Camouflage-aware Image-Text Retrieval via Expert Collaboration

Yao Jiang[1] Zhongkuan Mao[2] Xuan Wu[1] Keren Fu[1,2,*] Qijun Zhao[1,2]
[1]College of Computer Science, Sichuan University
[2]National Key Lab of Fundamental Science on Synthetic Vision, Sichuan University

Abstract

Camouflaged scene understanding (CSU) has attracted significant attention due to its broad practical implications. However, in this field, robust image-text cross-modal alignment remains under-explored, hindering deeper understanding of camouflaged scenarios and their related applications. To this end, we focus on the typical image-text retrieval task, and formulate a new task dubbed "camouflage-aware image-text retrieval" (CA-ITR). We first construct a dedicated camouflage image-text retrieval dataset (CamoIT), comprising ∼10.5K samples with multi-granularity textual annotations. Benchmark results conducted on CamoIT reveal the underlying challenges of CA-ITR for existing cutting-edge retrieval techniques, which are mainly caused by objects' camouflage properties as well as those complex image contents. As a solution, we propose a camouflage-expert collaborative network (CECNet), which features a dual-branch visual encoder: one branch captures holistic image representations, while the other incorporates a dedicated model to inject representations of camouflaged objects. A novel confidence-conditioned graph attention (C^2GA) mechanism is incorporated to exploit the complementarity across branches. Comparative experiments show that CECNet achieves ∼29% overall CA-ITR accuracy boost, surpassing seven representative retrieval models. The dataset and code will be available at https://github.com/jiangyao-scu/CA-ITR.

1. Introduction

Camouflaged scene understanding (CSU) is a comprehensive visual analysis task that aims to perceive and understand objects whose visual characteristics are highly similar to those of the surrounding background [14, 29]. As a challenging and rapidly advancing field within computer vision, it has attracted considerable attention owing to its wide-ranging applications in critical domains, including medicine, industry, and agriculture [4, 12–15, 38].

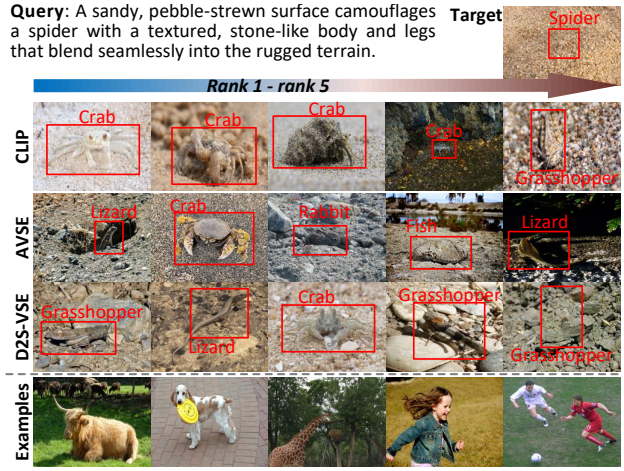

Figure 1. Qualitative results of three state-of-the-art (SOTA) retrieval methods (*i.e.*, CLIP [50], AVSE [40], and D2S-VSE [39]) on CA-ITR. Camouflaged objects in the images are marked with red bounding boxes. Below the dotted line are samples from the general ITR datasets (*i.e.*, MS-COCO [36] and Flickr30K [63]).

The field of CSU has undergone significant methodological evolution in recent years. Early approaches relied solely on visual cues, utilizing sophisticated neural networks to address the challenge of camouflaged object detection (COD) [14, 24, 26, 29, 44, 45, 55, 61, 67, 68]. Subsequent advancements integrated multimodal information to enhance detection performance or reduce reliance on dense annotations [21, 32, 49, 56, 64], thereby improving segmentation accuracy and expanding application scenarios. However, the core paradigm of these methods remains largely limited to segmentation tasks, generating pixel-wise masks that inherently lack the rich semantic information required for advanced cognitive tasks. Recently, emerging studies [51, 66] have explored multimodal large language models (MLLMs) in camouflaged scenarios, making a step further over the conventional segmentation frameworks. Despite such advances, these approaches still predominantly focus on the generative perspective, namely visual question answering and visual grounding, leaving explicit image-text cross-modal alignment largely under-explored.

Robust cross-modal alignment serves as the founda-

*Corresponding author: Keren Fu *(fkrsuper@scu.edu.cn)*.

tional bedrock for generalized multimodal understanding, underpinning critical capabilities such as visual question answering and detailed image captioning [3, 35]. Failure in such a step would inevitably undermine any downstream task. As illustrated in Fig. 1, for image-text retrieval, current state-of-the-art (SOTA) retrieval models frequently mismatch textual descriptions with incorrect visual objects in camouflaged scenes. Meanwhile, the struggles of advanced MLLMs on camouflaged scenes revealed by recent researches [27, 51], can also be likely attributed to this issue, since effective cross-modal understanding usually derives from robust initial alignment [33, 35].

In this spirit, we study the camouflage-aware image-text alignment problem via image-text retrieval, which is abbreviated as CA-ITR. Firstly, we construct a dedicated camouflage image-text retrieval dataset (CamoIT) by annotating images from existing COD datasets [11, 29, 43, 53] with comprehensive textual descriptions. CamoIT comprises ∼10.5K samples spanning diverse camouflage scenarios. Each image is paired with multi-granularity annotations, including category labels, detailed descriptions of camouflaged objects, and comprehensive image-level captions. Leveraging CamoIT, a comprehensive CA-ITR benchmark is conducted by evaluating open-source SOTA retrieval methods/models, whose results demonstrate that CA-ITR is not merely a domain transfer task but one that presents several distinct challenges, including the difficulty of perceiving camouflaged objects, complexity of image contents, and necessity for fine-grained understanding.

To study CA-ITR, we further propose a baseline model, namely camouflage-expert collaborative network (CECNet). It features a dual-branch visual encoder: one branch captures holistic image representations, while the other incorporates a dedicated COD model to assist representations of camouflaged objects. A novel confidence-conditioned graph attention (C^2GA) mechanism is used to exploit the complementarity across branches.

In a nutshell, our contributions are three-fold:

- **Task and dataset.** We formalize the new task of "camouflage-aware image-text retrieval" (CA-ITR) and construct CamoIT, which comprises ∼10.5K samples across 237 categories. CamoIT not only supports CA-ITR but also bridges high-level retrieval with low-level perception. To the best of our knowledge, this work is the first attempt in the field to propose and address image-text retrieval in camouflaged scenarios.
- **Benchmark and analyses.** We conduct a comprehensive CA-ITR benchmark, demonstrating that it is not merely a domain transfer task, but one characterized by unique challenges: perceiving camouflaged objects, handling complex image contents, and achieving fine-grained understanding.
- **New baseline model.** We propose a camouflage-expert collaborative network (CECNet) that incorporates a COD expert to enhance object perception and a novel confidence-conditioned graph attention (C^2GA) mechanism to refine feature fusion. Experimental results show that CECNet achieves ∼29% overall CA-ITR accuracy boost, surpassing seven representative retrieval models.

2. Related Work

2.1. Image-Text Retrieval

Image-Text Retrieval (ITR) methods can be divided into global alignment and local alignment. Global alignment methods [10, 17, 18, 22, 34, 39, 40, 52, 60, 65] match holistic image and text representations, but overlook fine-grained cross-modal correspondences, which limits their retrieval accuracy. In contrast, local alignment methods [7, 19, 30, 37, 46, 59] address this limitation to some extent by capturing fine-grained interactions between image regions and text words through mechanisms like cross-attention and graph-based matching. The implementation of these alignment strategies relies on different backbone architectures. Several approaches [18, 46, 60] adopt the BUTD framework [2], leveraging object-level visual representations to enhance retrieval. Other methods [19, 22, 39, 40] utilize learnable image encoders to extract image features. D2S-VSE [39] further explores a dense text distillation strategy to enhance cross-modal alignment. Although these methods have made significant progress in ITR, they are mainly trained on data with clearly separated foreground objects, leading to a notable performance drop when handling camouflaged objects that closely resemble the background.

2.2. Camouflaged Scene Understanding

Current research in CSU predominantly focuses on camouflaged object detection (COD) [4, 15], with extensions into related tasks such as camouflaged object instance segmentation [20, 41], camouflaged object counting [54], and camouflaged object ranking [43]. These COD methods rely solely on visual cues and employ diverse network architectures, including multi-stream frameworks [28, 47, 48, 61, 62, 67, 68], bottom-up and top-down integration [11, 14, 24, 26, 44, 45, 55], and auxiliary branches [29, 31, 43]. In recent years, multimodal textual and visual information has been increasingly applied in CSU. Several approaches [21, 32, 56] leverage MLLMs to reduce dependence on densely annotated data, utilizing visual reasoning and grounding capabilities to enable camouflaged object segmentation in open-world or zero-shot settings. ACUMEN [64] enhances segmentation by incorporating textual descriptions of camouflaged attributes, while OVCoser [49] enables open-vocabulary segmentation through target category integration. These methods are still largely confined to segmentation tasks, producing pixel-wise masks that lack the rich semantic information needed for advanced cognitive tasks.

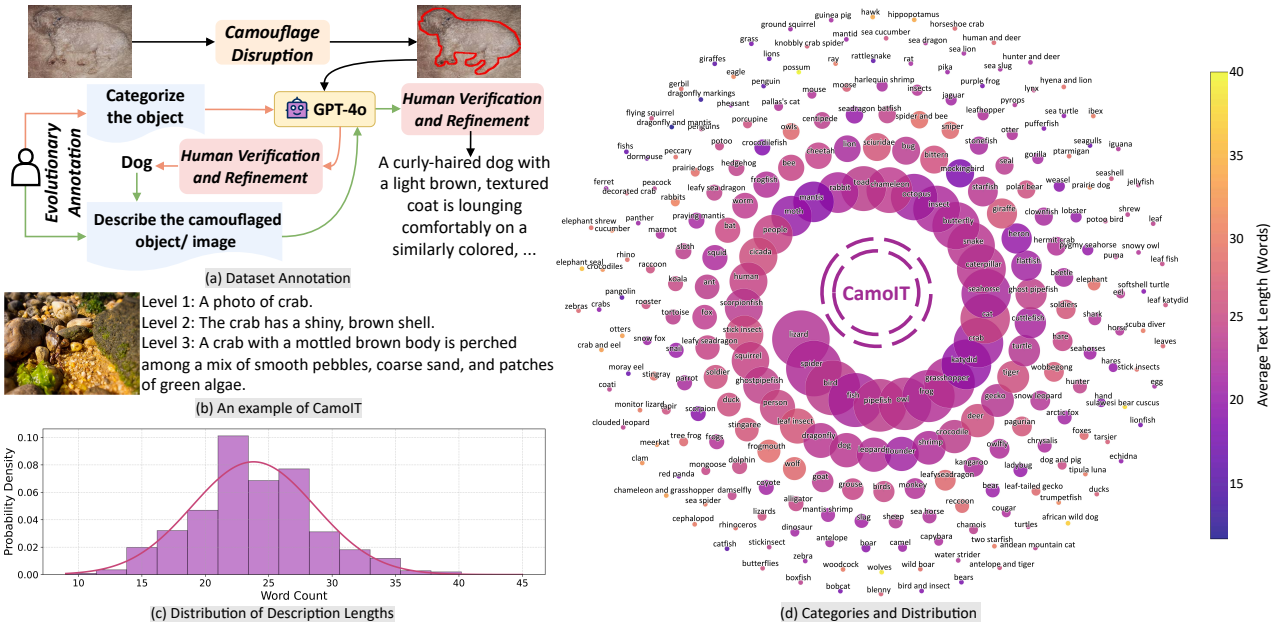

Figure 2. Data annotation process, an example, and statistical analyses of CamoIT.

Recently, several studies [51, 66] have developed multimodal large models for camouflaged scenarios. However, these approaches predominantly target generative tasks such as dialogue systems, leaving the core challenge of achieving robust cross-modal alignment in complex camouflaged environments largely unaddressed.

3. Construction of CamoIT

Our CamoIT is built upon existing COD datasets (*i.e.*, CHAMELEON [53], CAMO [29], COD10K [11], and NC4K [43]), from which we excluded low-quality samples, such as those with unrecognizable categories. As these datasets were designed originally for the segmentation task, they primarily provide mask annotations. To adapt them for the CA-ITR task, we utilized GPT-4o [25] to support data annotation, enabling the efficient generation of comprehensive textual descriptions. Our data annotation process consists of three components: camouflage disruption, evolutionary annotation protocol, and manual verification and refinement, as shown in Fig. 2 (a). The construction adheres to the annotation standards of MS-COCO [5].

Camouflage Disruption: To address GPT-4o's difficulty in perceiving camouflaged objects, we propose a camouflage disruption strategy. We extract object contours from COD masks and overlay them on the original images with a distinct color, effectively breaking the camouflage and enhancing object visibility for GPT-4o, as shown in Fig. 2 (a). It is important to note that the instruction "ignore the red marks" was included in the prompt in order to prevent the model from misinterpreting the artificially introduced highlights as an integral part of the scene.

Evolutionary Annotation: Given the complexity of describing an image with multiple components like key objects and the environment, it is difficult for a model to generate accurate descriptions for all parts at once. We therefore employ a multi-stage, progressive annotation strategy. As shown in Fig. 2 (a), this approach evolves from simple image classification to a description of the camouflaged object, and finally to a comprehensive image description. A key feature of this process is that the category names identified in the initial classification stage are explicitly incorporated into the prompts for subsequent stages, thereby enhancing the quality and accuracy of the final description.

Manual Verification and Refinement: To ensure high-quality annotations, we incorporate a mandatory manual verification and refinement stage. 16 trained annotators systematically conducted three rounds of description review and refinement in accordance with standardized guidelines. Their tasks included (1) correcting factual errors or hallucinations, and (2) improving conciseness by removing redundant content. This process required approximately two to four minutes per image.

CamoIT comprises a total of 10,464 samples, covering approximately 237 categories—including both animals and artificial camouflaged objects—across diverse environments such as marine and terrestrial scenes. Each image is assigned multi-granularity labels, as shown in Fig. 2 (b). It is noteworthy that CamoIT's multi-granularity annotations, accompanied by original masks, are suitable for a broad spectrum of tasks beyond ITR. For example, object-level annotations (Level 1 and 2) could benefit community's research in visual grounding and COD [49, 64], though we finally utilize Level 3 annotations for CA-ITR. The distribution of samples across these categories is visualized in

Fig. 2 (d), with larger circles closer to the center indicating higher sample frequencies and brighter colors reflecting a greater average number of annotated words per category. From Fig. 2 (d), it can be observed that the most common object categories in CamoIT correspond to relatively frequent camouflage objects in everyday life, such as lizards and spiders. The distribution of description lengths, as shown in Fig. 2 (c), ranges from 10 to 45 words, with most descriptions clustered around 25 words. We divided the samples into training and test sets using a stratified split with an approximate ratio of 7:3 per class, ensuring that both sets maintained a comparable class distribution. For categories with insufficient samples, we perform random allocation while ensuring a balanced distribution of category counts between the training and test sets. The final test set comprises exactly 3,000 samples.

By using CamoIT, we conducted a comprehensive CA-ITR benchmark study; details are in Section 5.

4. Camouflage-Expert Collaborative Network

CA-ITR faces the fundamental challenge of perceiving and representing camouflaged objects. A natural solution is to enhance the target region within a single encoder by employing learnable prompts or COD masks as soft attention mechanisms. However, when highly similar backgrounds interfere with the object during encoding, methods that merely adjust regional weights are fundamentally inadequate to counteract this interference. Although applying masks before encoding can disrupt the camouflage structure and accentuate the target, this strategy risks degrading image fidelity and introducing semantic misalignment, ultimately weakening cross-modal alignment.

As a solution, we propose a novel camouflage-expert collaborative network (CECNet), which introduces a dual-branch architecture designed to fundamentally prevent feature contamination. One branch processes the original image to preserve global context, while the other independently encodes the isolated camouflaged object to generate a purified feature representation. A novel confidence-conditioned graph attention (C^2GA) module is then proposed to intelligently fuse these complementary streams. As illustrated in Fig. 3, CECNet is built upon a standard VSE framework for global alignment. Below, each component will be discussed in detail.

4.1. Camouflage-Expert Branch

Given a camouflaged image I, it is processed by a COD method to generate the mask $\mathcal{M}$ of the camouflaged object. This mask is then applied to the original image via element-wise multiplication, yielding a refined image representation $I_c = \mathcal{M} \otimes I$ that primarily preserves the camouflaged object. Subsequently, I_c is fed to a camouflaged object encoder based on a visual Transformer [50] to extract multi-level features E^l, where $l = 1, \ldots, 12$ indicates the feature level. Each E^l consists of $(N+1)$ tokens: a CLS token E^l_0 and N image patch tokens E^l_i $(i = 1, \ldots, N)$.

4.2. Global Context Branch

This branch processes the original image I using a standard vision Transformer to generate multi-level features G^l, which effectively capture the holistic scene context. Similarly, G^l comprises $N+1$ tokens: G^l_i, where $i = 0, \ldots, N$. To progressively enhance the representation of camouflaged objects, we incorporate the C^2GA module (detailed in Sec. 4.3) after each block in the encoder. This allows the global branch to leverage features from the camouflage expert branch, resulting in a more effective global representation. The final feature output of this branch (CLS token) is then used as the ultimate visual representation.

4.3. Confidence-Conditioned Graph Attention

Directly fusing features from the two branches using simple strategies like addition or linear transformation is suboptimal (see Sec. 5.5). The concern is that these approaches mix all features indiscriminately, risking contamination of the expert branch's purified camouflage representation by dominant background features from the global branch.

To address this issue, we propose a Confidence-Conditioned Graph Attention (C^2GA) mechanism, which leverages camouflage confidence scores to separately aggregate foreground and background features, thereby preventing feature contamination and enhancing camouflage representation in a more effective manner. C^2GA is inspired by multi-head attention mechanisms, which have been shown to effectively capture diverse aspects of input through the construction of distinct subspaces. Differently, C^2GA explicitly constructs a foreground-relevant graph and a background-relevant graph, guiding the network to focus on each semantic category within its respective subspace.

Specifically, as shown in Fig. 3, C^2GA takes as input the global branch features G (hereafter, the layer superscript l is omitted for simplicity), the expert branch features E, and the mask $\mathcal{M}$. $\mathcal{M}$ is used to compute a camouflage confidence score for each patch feature by applying max pooling to the corresponding region. Following multi-head attention, C^2GA applies a linear projection to map G into two distinct subspaces, yielding the foreground-oriented representation G^{obj} and the background-oriented representation G^{env}. Similarly, E is projected through the same linear projection to generate E^{obj} and E^{env}. These representations, along with their corresponding camouflage confidence scores, are utilized to construct a foreground-relevant graph $\mathcal{G}^F$ and a background-relevant graph $\mathcal{G}^B$. The node set of $\mathcal{G}^F$ comprises all tokens from G^{obj} and E^{obj}, whereas that of $\mathcal{G}^B$ includes tokens from G^{env} and E^{env}.

Taking $\mathcal{G}^F$ as an example, we employ a confidence score

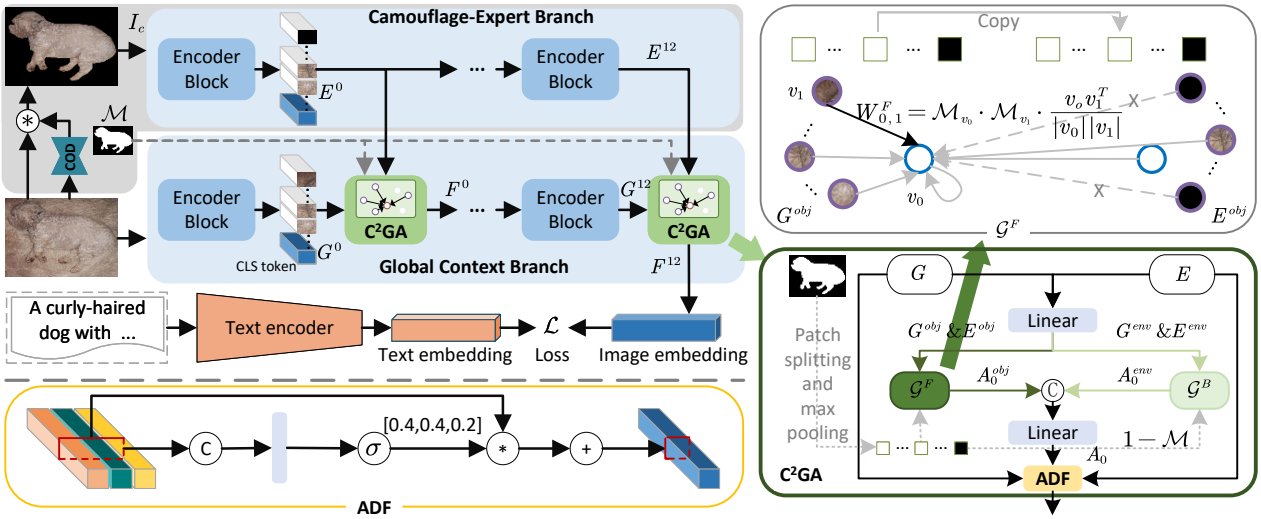

Figure 3. The overall pipeline of the proposed CECNet and C^{2}GA.

to modulate node relationships and suppress edges involving background nodes, thereby constructing a graph focused on foreground-related structures. In this graph, the edge weight between two nodes v_i and v_j is determined by their similarity and camouflage confidence score:

$$W^F_{i,j} = \mathcal{M}_{v_i} \cdot \mathcal{M}_{v_j} \cdot \frac{v_i v_j^T}{|v_i||v_j|}, \quad v_i, v_j \in \mathcal{V}^F \tag{1}$$

where $\mathcal{M}_{v_i} \in [0,1]$ and $\mathcal{M}_{v_j} \in [0,1]$ denote the camouflage confidence scores of the corresponding nodes. Notably, the camouflage confidence scores for the CLS tokens from G^{obj} and E^{obj} are set to 1, as these tokens contain information about the camouflaged objects. The camouflage confidence scores of E^{obj} are copied from the node at the corresponding position of G^{obj}, as camouflaged objects exhibit consistent representations across both original images. Once the graph is constructed, foreground-relevant information is aggregated through $\dot{v}_i = \sum_{v_j \in \mathcal{V}^F} W^F_{i,j} v_j$. Note that, due to the use of the global alignment method, we exclusively aggregate information into G^{obj}_0 (CLS token) to enhance its feature representation, resulting in A^{obj}_0.

Similarly, $\mathcal{G}^B$ is also constructed and performs background-relevant information aggregation, resulting in A^{env}_0. In this case, the confidence score of a node v_j from G^{env} is defined as the background confidence score $(1 - \mathcal{M}_{v_j})$, while the confidence score for G^{env}_0 is set to 1. The confidence scores of all nodes from E^{env} are assigned 0, as they do not contain any background information.

The enhanced foreground (A^{obj}_0) and background (A^{env}_0) representations are then concatenated and projected back to the original feature space, achieving enhanced global features A_0. To ensure feature stability, we employ adaptive gating fusion (ADF in Fig. 3) to integrate enhanced global features with the original global features:

$$F_0 = sum(\sigma(f([A_0, E_0, G_0])) \cdot [A_0, E_0, G_0]), \tag{2}$$

where $f([\cdot])$ utilizes a linear map to adaptively learn the fusion weights for each feature across all channel components, as shown in Fig. 3, and σ is the sigmoid function. Similarly, by using the global alignment method, we update only the CLS token in the global context branch, leaving the other patch tokens unchanged.

4.4. Network Optimization

The visual features extracted from CECNet are aligned with their textual counterparts. Our baseline is built by employing the standard Transformer [50] for text encoding, and the InfoNCE loss [57] for global cross-modal alignment, i.e. $\mathcal{L} = \mathcal{L}_{T2I} + \mathcal{L}_{I2T}$, where $\mathcal{L}_{T2I}$ is text-to-image loss:

$$\mathcal{L}_{T2I} = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{e^{(T_i V_i^\top/\tau)}}{\sum_{j=1}^{n} e^{(T_i V_j^\top/\tau)}}, \tag{3}$$

and $\mathcal{L}_{I2T}$ (image-to-text loss) is defined similarly. Here, V_i and T_i denote the visual and textual features of the i-th sample in a batch, n is the batch size, and the temperature parameter τ is set to 0.05. Notably, as CA-ITR is a very challenging cross-modal alignment task, facing challenges such as bridging the representation gap (where objects are visually similar to the background yet semantically distinct), exploring more elaborate text-side encoding or loss strategies would be a promising direction for future research.

5. Experiments and Results

5.1. Datasets and Protocols

We conducted a comprehensive benchmark for CA-ITR, evaluating both existing SOTA retrieval methods and our proposed CECNet. Our benchmark is conducted on the proposed CamoIT dataset, and two widely recognized benchmark datasets, *i.e.*, the test set of Flickr30K [63] and MS-COCO [36], are also included for reference. Following pre-

Table 1. Quantitative results ($R@K$, %) of models trained on MS-COCO [36] (left) and Flickr30K [63] (right). I2T means image-to-text retrieval, while T2I denotes text-to-image retrieval. "FG" and "BU" mean whether the model employs local alignment or utilizes the BUTD framework [2]. "N/A" means that the result is unavailable. The highest scores are marked in **bold**. "Pub." denotes the publication year.

Models	Pub.	FG	BU	MS-COCO [36]				CamoIT				Flickr30K [63]				CamoIT			
				I2T		T2I		I2T		T2I		I2T		T2I		I2T		T2I	
				$R@1$	$R@10$	$R@1$	$R@10$	$R@1$	$R@10$	$R@1$	$R@10$	$R@1$	$R@10$	$R@1$	$R@10$	$R@1$	$R@10$	$R@1$	$R@10$
CFM [60]	22	✗	✓	59.6	92.9	42.7	83.4	10.7	26.0	10.7	26.7	82.7	98.2	62.6	92.0	5.4	16.5	6.5	18.9
HREM [18]	23	✗	✓	**62.5**	**93.4**	44.0	83.5	11.3	28.1	10.5	27.1	**84.8**	97.9	62.4	92.2	6.7	19.2	5.7	17.9
CHAN [46]	23	✓	✓	59.8	93.3	44.9	84.2	10.9	29.7	13.2	30.1	80.6	97.8	63.9	92.6	6.9	20.0	7.8	21.7
DBL [7]	24	✓	✓	57.5	91.6	41.3	80.9	7.1	20.0	8.7	23.5	78.2	97.4	58.7	89.5	4.5	14.2	3.7	13.6
CUSA [22]	24	✗	✗	57.4	90.1	44.3	82.0	**15.1**	**37.0**	**13.5**	**35.7**	81.0	97.9	66.6	94.0	**12.6**	**34.0**	**10.5**	**27.6**
LAPS [19]	24	✓	✗	56.1	91.2	43.9	83.4	11.8	29.8	10.6	27.3	75.8	97.2	62.5	92.9	5.9	17.2	4.9	15.7
AVSE [40]	25	✗	✗	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	75.9	97.8	62.1	93.0	5.7	16.0	4.8	14.8
D2S-VSE [39]	25	✗	✗	60.1	92.5	**46.3**	**85.2**	13.3	33.1	12.7	30.9	83.1	**98.3**	**68.5**	**95.0**	7.5	20.9	6.5	18.6

vious methods [7, 19, 22], we assess the retrieval performance using recall at K ($R@K$), defined as the proportion of queries for which the correct instance is retrieved among the K nearest neighbors to the query. Following MS-COCO and Flickr30K, we use image-level descriptions (Level 3) in CamoIT. For a comprehensive evaluation, we use the full test sets: the full 5K set for MS-COCO, the full 3K set for CamoIT, and the full 1K set for Flickr30K.

5.2. Implementation Details

Our CECNet is built upon CLIP (ViT-B/32) [50], with the text encoder and the global context branch directly adopted from the corresponding components. The camouflage-expert encoder *shares* parameters with the encoder in the global context branch, thereby ensuring consistent feature representations across both pathways. ZoomNeXt [48] is selected as the COD expert model. A two-stage training strategy is employed to optimize the network: C^2GA modules are trained initially while maintaining the rest as CLIP weights, followed by end-to-end fine-tuning of the entire CECNet (excluding the COD model). The overall training process is accelerated using an NVIDIA RTX 4090 GPU. Following the settings in previous works [18, 50], we use a batch size of 128 and an image size of 224. CECNet is optimized using the Adam optimizer, with the learning rate initially set to 1e-4 in the first stage and subsequently reduced to 1e-5 in the second stage.

The benchmark includes eight SOTA open-source retrieval methods, comprising five global alignment approaches (*i.e.*, CFM [60], HREM [18], CUSA [22], D2S-VSE [39], and AVSE [40]) and three local alignment approaches (*i.e.*, CHAN [46], DBL [7], LAPS [19]). Among these methods, CFM, HREM, CHAN, and DBL adopt the SCAN [30] architecture and utilize the frozen bottom-up attention (BUTD) [2] for image encoding, whereas the remaining approaches employ trainable visual encoders. Note that, since only CUSA provides a CLIP-based implementation along with weights, we evaluate its CLIP-based variant to enable a direct comparison with our method. For the other methods (*i.e.*, LAPS, D2S-VSE, AVSE), we evaluate their implementation versions based on the ViT-Base-224 [8] (pre-trained on ImageNet-21K [6]). To comprehensively evaluate the performance of the aforementioned methods on CA-ITR, we retrain them on CamoIT.

5.3. Benchmark on CA-ITR

Experimental results are summarized in Table 1 and Table 2. As shown in Table 1, methods trained on MS-COCO achieve higher performance on CamoIT than those trained on Flickr30K, likely due to the larger scale of MS-COCO (approximately 4 times that of Flickr30K). Among these methods, CUSA demonstrates superior generalization on CamoIT due to its foundation on CLIP. However, the performance of all methods, as measured by $R@1$, consistently lies under 15%, significantly underperforming relative to their results on MS-COCO and Flickr30K. The results presented in Table 2 demonstrate that retraining these models on CamoIT leads to significant performance improvements. Nevertheless, even after retraining, all methods underperform compared to their results in conventional retrieval, indicating that *CA-ITR is not simply a domain adaptation task, but one that entails multiple unique challenges.*

It is noteworthy that following retraining, CUSA exhibited only marginal improvement, whereas D2S-VSE demonstrated the most substantial performance gain. The limited performance improvement of CUSA may stem from its use of a frozen image encoder [1] to correct potential noise in image-text correspondence labels. However, as the encoder is pre-trained on daily scenes, it fails to provide meaningful features for camouflaged scenes, undermining this objective. In contrast, D2S-VSE emphasizes the key role of information capacity in cross-modal retrieval and introduces a dense text distillation strategy to enhance the information density of sparse text representations. Its significant performance gain demonstrates the crucial role of information capacity in CA-ITR and, more fundamentally, *reveals the high complexity of image contents (as can be observed in Fig. 1), emphasizing the need for a more granular understanding of CA-ITR* through richer textual descriptions to enable accurate modeling.

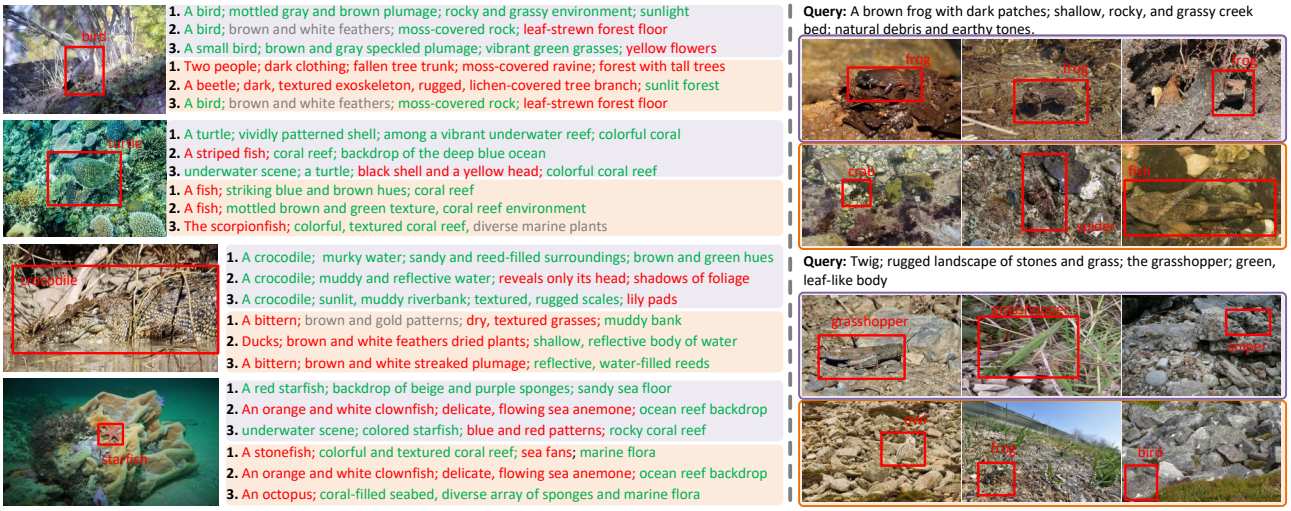

Figure 4. Qualitative results of CECNet (top, light purple) and CLIP (bottom, light orange) on sentence retrieval (left) and image retrieval (right). For each query, we present the top three relevant cross-modal instances. To enhance readability, we retain only the essential components of the sentence. The accurate, inaccurate, and ambiguous portions of the sentences in the search results are highlighted in green, red, and gray, respectively. Camouflaged objects are marked with red bounding boxes.

Table 2. Quantitative results ($R@K$, %) on CA-ITR. All models are *retrained* on CamoIT. The highest scores are marked in **bold**.

Models	I2T			T2I		
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$
CFM [60]	30.8	59.9	70.3	28.9	57.2	68.3
HREM [18]	34.3	62.7	74.0	31.5	59.9	71.4
CUSA [22]	23.9	53.5	66.7	23.5	51.3	64.3
LAPS [19]	27.8	62.0	73.9	28.2	58.0	70.7
D2S-VSE [39]	37.1	68.4	79.5	35.5	67.5	78.4
AVSE [40]	28.1	59.7	72.2	26.1	56.3	69.7
CLIP [50]	41.3	69.2	79.0	41.1	67.7	78.4
D2S-VSE + CEC	39.0	69.4	81.0	37.0	68.3	79.7
AVSE + CEC	29.9	60.7	73.6	28.5	59.2	71.0
CECNet (Ours)	**45.8**	**74.5**	**83.5**	**44.6**	**73.9**	**83.1**

More interestingly, methods based on the frozen BUTD framework (CFM, HREM) outperform several trainable ViT-based approaches (CUSA, LAPS, AVSE) on CamoIT. This performance advantage can be attributed to BUTD's object-level representation paradigm, which decomposes an image into explicit object regions and encodes them accordingly. During object proposal generation, BUTD may capture camouflaged objects—even if only partially—particularly when the camouflage is not highly effective, thereby alleviating some of the perceptual challenges associated with such objects. In contrast, ViT relies on holistic features that are entangled with background context, making it more difficult to isolate object-specific signals from uniform or cluttered surroundings. This finding not only demonstrates *the distinct challenges associated with camouflaged object perception in CA-ITR*, but also substantiates the design rationale behind CECNet's dual-branch architecture and its explicit modeling of camouflaged objects. However, BUTD struggles to use background information [42], which actually highlights the rationale for CECNet's integration of the global context branch.

Overall, the benchmark results demonstrate that CA-ITR is not merely a domain transfer task but one that presents several distinct challenges, including the difficulty of perceiving camouflaged objects, complexity of image contents, and necessity for fine-grained understanding.

5.4. Comparisons with State-of-the-Arts

The quantitative results of CECNet and fine-tuned CLIP are presented in Table 2. On the benchmark, our CECNet achieves an average R@1 improvement of 8.9% over the strongest general retrieval model, D2S-VSE. This superiority is consistent across different models: CECNet advances the fine-tuned CLIP by an average of 4% and notably surpasses another CLIP-based model, CUSA, by a large margin of 21.5%. Also, we further apply the camouflage-expert collaborative scheme to D2S-VSE and AVSE ("D2S-VSE + CEC" and "AVSE + CEC" in Table 2), resulting in performance enhancements. The consistent and significant gains demonstrate that CECNet effectively tackles the specific challenges of CA-ITR. As a pioneering exploration in CA-ITR, the primary contribution of CECNet lies in establishing a foundational baseline and validating the feasibility of the task. The achieved performance, which mirrors the early development phase of traditional ITR [9, 23, 58], exemplified by 38.1% on $R@1$ on Flickr30K, should be viewed as a promising starting point that enables and motivates future research in this challenging domain.

We further present qualitative results of CECNet and CLIP on both sentence retrieval and image retrieval tasks, as shown in Fig. 4. The results demonstrate that CECNet achieves retrieval outcomes with higher semantic relevance to the query, especially in cross-modal scenarios where the primary objectives closely align with the query. In contrast,

Table 3. Ablation study of $R@K$ (%) on dual-branch visual encoder and C^{2}GA. The highest scores are marked in **bold**.

Settings	CamoIT					
	I2T			T2I		
	R@1	R@5	R@10	R@1	R@5	R@10
Baseline	41.3	69.2	79.0	41.1	67.7	78.4
A1	28.5	58.3	70.9	30.0	58.0	70.4
A2	42.1	69.9	79.2	41.2	69.5	79.0
A3	41.8	69.2	79.3	42.2	68.9	79.5
B1	42.5	71.7	80.7	42.9	71.1	80.3
B2	42.4	70.4	80.7	41.7	70.2	79.9
B3	42.9	70.7	80.6	42.3	68.9	78.8
CECNet	**45.8**	**74.5**	**83.5**	**44.6**	**73.9**	**83.1**

although CLIP retrieves results that are strongly tied to the background semantics of the query, the primary objectives in these results frequently mismatch the queried content. The limitations of CLIP highlight the unique challenges in camouflaged object perception within CA-ITR, whereas CECNet achieves superior performance by enhancing the perception and understanding of camouflaged objects.

5.5. Ablation Study

Rationality Behind the Dual-Branch Architecture. We conducted comprehensive ablation studies to validate the necessity of dual-branch encoder. As summarized in Table 3, using the standard CLIP model as the baseline ("Baseline"), We evaluated CECNet against several variants specifically designed to enhance attention toward camouflaged object regions within a single encoder. Specifically, "A1" applies a mask to adaptively modulate the input image via convolutional operations before encoding; "A2" integrates the mask with image features through convolutional fusion; "A3" introduces a trainable prompt to learn feature representations of camouflaged objects, computing the similarity between the prompt and image features to generate a class activation mapping (CAM)-like attention map, which is supervised by the ground-truth segmentation mask during training. Results show that, apart from "A1", which causes performance degradation, all other schemes achieve performance improvements, with CECNet demonstrating the most substantial enhancement. These results are consistent with the analyses presented in Section 4, thereby confirming the effectiveness and necessity of the dual-branch encoder, and further supporting the rationality of its architectural design as discussed in Section 5.3.

Effects of C^{2}GA. We compared C^{2}GA against several fusion strategies, including a simple addition, linear fusion (via channel concatenation followed by a linear layer), and a vanilla graph attention mechanism, denoted as "B1", "B2", and "B3", respectively. As summarized in Table 3, both alternatives yield performance gains; however, their improvements are limited relative to C^{2}GA. The alternative solutions improve performance by leveraging camouflage object information. However, they indiscriminately integrate all features, potentially allowing dominant background features to corrupt the expert branch's refined camouflage representation. In contrast, C^{2}GA selectively integrates camouflage object information into the global representation, preventing feature contamination and enhancing the camouflaged object's representation.

Table 4. Retrieval performance of CECNet and segmentation performance of corresponding COD models. "White" denotes the replacement of COD detection results with a white image. Segmentation performance is measured by S_α and MAE (see [16]).

COD Models	Retrieval				Segmentation	
	I2T		T2I		S_α	MAE
	R@1	R@5	R@1	R@5		
White	41.6	69.6	41.3	69.4	0.443	0.120
SINet [11]	42.3	70.4	42.1	69.9	0.808	0.049
SINet-v2 [14]	43.3	72.7	43.1	71.7	0.843	0.037
ZoomNet [47]	44.0	71.6	42.9	70.5	0.851	0.033
ZoomNeXt [48]	**45.8**	**74.5**	**44.6**	**73.9**	**0.906**	**0.021**

Influence of Different COD Models. Our method employs a COD model as a dedicated expert to provide knowledge for camouflaged scene understanding. To evaluate how this expert guidance influences performance, we integrate four COD backbones: SINet [11], SINet-v2 [14], ZoomNet [47], and ZoomNeXt [48]. We establish performance bounds using all-white images ("White") to simulate scenarios where object perception is unavailable. All COD models are retrained on the CamoIT training set to prevent potential information leakage from prior training data appearing in the test set. We follow each method by initializing training directly from pre-trained weights (e.g., ImageNet [6]). As shown in Table 4, CECNet effectively translates improvements in expert capability into enhanced overall performance, which validate CECNet's ability to leverage expert knowledge. *By the way, our validation is the first to verify that incorporating COD models into the image-text retrieval paradigm can truly boost the retrieval accuracy.*

6. Conclusion

This paper is the first to study cross-modal retrieval in camouflaged scenarios, formally defining the task of camouflage-aware image-text retrieval (CA-ITR). We construct a dedicated camouflage image-text retrieval dataset (CamoIT) as support, and the benchmark conducted on CamoIT reveals the underlying challenges of CA-ITR. To address CA-ITR, we propose a camouflage-expert collaborative network (CECNet), which incorporates a COD expert and a novel confidence-conditioned graph attention (C²GA) mechanism into a dual-branch visual encoder to enhance camouflaged object representations. Extensive experiments show that CECNet achieves encouraging performance over seven representative retrieval models.

Acknowledgments. This work was supported by the NSFC, under No. 62176169, and the Sichuan Science and Technology Program (2025ZNSFSC0469).

References

[1] Xiang An, Jiankang Deng, Kaicheng Yang, Jaiwei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Unicom: Universal and compact representation learning for image retrieval. In *ICLR*, 2023. 6

[2] 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*, pages 6077–6086, 2018. 2, 6

[3] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI*, 41(2):423–443, 2019. 2

[4] Hongbo Bi, Cong Zhang, Kang Wang, Jinghui Tong, and Feng Zheng. Rethinking camouflaged object detection: Models and datasets. *IEEE TCSVT*, 32(9):5708–5724, 2022. 1, 2

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. 3

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 6, 8

[7] Haiwen Diao, Ying Zhang, Shang Gao, Xiang Ruan, and Huchuan Lu. Deep boosting learning: A brand-new cooperative approach for image-text matching. *IEEE TIP*, 33: 3341–3352, 2024. 2, 6

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 6

[9] Aviv Eisenschtat and Lior Wolf. Linking image and text with 2-way nets. In *CVPR*, pages 1855–1865, 2017. 7

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

[11] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, pages 2774–2784, 2020. 2, 3, 8

[12] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *MICCAI*, pages 263–273, 2020. 1

[13] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic COVID-19 lung infection segmentation from CT images. *IEEE Trans. Medical Imaging*, 39(8):2626–2637, 2020.

[14] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE TPAMI*, 44(10): 6024–6042, 2022. 1, 2, 8

[15] Deng-Ping Fan, Ge-Peng Ji, Peng Xu, Ming-Ming Cheng, Christos Sakaridis, and Luc Van Gool. Advances in deep concealed scene understanding. *Vis. Intell.*, 1(1), 2023. 1, 2

[16] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and A. Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, pages 4558–4567, 2017. 8

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

[18] Zheren Fu, Zhendong Mao, Yan Song, and Yongdong Zhang. Learning semantic relationship among instances for image-text matching. In *CVPR*, pages 15159–15168, 2023. 2, 6, 7

[19] Zheren Fu, Lei Zhang, Hou Xia, and Zhendong Mao. Linguistic-aware patch slimming framework for fine-grained cross-modal alignment. In *CVPR*, pages 26297–26306, 2024. 2, 6, 7

[20] Zhentao He, Changqun Xia, Shengye Qiao, and Jia Li. Text-prompt camouflaged instance segmentation with graduated camouflage learning. In *ACM MM*, pages 5584–5593, 2024. 2

[21] Jian Hu, Jiayi Lin, Shaogang Gong, and Weitong Cai. Relax image-specific prompt requirement in SAM: A single generic prompt for segmenting camouflaged objects. In *AAAI*, pages 12511–12518, 2024. 1, 2

[22] Hailang Huang, Zhijie Nie, Ziqiao Wang, and Ziyu Shang. Cross-modal and uni-modal soft-label alignment for image-text retrieval. In *AAAI*, pages 18298–18306, 2024. 2, 6, 7

[23] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal LSTM. In *CVPR*, pages 7254–7262, 2017. 7

[24] Zhou Huang, Hang Dai, Tian-Zhu Xiang, Shuo Wang, Huai-Xin Chen, Jie Qin, and Huan Xiong. Feature shrinkage pyramid for camouflaged object detection with transformers. In *CVPR*, pages 5557–5566, 2023. 1, 2

[25] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 3

[26] Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, and Zhongxuan Luo. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In *CVPR*, pages 4703–4712, 2022. 1, 2

[27] Yao Jiang, Xinyu Yan, Ge-Peng Ji, Keren Fu, Meijun Sun, Huan Xiong, Deng-Ping Fan, and Fahad Shahbaz Khan. Effectiveness assessment of recent large vision-language models. *Vis. Intell.*, 2(1):17, 2024. 2

[28] Nobukatsu Kajiura, Hong Liu, and Shin'ichi Satoh. Improving camouflaged object detection with the uncertainty of pseudo-edge labels. In *ACM Multimedia Asia*, pages 7:1–7:7, 2021. 2

[29] Trung-Nghia Le, Tam V. Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Comput. Vis. Image Underst.*, 184:45–56, 2019. 1, 2, 3

[30] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 212–228, 2018. 2, 6

[31] Aixuan Li, Jing Zhang, Yunqiu Lv, Bowen Liu, Tong Zhang, and Yuchao Dai. Uncertainty-aware joint salient object and camouflaged object detection. In *CVPR*, 2021. 2

[32] Haoran Li, Chun-Mei Feng, Yong Xu, Tao Zhou, Lina Yao, and Xiaojun Chang. Zero-shot camouflaged object detection. *IEEE TIP*, 32:5126–5137, 2023. 1, 2

[33] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, pages 9694–9705, 2021. 2

[34] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Image-text embedding learning via visual and textual semantic reasoning. *IEEE TPAMI*, 45(1):641–656, 2023. 2

[35] Songtao Li and Hao Tang. Multimodal alignment and fusion: A survey. *CoRR*, abs/2411.17040, 2024. 2

[36] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014. 1, 5, 6

[37] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *CVPR*, pages 10918–10927, 2020. 2

[38] Maozhen Liu and Xiaoguang Di. Extraordinary mhnet: Military high-level camouflage object detection network and dataset. *Neurocomputing*, 549:126466, 2023. 1

[39] Yang Liu, Wentao Feng, Zhuoyao Liu, Shudong Huang, and Jiancheng Lv. Aligning information capacity between vision and language via dense-to-sparse feature distillation for image-text matching. In *ICCV*, pages 21679–21688, 2025. 1, 2, 6, 7

[40] Yang Liu, Mengyuan Liu, Shudong Huang, and Jiancheng Lv. Asymmetric visual semantic embedding framework for efficient vision-language alignment. In *AAAI*, pages 5676–5684, 2025. 1, 2, 6, 7

[41] Naisong Luo, Yuwen Pan, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Camouflaged instance segmentation via explicit de-camouflaging. In *CVPR*, pages 17918–17927, 2023. 2

[42] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. Dual-level collaborative transformer for image captioning. In *AAAI*, pages 2286–2293, 2021. 7

[43] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *CVPR*, pages 11591–11601, 2021. 2, 3

[44] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *CVPR*, pages 8772–8781, 2021. 1, 2

[45] Haiyang Mei, Xin Yang, Yunduo Zhou, Ge-Peng Ji, Xiaopeng Wei, and Deng-Ping Fan. Distraction-aware camouflaged object segmentation. *SCIENTIA SINICA Informationis*, 3(7), 2023. 1, 2

[46] Zhengxin Pan, Fangyu Wu, and Bailing Zhang. Fine-grained image-text matching by cross-modal hard aligning network. In *CVPR*, pages 19275–19284, 2023. 2, 6

[47] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *CVPR*, pages 2150–2160, 2022. 2, 8

[48] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoomnext: A unified collaborative pyramid network for camouflaged object detection. *IEEE TPAMI*, 46(12):9205–9220, 2024. 2, 6, 8

[49] Youwei Pang, Xiaoqi Zhao, Jiaming Zuo, Lihe Zhang, and Huchuan Lu. Open-vocabulary camouflaged object segmentation. In *ECCV*, pages 476–495, 2024. 1, 2, 3

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1, 4, 5, 6, 7

[51] Jiacheng Ruan, Wenzhen Yuan, Zehao Lin, Ning Liao, Zhiyu Li, Feiyu Xiong, Ting Liu, and Yuzhuo Fu. Mm-camobj: A comprehensive multimodal dataset for camouflaged object scenarios. In *AAAI*, pages 6740–6748, 2025. 1, 2, 3

[52] Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris. Adversarial representation learning for text-to-image matching. In *ICCV*, pages 5813–5823, 2019. 2

[53] Przemysław Skurowski, Hassan Abdulameer, Jakub Błaszczyk, Tomasz Depta, Adam Kornacki, and Przemysław Kozieł. Animal camouflage analysis: Chameleon database. *Unpublished manuscript*, 2(6):7, 2018. 2, 3

[54] Guolei Sun, Zhaochong An, Yun Liu, Ce Liu, Christos Sakaridis, Deng-Ping Fan, and Luc Van Gool. Indiscernible object counting in underwater scenes. In *CVPR*, pages 13791–13801, 2023. 2

[55] Yujia Sun, Geng Chen, Tao Zhou, Yi Zhang, and Nian Liu. Context-aware cross-level fusion network for camouflaged object detection. In *IJCAI*, pages 1025–1031, 2021. 1, 2

[56] Lv Tang, Peng-Tao Jiang, Zhihao Shen, Hao Zhang, Jin-Wei Chen, and Bo Li. Chain of visual perception: Harnessing multimodal large language models for zero-shot camouflaged object detection. In *ACM MM*, pages 8805–8814, 2024. 1, 2

[57] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. 5

[58] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *ICLR*, 2016. 7

[59] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. CAMP: cross-modal adaptive message passing for text-image retrieval. In *ICCV*, pages 5763–5772, 2019. 2

[60] Hao Wei, Shuhui Wang, Xinzhe Han, Zhe Xue, Bin Ma, Xiaoming Wei, and Xiaolin Wei. Synthesizing counterfactual samples for effective image-text matching. In *ACM MM*, pages 4355–4364, 2022. 2, 6, 7

[61] Zongwei Wu, Danda Pani Paudel, Deng-Ping Fan, Jingjing Wang, Shuo Wang, Cédric Demonceaux, Radu Timofte, and Luc Van Gool. Source-free depth for object pop-out. In *ICCV*, pages 1032–1042, 2023. 1, 2

[62] Jinnan Yan, Trung-Nghia Le, Khanh-Duy Nguyen, Minh-Triet Tran, Thanh-Toan Do, and Tam V. Nguyen. Mirrornet: Bio-inspired camouflaged object segmentation. *IEEE Access*, 9:43290–43300, 2021. 2

[63] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78, 2014. 1, 5, 6

[64] Hong Zhang, Yixuan Lyu, Qian Yu, Hanyang Liu, Huimin Ma, Ding Yuan, and Yifan Yang. Unlocking attributes' contribution to successful camouflage: A combined textual and visual analysis strategy. In *ECCV*, pages 315–331, 2024. 1, 2, 3

[65] Yan Zhang, Zhong Ji, Di Wang, Yanwei Pang, and Xuelong Li. USER: unified semantic enhancement with momentum contrast for image-text retrieval. *IEEE TIP*, 33:595–609, 2024. 2

[66] Pancheng Zhao, Deng-Ping Fan, Shupeng Cheng, Salman H. Khan, Fahad Shahbaz Khan, David A. Clifton, Peng Xu, and Jufeng Yang. Deep learning in concealed dense prediction. *CoRR*, abs/2504.10979, 2025. 1, 3

[67] Dehua Zheng, Xiaochen Zheng, Laurence T. Yang, Yuan Gao, Chenlu Zhu, and Yiheng Ruan. MFFN: multi-view feature fusion network for camouflaged object detection. In *WACV*, pages 6221–6231, 2023. 1, 2

[68] Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *CVPR*, pages 4494–4503, 2022. 1, 2