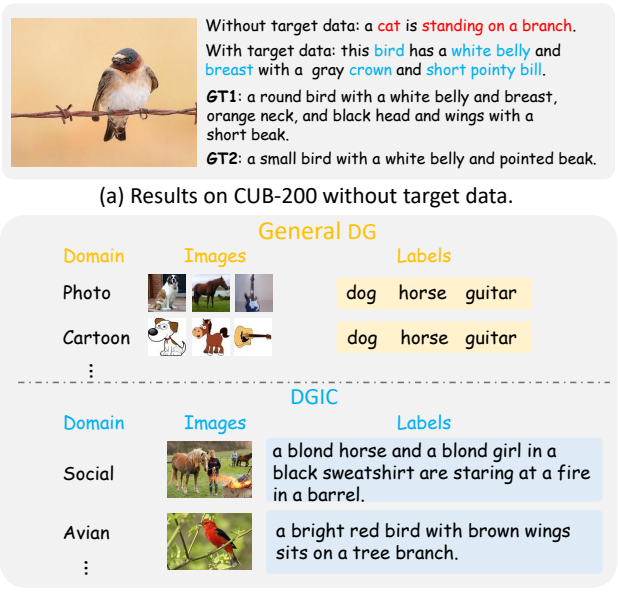



Figure 1. (a) Without using the target data, the model will overfit the domain-specific bias. (b) DGIC often has more complex images and labels compared to the general DG task.

under a domain-specific setting, *i.e.*, training and testing on the same domain. A few works attempt to relax the reliance on new domain image-caption data of different domains. Gu *et al.* [19] proposed the pivot image captioning task to generate captions in a pivot language and then translate the pivot language to a target language. Some works [1, 24] aim to describe novel objects not appearing in image captioning training data but existing in image recognition training datasets, which is referred to as novel object captioning. [10, 61, 67, 68] proposed to transfer the knowledge to a new domain under a domain adaption setting [49, 59, 60, 65], assuming that unpaired target domain data is available in training. Despite the improved performance, the methods mentioned above assume that the target domain is seen, which is impractical as new domain data is often unavailable in real-world scenarios for image captioning systems. Therefore, it is necessary to explore a method with a good generalization ability for unseen domains.

**Domain generalization:** Recently, some DG methods show impressive generalization ability for object recognition [6, 7, 31, 32, 38, 39, 56]. Li *et al.* [32] proposed to use Maximum Mean Discrepancy (MMD) measure to align latent features among different domains based on adversarial auto-encoder. Several methods [6, 56] attempted to introduce self-supervised learning to learn generic features and hence less over-fitting to domain-specific biases. Carlucci *et al.* [6] predicted relative positions of image patches to solve the Jigsaw puzzles problem, which can learn more generalizable semantic features among different domains. Similarly, Wang *et al.* [56] combined intrinsic self-supervision

| | M | V | F | C | O | |
|---|---|---|---|---|---|---|
| **M** | – | 0.0250 | 0.0090 | 0.0942 | 0.1385 | |
| **V** | 0.0002 | – | 0.0326 | 0.1262 | 0.1618 | |
| **F** | 0.0002 | 0.0004 | – | 0.1050 | 0.1506 | **Var** |
| **C** | 0.0067 | 0.0067 | 0.0067 | – | 0.1959 | 0.1034 |
| **O** | 0.0133 | 0.0100 | 0.0133 | 0.0067 | – | 0.1290 |

Table 2. Measuring the domain gaps of the DGIC benchmark with MMD. Orange is visual domain gaps over 2048-D ResNet [23] embedding, and green is linguistic domain gaps over 768-D BERT [16] embeddings.

and extrinsic relationship supervision using metric learning. There are also some methods that are specific to particular vision tasks, such as semantic segmentation [11, 12, 42, 63] and person Re-Identification [29,33,48,70]. However, to the best of our knowledge, no previous work has explored domain generalization for image captioning, so the rich contextual information is not well exploited in these works.

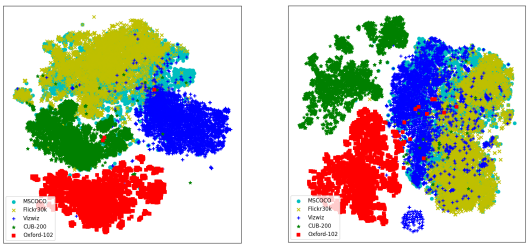
## 3. DGIC Benchmark

**Dataset construction:** To better analyze the domain generalization ability of different image captioning models, we propose a new benchmark called **D**omain **G**eneralization for **I**mage **C**aptioning (DGIC). We propose this DGIC benchmark because most domain generalization benchmarks focus on the image-based task without related textual information. Specifically, we employ existing popular datasets from five domains: common domain sourced from MSCOCO [35], assistive domain sourced from Vizwiz [21], social domain sourced from Flickr30k [62], avian domain sourced from CUB-200 [44, 57], and floral domain sourced from Oxford-102 [44, 57]. To explore the DGIC, we divide these five datasets into two parts: four domains as source domains for training and the other one as target domain for testing. More detailed information of each domain can be seen in the supplementary material.

**Measuring the domain gaps:** The current mainstream object recognition domain generalization task aims to target images of different styles, such as normal-style images, cartoon-style images, sketch-style images, and artistic-style images. However, our task targets different image-language scenarios, and their image-language semantics have respective focus scenarios. So we employ them as benchmarks. To analyze the domain gaps among different domains in our DGIC benchmark dataset, we follow many works [8, 64] to measure the domain gaps by using the Maximum Mean Discrepancy (MMD) [18]. The MMD distance between domains $\mathcal{D}^S$ and $\mathcal{D}^T$ can be measured according to the following equation:

$$\mathrm{MMD}(\mathcal{D}^S, \mathcal{D}^T)^2 = \|\mu_{\mathbb{P}_s} - \mu_{\mathbb{P}_t}\|^2_{\mathcal{H}}\,, \tag{1}$$

where $\mu_{\mathbb{P}_s} := \mathbb{E}_{\mathbf{s}\sim\mathcal{D}^S}[\phi(\mathbf{s})]$ and $\mu_{\mathbb{P}_t} := \mathbb{E}_{\mathbf{t}\sim\mathcal{D}^T}[\phi(\mathbf{t})]$ are the samples projected in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, and $\phi(\cdot) : \mathbb{R}^d \to \mathcal{H}$ represents a mapping operation. We study the MMD distance on the DGIC benchmark and report the results in Tab. 2. Also, we visualize the data and label from different domains by using t-SNE [52] in Fig. 2, from which we can see there are significant domain gaps. We also see that the label distributions of different domains are very diverse, and it is difficult to learn domain-independent features through representation learning of samples with the same label. Therefore, we propose a generalization framework to learn the model with better domain generalization ability.




(a) Visual domain gaps (b) Linguistic domain gaps

Figure 2. The t-SNE [52] visualization of the DGIC benchmark. Results reflect large visual domain gaps (a) and linguistic domain gaps (b) of cross-dataset settings.

## 4. Language-guided Semantic Metric Learning

### 4.1. Generalization for IC models

**Formulation:** For the domain generalizable image captioning, we consider a set of source domains $\mathcal{D}^S = \{\mathcal{D}^S_1, \ldots, \mathcal{D}^S_j, \ldots, \mathcal{D}^S_{N^S}\}$, with the j-th domain $\mathcal{D}^S_j$ having $N_j$ image-caption pairs $\{\mathbf{I}^j_i, \mathbf{C}^j_i\}$, where $\mathbf{I}^j_i$ is the i-th image in $\mathcal{D}^S_j$ and $\mathbf{C}^j_i$ is the corresponding caption describing $\mathbf{I}^j_i$. In the test stage, we aim to evaluate the model directly on a given unseen domain $\mathcal{D}^T$. The goal is to construct a model with the source domains which is able to generalize for the unseen target domain captioning.

**Challenges:** Given an input image $\mathbf{I}$ and model parameters $\theta$, an image captioning model with maximum likelihood estimation factors the distribution over possible output token sequence $\mathbf{C} = \{w_1, \ldots, w_t\}$ into a chain of conditional probabilities can be formulated as:

$$p_\theta(\mathbf{C} \mid \mathbf{I}) = \prod_{n=1}^{t} p_\theta(w_n \mid w_{0:n-1}, \mathbf{I}, \theta)\,. \tag{2}$$

Given a target caption $\mathbf{C}^* = \{w^*_1, \ldots, w^*_t\}$, Equation 2 can be modified with a cross-entropy loss the $\mathcal{L}_{XE}$:

$$\mathcal{L}_{XE}(\theta) = -\sum_{n=1}^{t} \log p(w^*_n | w^*_{1:n-1}, \mathbf{I}, \theta)\,. \tag{3}$$

To accelerate training, existing image captioning methods [13, 26] generate the text at time $t$ based on the ground-truth label of previous $t-1$ steps. However, when testing, the ground-truth label is unavailable. Therefore, we need to generate the text at time $t$ based on the predicted text of the previous $t-1$ steps. So the error will accumulate with the time step. This phenomenon is severer for unseen domains, because it is difficult to distinguish whether the generated text is semantically meaningful on the unseen

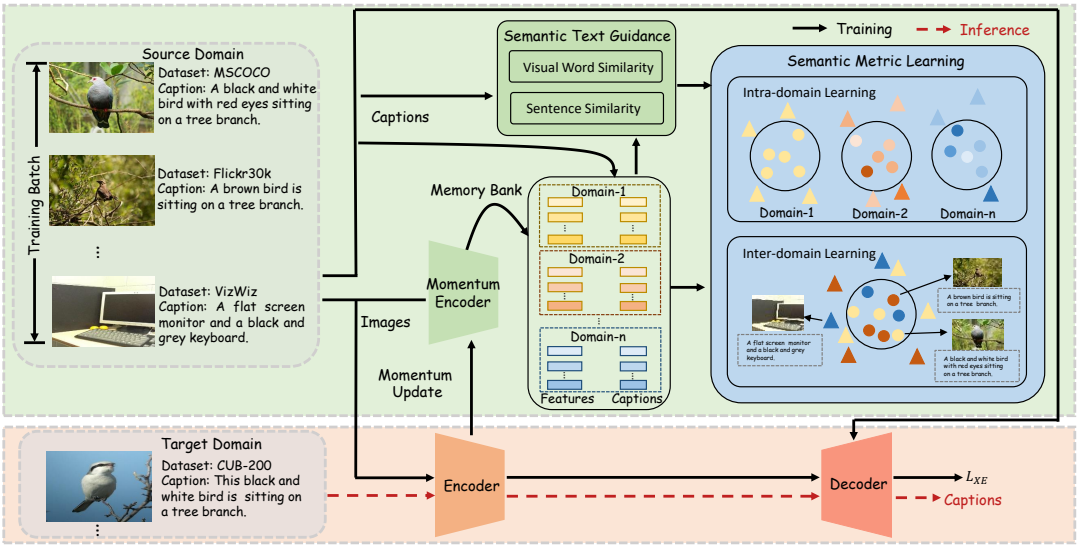



Figure 3. Our language-guided semantic metric learning framework.

domain. Therefore, the potential solution is to design new objectives that can improve discriminative semantically capability, such as metric/contrastive learning.

The second challenge is that most existing domain generalization methods cannot be directly applied to image captioning, because they are flawed for exploiting labels that are rich in continuous linguistic information. So we need to make reasonable use of fine-grained visual-linguistic information (e.g., object, attribute and relation) to improve semantic metric learning from the source domain without using target data.

To cope with the challenges, we propose our Language-guided semantic metric learning (LSML) framework based on language-guided semantic metric learning (Fig. 3).

## 4.2. Bridging domain gaps

The overall framework is shown in Fig. 3. We first store the features obtained by encoding multiple source domain images in a memory bank, then sample the positive and negative samples of multiple domains using semantic text guidance, and then use metric learning to constrain positive and negative samples among intra and inter domains, allowing the model to better learn discriminative information to improve generalization.

### 4.2.1 Semantic metric learning

**Inter-domain learning:** In order to better learn domain-independent visual components, we design inter-domain metric learning to improve model generalization with the triplet loss among multiple source domains. The triplet loss measures the similarity of samples and ensures that the distance between similar sample pairs is closer, and the negative samples are far away from the anchor sample, so as to learn a discriminative feature representation. In most cases, we can first choose a feature obtained from an encoder or decoder from a multi-source domain as the anchor sample $v_a$. Then, following [56], we choose other samples from the same category as the positive samples $v_p$, while samples from different categories as the negative samples $v_n$. Finally, these samples constitute the triplet that can be used for metric learning as:

$$Tri = \{(v_a, v_p, v_n) | y_a = y_p,\ y_a \neq y_n\}. \tag{4}$$

Unfortunately, unlike the classification task [17, 56], the DGIC task has different label spaces across datasets. We can simply use the cosine similarity of the image features to approximate whether the categories are the same or not here. The objective of the inter-domain metric learning can be formulated as:

$$\mathcal{L}_{inter} = \frac{1}{N_p}\frac{1}{N_n}\sum_{i=1}^{N_p}\sum_{j=1}^{N_n}[d(v_a, v_{p_i})^2 - d(v_a, v_{n_j})^2 + \delta]_+, \tag{5}$$

where $N_p$ and $N_n$ are the numbers of positive samples and negative samples, respectively, $d(\cdot,\cdot)$ denotes the Euclidean distance between two samples, $[\cdot]_+$ represents $max(0,\cdot)$, and $\delta$ is the margin of triplet loss.

**Intra-domain learning:** To ensure each domain has smooth representations for semantically related instances, we design intra-domain metric learning to assist inter-domain metric learning. Intra-domain metric learning treats each domain as a training task and aligns the features with multi-stream metric learning tasks. We can ensure semantically relevant samples from different domains share the similar embedding space, where different domain streams share the same backbone:

$$\mathcal{L}_{intra} = \sum_{D=1}^{N_s}\{\frac{1}{N_p}\frac{1}{N_n}\sum_{i=1}^{N_p}\sum_{j=1}^{N_n}[d(v_a, v_{p_i})^2 - d(v_a, v_{n_j})^2 + \delta]_+, v_a, v_{p_i}, v_{n_j} \in \mathcal{D}_j^S\}. \tag{6}$$

**Discussion:** For captioning models designed with the se-

mantic discriminative objective, the naive idea could be to use image data augmentation as positive samples and others as negative samples as the popular contrastive learning [9, 22]. But image data augmentation such as flipping or color jittering would cause the corresponding caption to change as well, and would not learn the semantic relationships between instances. Alternatively, like [15], matched image-caption pairs are used as positive samples and the unmatched pairs are used as negative samples, which only increase the uniqueness and cannot learn the similarity among instances to apply to unseen domains. Different from these methods, we focus on modeling the relation between instances among domains. In this way, we can capture domain-independent features for generalization.

#### 4.2.2 Semantic text guidance

**Visual word guidance:** Intuitively, when humans describe an image, the words of the visually obvious parts first appear in the mind, such as objects, actions, and attributes. Based on this inspiration, we employ the nouns, verbs, and adjectives in the caption to determine the degree of similarity between the two instances in the multi-source domain, and divide them into the anchor-positive sample pair or the anchor-negative sample pair. First, we parse the ground-truth caption $\mathbf{C}^* = \{w_1^*, \ldots, w_t^*\}$ of the source domain visual features $v$, and obtain the visual words group corresponding to each image through the part-of-speech:

$$Group_{vw} = \{w_n^* | \,\mathrm{POS}(w_n^*) \in \{\text{ noun, verb, adj }\}, \\ n \in \{0, \ldots, t\}\}, \tag{7}$$

where $\mathrm{POS}(\cdot)$ represents the part-of-speech of a word, the part-of-speech types of objects, actions and attributes are identified such as 'VB,' 'VBG,' 'NN,' 'NNS,' *etc*., also placing some words restrictions such as excluding 'is,' 'are,' *etc*. Next, we design the intersection of a visual word over union (IoU) $vwIoU$ to compute the distance between two samples $v^i$ and $v^j$, judge whether it is an anchor-positive sample pair or an anchor-negative sample pair according to the degree of visual words overlap:

$$vwIoU(v^i, v^j) = \frac{|Group_{vw}^i \cap Group_{vw}^j|}{|Group_{vw}^i \cup Group_{vw}^j|}. \tag{8}$$

**Sentence guidance:** With fine-grained text-guided alignment, we also design sentence-level guidance from a coarse-grained level. Sentence-level guidance can complement word temporal order and synonymy (e.g., car and vehicle). We use the sentence BERT embedding for cosine similarity calculation to get the sentence similarity, combine with the visual word similarity to get a total semantic similarity, and together to decide the triplet sampling:

$$sim_t(v^i, v^j) = max(vwIoU(v^i, v^j), s * sim_s), \tag{9}$$

where $s$ is the scaling factor. After obtaining the total semantic similarity $sim_t$ between the two samples, we can compare it with a threshold $p_{th}$. Those larger than the threshold are considered similar samples, and those smaller than the threshold are considered dissimilar samples. Then we propose the triplet selector can be formulated as:

$$Tri = \{(v_a, v_p, v_n) | sim_t(v^a, v^p) >= p_{th}, \\ sim_t(v^a, v^n) < p_{th}\}. \tag{10}$$

**Discussion:** It is worth mentioning that we use text as the guidance of the triple sampling instead of using images. The reasons are the following: (1) Not only language is ground truth at training and contains accurate object information, but also language contains rich knowledge that is easier to be measured than images. So, even for the text-to-image generation task where images are ground-truth at training time, the text is used as the guidance of the triple sampling. (2) By measuring modalities' domain gap, we find that the overall domain gap of languages is relatively smaller than that of images (see Fig. 2). By introducing the guidance from the text, we can effectively use the rich linguistic information to improve semantic metric learning.

#### 4.2.3 Good-hard negative samples mining

Moreover, [41, 46, 47, 50, 56] shows that mining hard negative samples in metric learning can effectively help the model correct its mistakes more quickly. There are some selection strategies such as semi-hard [47] and K-hard [56] negative selection. So the critical problem is: what is a **good** and **hard** negative sample for DGIC? We consider that for good-hard samples, visual words contained in the caption of the negative sample and the anchor do not overlap as much as possible. Because we hope that the image is more irrelevant to the content of the anchor image, which will increase the discriminative information of the model and allow the model to quickly learn more useful information. Therefore, we choose the **good-hard** negative samples with the smaller $sim_t$ of the anchor sample as the good-hard negative sample, and our language guidance can easily implement this strategy. We only need to set an upper bound $n_u$ and a lower bound $n_l$ for the negative samples $sim_t$. The new selection of the triplet can be formulated as:

$$Tri = \{(v_a, v_p, v_n) | sim_t(v^a, v^p) >= p_{th}, \\ n_l <= sim_t(v^a, v^n) < n_u\}, \tag{11}$$

where $n_u <= p_{th}$, our proposed metric learning equipped with the good-hard negative samples mining can select more informative hard negatives for each anchor, thus guiding the model to learn more discriminative features.

#### 4.2.4 Memory bank and momentum update

Due to the limitation of computational resources on batch size, insufficient positive and negative samples in a batch may lead to poor metric learning performance. We build a vision-language memory bank and momentum encoder to improve the efficiency of metric learning through better negative example mining (more details in the appendix). Moreover, unlike most existing contrastive learning methods that only sample negative instances, we also use the

| Method | Source→ Target | BLEU1 | BLEU2 | BLEU3 | BLEU4 | METEOR | ROUGE | CIDEr | SPICE |
|---|---|---|---|---|---|---|---|---|---|
| Up-Down [3] | | 29.00 | 15.26 | 8.12 | 4.47 | 11.42 | 27.78 | 38.63 | 12.12 |
| AoANet [26] | | 29.55 | 15.38 | 8.22 | 4.69 | 11.71 | 28.21 | 41.19 | 13.51 |
| $M^2$Transformer [13] | F+V+C+O→ M | 30.29 | 15.70 | 8.47 | 4.89 | 11.38 | 27.65 | 41.76 | 12.96 |
| EISNet [56] | | 29.12 | 15.10 | 8.14 | 4.61 | 11.95 | 27.88 | 42.39 | 13.42 |
| LSML (Ours) | | **32.22** | **17.31** | **9.66** | **5.53** | **12.91** | **29.63** | **53.24** | **16.08** |
| Method | Source→ Target | BLEU1 | BLEU2 | BLEU3 | BLEU4 | METEOR | ROUGE | CIDEr | SPICE |
| Up-Down [3] | | 20.64 | 9.31 | 4.11 | 1.99 | 6.16 | 20.07 | 13.75 | 3.65 |
| AoANet [26] | | 21.39 | 10.04 | 4.71 | 2.37 | 6.68 | 20.83 | 17.32 | 4.67 |
| $M^2$Transformer [13] | M+F+C+O→ V | 21.15 | 9.79 | 4.76 | 2.52 | 6.81 | 20.54 | 18.37 | 5.43 |
| EISNet [56] | | 20.70 | 9.91 | 4.73 | 2.28 | 6.79 | 20.52 | 17.89 | 4.78 |
| LSML (Ours) | | **21.93** | **10.42** | **5.21** | **2.78** | **7.41** | **21.09** | **21.25** | **6.14** |
| Method | Source→ Target | BLEU1 | BLEU2 | BLEU3 | BLEU4 | METEOR | ROUGE | CIDEr | SPICE |
| Up-Down [3] | | 25.60 | 13.54 | 7.09 | 3.85 | 11.36 | 26.41 | 36.04 | 14.13 |
| AoANet [26] | | 24.95 | 13.35 | 7.11 | 4.06 | 11.44 | 26.14 | 37.37 | 13.37 |
| $M^2$Transformer [13] | M+V+C+O→ F | 25.85 | 13.95 | 7.49 | 4.12 | 11.86 | 26.87 | 39.98 | 14.16 |
| EISNet [56] | | 27.15 | 14.70 | 7.91 | 4.36 | 13.04 | 28.62 | 45.99 | 15.64 |
| LSML (Ours) | | **28.08** | **15.80** | **8.76** | **4.69** | **14.03** | **30.03** | **53.51** | **17.01** |
| Method | Source→ Target | BLEU1 | BLEU2 | BLEU3 | BLEU4 | METEOR | ROUGE | CIDEr | SPICE |
| Up-Down [3] | | 16.57 | 5.14 | 0.85 | 0.27 | 7.99 | 20.18 | 3.70 | 14.56 |
| AoANet [26] | | 18.53 | 6.01 | 1.94 | 0.79 | 8.58 | 20.78 | 4.84 | 15.47 |
| $M^2$Transformer [13] | M+V+F+O→ C | 18.46 | 6.22 | 2.14 | 0.82 | 8.68 | 20.61 | 7.78 | 15.17 |
| EISNet [56] | | 19.47 | 6.38 | 2.44 | 1.07 | 8.82 | 21.16 | 6.83 | 15.20 |
| LSML (Ours) | | **20.39** | **8.01** | **3.19** | **1.31** | **10.24** | **20.93** | **9.60** | **15.72** |
| Method | Source→ Target | BLEU1 | BLEU2 | BLEU3 | BLEU4 | METEOR | ROUGE | CIDEr | SPICE |
| Up-Down [3] | | 19.01 | **6.07** | 1.57 | 0.33 | 8.07 | 17.34 | 9.04 | 12.38 |
| AoANet [26] | | 17.34 | 4.74 | 1.31 | 0.13 | 7.61 | 16.99 | 10.29 | 11.92 |
| $M^2$Transformer [13] | M+V+F+C→ O | 17.12 | 4.86 | 1.23 | 0.37 | 8.28 | 16.93 | 11.12 | 13.95 |
| EISNet [56] | | 17.70 | 5.13 | 1.57 | 0.63 | 8.62 | 17.44 | 11.38 | 12.52 |
| LSML (Ours) | | **19.41** | 5.90 | **2.17** | **0.85** | **9.72** | **17.95** | **14.35** | **15.23** |

Table 3. Comparison with state-of-the-arts image captioning and domain generalization methods on five settings in DGIC benchmark.

memory bank to sample positive instances, which constrains the semantics intra and inter domains to increase the diversity of domain samples.

#### 4.2.5 Objective function

We consider training our LSML framework with the objective function formulated as follows:

$$\mathcal{L}_{Total} = \alpha * \mathcal{L}_{XE} + \beta * \mathcal{L}_{inter} + \gamma * \mathcal{L}_{intra} \quad (12)$$

where $\alpha, \beta$and $\gamma$ are hyper-parameters to balance the weights of the caption generation supervision, the visual words metric learning for inter domain, and the visual words metric learning for intro domain, respectively. After the training process, we obtain our LSML model.

## 5. Experiments and analysis

### 5.1. The DGIC benchmark

**Datasets:** The DGIC benchmark consists of 253K images and 1,365K captions, sourced from MSCOCO [35], VizWiz [21], Flickr30K [62], CUB-200 [44, 57] and Oxford-102 [40, 44]. For exploring the DGIC, we divide these five datasets into two parts: four domains as source domains for training and the other one as target domain for testing. After permutation and combination, we can obtain five settings for experiments, and we follow the data split provided by Karpathy *et al*. [30]. Due to space limitations, more details can be found in the supplementary materials.

**Evaluation:** Following the standard evaluation protocol, we employ the full set of captioning metrics: BLEU [43], METEOR [4], ROUGE [34], CIDEr [54], and SPICE [2].

| Source | MSCOCO | | Source | Oxford-102 | |
|---|---|---|---|---|---|
| | CIDEr | SPICE | | CIDEr | SPICE |
| V+C | 24.12 | 8.56 | M+V | 10.37 | 12.62 |
| F+C | 44.68 | 13.7 | M+F | 12.10 | 13.32 |
| F+V+O | 51.43 | 15.62 | M+F+C | 12.71 | 13.82 |
| F+V+C | 51.91 | 15.68 | M+F+V | 14.13 | 14.87 |
| F+V+C+O | **53.24** | **16.08** | M+F+V+C | **14.35** | **15.23** |

Table 4. Comparison of training with different source domains. Experiments are conducted with our method.

| Method | BLEU4 | METEOR | ROUGE | CIDEr | SPICE |
|---|---|---|---|---|---|
| CCSA [38] | 4.52 | 11.57 | 27.72 | 40.48 | 13.03 |
| MMD-AAE [32] | 4.46 | 11.62 | 27.74 | 41.74 | 13.22 |
| EISNet [56] | **4.61** | **11.95** | **27.88** | **42.39** | **13.42** |

Table 5. Different DG methods on the MSCOCO target doamin.

| Method | CIDEr | SPICE | Object | Relation | Attribute |
|---|---|---|---|---|---|
| EISNet [56] | 42.39 | 13.42 | 25.23 | 1.60 | 3.60 |
| CL [15] | 42.04 | 13.38 | 24.32 | 1.54 | 4.04 |
| CompCap [14] | 43.51 | 13.28 | 23.90 | 1.60 | 3.47 |
| LSML (Ours) | **53.24** | **16.08** | **28.37** | **1.96** | **5.25** |

Table 6. CIDEr and breakdown of SPICE metrics.

### 5.2. Empirical studies and observations

We conduct extensive studies on DGIC and try to answer the following questions: Q1: What is the property of the DGIC benchmark? Q2: How about existing methods' generalization performance on DGIC? Q3: How is our LSML framework performed on DGIC? Q4: How is the effectiveness of different components in our LSML framework?

#### 5.2.1 DGIC benchmark properties (Q1)

**Domain gaps between different datasets:** Tab. 2 and Fig. 2 respectively show the visual and linguistic domain gaps of benchmarks across datasets from the perspectives of objective quantification and subjective visualization. In Tab. 2, we use Maximum Mean Discrepancy (MMD) to calculate the difference between the two distributions. The

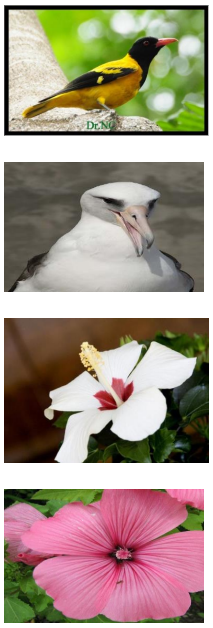


Figure 4. Examples of image captioning results by EISNet and our method, coupled with the corresponding ground-truth sentences.

greater the difference, the greater the value. In the orange triangle, we observe that the MMD values between MSCOCO, Flickr30k, and VizWiz are the smallest, which means that they are visually similar. However, from the visualization results, VizWiz still has a large discrepancy with MSCOCO and Flickr30k. Because Vizwiz's images come from low-quality photos taken by the phones of visually impaired people. The dataset with the biggest gap with other domains is Oxford-102 and then CUB-200. It can also be seen from the visualization results that they are obviously far away from other domains. In terms of linguistic domain gaps, MSCOCO and Flickr30k are similar, as expected, because they are both more generic. However, VizWiz also has a relatively small gap with them. We argue because although the images of VizWiz are blurry, many images are concerned with everyday activities, so the captions are closer to MSCOCO and Flickr30k covering daily life. CUB-200 and Oxford-102 still have a certain language gap with other datasets because there are all descriptions of birds or flowers. However, in general, we can see from the visualization and normalized variance that the gap of the language is still slightly smaller than that of the image, because the image is filled with a lot of complex scene noise in different domains.

**Analysis on the number of source domains:** Tab. 4 shows that more source domains can improve the generalization ability of the model. Especially, the improvement is very significant for generic domains such as MSCOCO, and there is also a small improvement for specific domains such as Oxford-102. It indicates multiple source domains have the potential to provide diverse patterns to facilitate model generalization. Therefore, we include multiple domains when constructing our DGIC benchmark.

#### 5.2.2 Existing methods' generalization on DGIC (Q2)

In Tab. 3, we report the results of different state-of-the-art image captioning and domain generalization methods on DGIC. We evaluate several most popular SOTA image captioning methods, including Up-Down [3], AoANet [26], $M^2$Transformer [13]. We see an unsatisfactory performance in all settings for these methods, where the model is trained without any target domain data or even images. This verifies that most existing image captioning models lack generalization capabilities on unseen domains. We note that the Transformer method generalizes the best due to modeling strong relationships by self-attention, achieving the best results under all five different settings of DGIC.

In Tab. 5, we also analyze some different classic DG methods: (1) Alignment-based methods (CCSA [38]); (2) Maximum Mean Discrepancy-based methods (MMD-AAE [32]); (3) Metric learning methods (EISNet [56]). To cope with the challenge of applying domain adaptation over image captioning, we employ the vanilla Transformer as the encoder-decoder backbone, combined with these DG methods. Experiments show that better generalization performance can be achieved by metric learning, which approximates semantically similar features for image captioning.

#### 5.2.3 Performance of LSML (Q3)

**Quantitative analysis:** To verify that our framework is more suitable for DGIC than existing methods, we compare our LSML method with the representative method EISNet [56] on all settings. EISNet [56] uses simple image category labels to guide metric learning for a better generalization. For a fair comparison, both methods use a basic Transformer [53] as the backbone. In Tab. 3, we observe that the performance of our method is better than other existing methods across the five settings. Specifically, our method outperforms prior methods by a considerable margin on CIDEr and SPICE, such as 10.85% on CIDEr and 3.56% on SPICE under MSCOCO target domain setting, which are specially designed metrics for image captioning and can more accurately evaluate sentence generation in the multimodal task.

Under the setting when the target domain is CUB-200 or Oxford-102, the domain gaps between these two datasets and other datasets are very large (see 5.2.1), where other datasets rarely contain captions that specifically describe the appearance of birds and flowers. So it is relatively difficult to generate such specific domain words for the challenging domains. However, our LSML method still surpasses existing methods by a considerable margin, which demonstrates the effectiveness of our LSML framework.

**Qualitative analysis:** Because CUB-200 and Oxford-102 are very challenging specific domains, we show some

| $\mathcal{L}_{inter}$ | $\mathcal{L}_{intra}$ | B@4 | M | R | C | S |
|---|---|---|---|---|---|---|
| × | × | 4.46 | 12.23 | 27.79 | 43.47 | 13.77 |
| ✓ | × | 5.53 | 12.65 | 29.34 | 51.25 | 15.13 |
| × | ✓ | 4.86 | 12.55 | 28.62 | 47.36 | 14.55 |
| ✓ | ✓ | **5.53** | **12.91** | **29.63** | **53.24** | **16.08** |

Table 7. Comparison of loss function components.

| Method | Bleu@4 | Meteor | ROUGE | CIDEr | SPICE |
|---|---|---|---|---|---|
| w/ image | 4.46 | 12.23 | 27.79 | 43.47 | 13.77 |
| w/ sentence | 5.41 | 12.58 | 29.39 | 50.88 | 15.39 |
| w/ visual word | 5.44 | 12.90 | 29.41 | 51.55 | 15.85 |
| w/word & sent | **5.53** | **12.91** | **29.63** | **53.24** | **16.08** |

Table 8. Comparison of guidance components.

| Method | Ours | w/o MB | w/o MU | w/o MB&MU |
|---|---|---|---|---|
| CIDEr | **53.24** | 49.33 | 51.55 | 48.10 |

Table 9. Ablation analyes on the memory bank (MB) and momentum update (MU).

cases on the settings that target domains are CUB-200 and Oxford-102 to prove the better performance of our partial method in Fig. 4. In the top two images, the two captions generated by our model contain "a small yellow bird" and "a white and black bird" respectively, while the object detector cannot provide the appearance and color attributes of the object. This can reflect that our method can sufficiently mine the domain-independent features corresponding to the objects, actions, and attributes in the source domain to improve the generalization of unseen domains.

**Analysis on semantic improvement:** To better understand the improved fine-grained semantic performance, we show the breakdown of SPICE metric that evaluates objects, attributes and relations with several methods that have the most potential to apply to DGIC. CL [15] focus on matched or unmatched image-caption pairs, and CompCap [14] learns to predict compositional phrases. As shown in Tab. 6, LSML can better capture fine-grained semantics including complex attributes and relations of objects from semantic text guidance.

#### 5.2.4 Effectiveness of model's components (Q4)

We take MSCOCO as the target domain and conduct detailed studies on the effectiveness of different components.

**Loss function components:** We investigate the contributions of different components in Tab. 7. Inter-domain learning plays the most important role, capturing discriminative relationships between instances and providing the model with the ability to learn domain invariant features for captioning. Intra-domain learning can complement inter-domain learning by making features smoother across domains to enhance the robustness of discriminative learning.

**Memory bank and momentum update:** We also list the gains of the memory bank (MB) and the momentum update (MU) in Tab. 9. The results demonstrate that these two modules can significantly improve the efficiency of metric learning and the diversity of triplet sampling.

**Semantic guidance:** We explore the effectiveness of different semantic guidance approaches. Tab. 8 shows that the text guidance is able to sample more relevant samples than the image guidance, facilitating the mining of truly discriminative triples in multiple domains. We observe that visual word guidance has the best results through fine-grained alignment, which is not susceptible to the syntactic style and word position because high-level sentence guidance can help word guidance to complement synonymy.

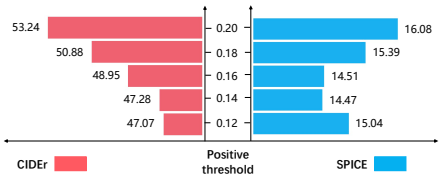



Figure 5. Comparison of different positive sampling thresholds.

| Range of Bounds | CIDEr | SPICE | Range of Bounds | CIDEr | SPICE |
|---|---|---|---|---|---|
| 0-0.02 | **53.24** | **16.08** | 0.02-0.06 | 51.02 | 14.95 |
| 0.02-0.04 | 52.62 | 15.70 | 0.06-0.1 | 49.78 | 15.76 |
| 0-0.2 | 48.63 | 14.97 | 0-0.1 | 49.96 | 14.61 |

Table 10. Comparison of different negative sampling' ranges of lower and upper bounds. When lower-upper bounds are 0-0.2, it means that the negative sample is randomly selected in the case of $sim_t(v^a, v^n) < p_{th}$. Other lower and upper bounds indicate good-hard negative sample mining.

**Triplet sampling:** We show the performance of different positive and negative sampling thresholds for triplet sampling in Fig. 5 and Tab. 10. The larger the sampling threshold $p_{th}$ of the positive sample, the better the model effect. Nonetheless, we need to make a trade-off because too large the threshold may lead to selecting no anchor-positive pair. When performing good-hard negative sample mining, the smaller the good-hard negative sampling range of the lower $n_l$ and upper $n_u$ bounds, the easier it is for the model to learn the discriminative information.

## 6. Conclusion

In this work, we propose a novel setting called domain generalizable image captioning (DGIC), where the data from the target domain is inaccessible. We first construct a benchmark dataset under this setting and analyze the limitations of existing methods for DGIC. With the analysis, we introduce the improved language-guided semantic metric learning framework called LSML, which can better generalize to the unseen image captioning domain. We can make better generalizations in the future through large-scale pre-trained models with some techniques such as prompt tuning, and our benchmarks can have a good role in facilitating and evaluating these models.

**Acknowledgements** We really thank Prof. Jinyang Guo for the help with many useful discussions and revising. This work was done during Yuchen's internship at Shanghai AI Lab and partially supported by the National Key R&D Program of China (No.2022ZD0160100), Shanghai Committee of Science and Technology (Grant No.21DZ1100100), the National Natural Science Foundation of China (Grant No.62222212 and No.62102384), Science Fund for Creative Research Groups Grant No.62121002.

## References


[1] Harsh Agrawal, Karan Desai, Xinlei Chen, Rishabh Jain, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *ICCV*, 2019. 2

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic Propositional Image Caption Evaluation. In *ECCV*, 2016. 6

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 1, 2, 6, 7

[4] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshops*, 2005. 6

[5] Shuyang Cao and Lu Wang. Cliff: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *EMNLP*, 2021. 1

[6] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, 2019. 2

[7] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *ECCV*, 2020. 2

[8] Qingchao Chen, Yang Liu, and Samuel Albanie. Mind-the-gap! unsupervised domain adaptation for text-video retrieval. In *AAAI*, 2021. 3

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*. 5

[10] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. Show, adapt and tell: Adversarial training of cross-domain image captioner. In *ICCV*, 2017. 1, 2

[11] Wuyang Chen, Zhiding Yu, Zhangyang Wang, and Animashree Anandkumar. Automated synthetic-to-real generalization. In *ICML*, 2020. 3

[12] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *CVPR*, 2021. 3

[13] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-Memory Transformer for Image Captioning. In *CVPR*, 2020. 1, 2, 3, 6, 7

[14] Bo Dai, Sanja Fidler, and Dahua Lin. A neural compositional paradigm for image captioning. In *NeurIPS*, 2018. 6, 8

[15] Bo Dai and Dahua Lin. Contrastive learning for image captioning. In *NeurIPS*, 2017. 5, 6, 8

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2018. 3

[17] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *ICCV*, 2013. 4

[18] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. In *NeurIPS*, 2006. 3

[19] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. Unpaired image captioning by language pivoting. In *ECCV*, 2018. 1, 2

[20] Jinyang Guo, Wanli Ouyang, and Dong Xu. Multi-dimensional pruning: A unified framework for model compression. In *CVPR*, 2020. 1

[21] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning Images Taken by People Who Are Blind. In *ECCV*, 2020. 1, 3, 6

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 5

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[24] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, 2016. 1, 2

[25] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *ICLR*, 2020. 1

[26] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on Attention for Image Captioning. In *ICCV*, 2019. 1, 2, 3, 6, 7

[27] Yiqing Huang, Jiansheng Chen, Wanli Ouyang, Weitao Wan, and Youze Xue. Image captioning with end-to-end attribute detection and subsequent attributes prediction. *IEEE TIP*, 2020. 1

[28] J. Guo, W. Ouyang, and D. Xu. Channel pruning guided by classification loss and feature importance. In *AAAI*, 2020. 1

[29] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *CVPR*, 2020. 3

[30] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 6

[31] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018. 2

[32] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018. 2, 6, 7

[33] Shengcai Liao and Ling Shao. Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. In *ECCV*, 2020. 3

[34] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL Workshops*, 2004. 6

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 1, 3, 6

[36] Aishan Liu, Jiakai Wang, Xianglong Liu, bowen Cao, Chongzhi Zhang, and Hang Yu. Bias-based universal adversarial patch attack for automatic check-out. In *ECCV*, 2020. 1

[37] Yuqing Ma, Shihao Bai, Wei Liu, Shuo Wang, Yue Yu, Xiao Bai, Xianglong Liu, and Meng Wang. Transductive relation-propagation with decoupling training for few-shot learning. *IEEE transactions on neural networks and learning systems*, 33(11):6652–6664, 2021. 1

[38] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*, 2017. 2, 6, 7

[39] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, 2013. 2

[40] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 6

[41] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, 2016. 5

[42] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*, 2018. 3

[43] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 2002. 6

[44] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 2016. 1, 3, 6

[45] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017. 1, 2

[46] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *ICLR*, 2021. 5

[47] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 5

[48] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *CVPR*, 2019. 3

[49] Peng Su, Shixiang Tang, Peng Gao, Di Qiu, Ni Zhao, and Xiaogang Wang. Gradient regularized contrastive learning for continual domain adaptation. In *AAAI*, 2020. 2

[50] Yumin Suh, Bohyung Han, Wonsik Kim, and Kyoung Mu Lee. Stochastic class-based hard example mining for deep metric learning. In *CVPR*, 2019. 5

[51] Shiyu Tang, Ruihao Gong, Yan Wang, Aishan Liu, Jiakai Wang, Xinyun Chen, Fengwei Yu, Xianglong Liu, Dawn Song, Alan Yuille, Philip H.S. Torr, and Dacheng Tao. Robustart: Benchmarking robustness on architecture design and training techniques. *https://arxiv.org/pdf/2109.05211.pdf*, 2021. 1

[52] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 3

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 7

[54] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In *CVPR*, 2015. 6

[55] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 1, 2

[56] Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. Learning from extrinsic and intrinsic supervisions for domain generalization. In *ECCV*, 2020. 2, 4, 5, 6, 7

[57] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-UCSD Birds 200. Technical report, California Institute of Technology, 2010. 1, 3, 6

[58] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 1, 2

[59] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *CVPR*, 05 2017. 2

[60] Hongliang Yan, Zhetao Li, Qilong Wang, Peihua Li, Yong Xu, and Wangmeng Zuo. Weighted and class-specific maximum mean discrepancy for unsupervised domain adaptation. *IEEE TMM*, 22(9):2420–2433, 2020. 2

[61] Min Yang, Wei Zhao, Wei Xu, Yabing Feng, Zhou Zhao, Xiaojun Chen, and Kai Lei. Multitask learning for cross-domain image captioning. *IEEE TMM*, 2018. 2

[62] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2014. 1, 3, 6

[63] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *ICCV*, 2019. 3

[64] Mingda Zhang, Tristan Maidment, Ahmad Diab, Adriana Kovashka, and Rebecca Hwa. Domain-robust vqa with diverse datasets and methods but no target labels. In *CVPR*, 2021. 3

[65] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *CVPR*, 2018. 2

[66] Zhiwang Zhang, Dong Xu, Wanli Ouyang, and Chuanqi Tan. Show, tell and summarize: Dense video captioning using visual cue aided sentence summarization. *IEEE TCSVT*, 30(9):3130–3139, 2020. 2

[67] Wentian Zhao, Xinxiao Wu, and Jiebo Luo. Cross-domain image captioning via cross-modal retrieval and model adaptation. *IEEE TIP*, 2020. 2

[68] Wei Zhao, Wei Xu, Min Yang, Jianbo Ye, Zhou Zhao, Yabing Feng, and Yu Qiao. Dual learning for cross-domain image captioning. In *CIKM*, 2017. 2

[69] Xiaowei Zhao, Xianglong Liu, Yuqing Ma, Shihao Bai, Yifan Shen, Zeyu Hao, and Aishan Liu. Temporal speciation network for few-shot object detection. *TMM*, 2023. 1

[70] Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Nicu Sebe. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *CVPR*, 2021. 3