DATE: Domain Adaptive Product Seeker for E-commerce

Haoyuan Li[1], Hao Jiang[2,*], Tao Jin[1], Mengyan Li[2], Yan Chen[2], Zhijie Lin[1], Yang Zhao[1], Zhou Zhao[1,*]

[1]Zhejiang University, [2]Alibaba Group

{lihaoyuan, jint_zju, linzhijie, awalk, zhaozhou}@zju.edu.cn

{aoshu.jh, yian.lmy, cy270543}@alibaba-inc.com

Abstract

*Product Retrieval (PR) and Grounding (PG), aiming to seek image and object-level products respectively according to a textual query, have attracted great interest recently for better shopping experience. Owing to the lack of relevant datasets, we collect two large-scale benchmark datasets from Taobao Mall and Live domains with about 474k and 101k image-query pairs for PR, and manually annotate the object bounding boxes in each image for PG. As annotating boxes is expensive and time-consuming, we attempt to transfer knowledge from annotated domain to unannotated for PG to achieve un-supervised Domain Adaptation (PG-DA). We propose a **D**omain **A**daptive Produc**t** S**e**eker (**DATE**) framework, regarding PR and PG as Product Seeking problem at different levels, to assist the query **date** the product. Concretely, we first design a semantics-aggregated feature extractor for each modality to obtain concentrated and comprehensive features for following efficient retrieval and fine-grained grounding tasks. Then, we present two cooperative seekers to simultaneously search the image for PR and localize the product for PG. Besides, we devise a domain aligner for PG-DA to alleviate uni-modal marginal and multi-modal conditional distribution shift between source and target domains, and design a pseudo box generator to dynamically select reliable instances and generate bounding boxes for further knowledge transfer. Extensive experiments show that our DATE achieves satisfactory performance in fully-supervised PR, PG and un-supervised PG-DA. Our desensitized datasets will be publicly available here[1].*

1. Introduction

Nowadays, with the rapid development of e-commerce and livestreaming, consumers can enjoy shopping on e-mall or various livestreaming platforms. Although the fact that

---
*Corresponding author.

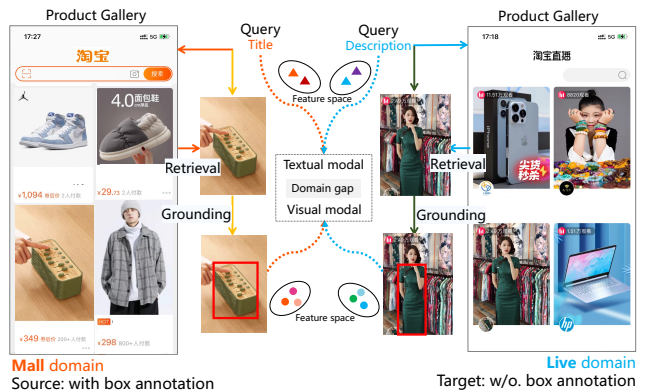[1]https://github.com/Taobao-live/Product-Seeking

Figure 1. Illustration of Product Retrieval (PR) and Grounding (PG) problems on two datasets collected from Taobao Mall and Live. (1) Given a text query (i.e. Chinese title or description of a product), PR is to seek the corresponding image-level product from gallery while PG is to seek the object-level product from an image. (2) We further explore **PG-DA**, which aims to transfer knowledge from the annotated source domain to the unannotated target domain under the influence of multi-modal domain gap to achieve un-supervised PG.

diverse products can be presented and purchased on screen brings us convenience, we are immersed in this miscellaneous product world. Therefore, cross-modal Retrieval [1, 3, 15, 21, 41, 43, 55] for Product (PR), aiming to seek the corresponding image based on a text query, is significant for boosting holistic product search engine and promoting consumers' shopping experience.

Besides, provided that the object-level product can be localized on the target product image or live room image according to a query, it will help consumers focus on the desired product and also benefit the downstream vision-to-vision retrieval. And we name this interesting task as Product Grounding (PG) like Visual Grounding [29, 36, 40, 45, 56]. Generally, PR and PG are seen as two separate tasks, but we consider mining the commonalities of PR and PG and regard them as Product Seeking at image-

level and object-level respectively. And we design a unified architecture to simultaneously solve PR and PG, which is more time-saving and memory-economical than separate methods.

To research the PR and PG with great practical application value, we collect two large-scale benchmark Product Seeking datasets TMPS and TLPS from Taobao Mall and Taobao Live domains with about 474k image-title pairs and 101k frame-description pairs respectively, and the locations of object-level products in images are manually annotated. As annotating bounding box of product is time-consuming and expensive, we explore how to transfer knowledge from an annotated domain to the unannotated one, and achieve un-supervised PG in domain adaptation setting (PG-DA). Thus, we propose the **D**omain **A**daptive Produc**t** **Se**eker (**DATE**) to solve the following aspects of the challenging PR, PG and PG-DA problems.

Firstly, due to the complexity of the mall and live scenarios, discriminative representations of the image and query are prerequisite to accurately localize the object. Considering conventional CNNs are hard to achieve long-distance relation reasoning and full-scale understanding, we utilize and improve the Swin-TF [37] to extract hierarchical and comprehensive features. As large-scale image seeking is demanding for PR, it is vital to ensure seeking inference is of trivial cost. Thus, we inject [REP] token into Swin-TF to absorb the weighted global semantics, and condense them into a single vector, which will be discriminative and concentrated for following efficient image seeking. And we perform the same semantics-aggregated technique for query feature extraction.

Secondly, the capacity of both macroscopic image seeking and microcosmic fine-grained object seeking is necessary for PR and PG. Therefore, we present two cooperative seekers, where image seeker calculates the cosine similarity between visual and textual concentrated features for PR, and object seeker based on cross-modal interaction transformer directly predicts the coordinates of the product by comprehensive features for PG. We validate the reasonableness of such cooperative strategy through experiments.

Thirdly, due to the domain gap between two datasets as Figure 1 shown, applying the model straightway to test on target domain will cause performance degeneration severely for PG-DA. To the best of our knowledge, this is the first work to consider un-supervised Visual Grounding in domain adaptation setting, and most uni-modal DA [8, 34, 38] and multi-modal DA [5,7] methods are not directly applicable in our complicated object seeking. Therefore, we devise a domain aligner based on Maximum Mean Discrepancy to align the domain by minimizing uni-modal marginal distribution and multi-modal conditional distribution divergence between source and target domains, and design a dynamic pseudo bounding box generator to select similar instances in target domain and generate reliable boxes for knowledge transfer.

To summarize, the contributions of this paper are as follows:

- We collect and manually annotate two large-scale benchmark datasets for PR and PG with great practical application value.

- We propose a unified framework with semantics-aggregated feature extractor and cooperative seekers to simultaneously solve fully-supervised PR and PG.

- We explore un-supervised PG in domain adaptation setting and design the multi-modal domain aligner and dynamic box generator to transfer knowledge.

- We conduct extensive experiments which shows that our methods achieve satisfactory performance in fully-supervised PR, PG and un-supervised PG-DA.

2. Related Work

2.1. Visual Retrieval

Given a text query, Visual Retrieval (VR) [1, 3, 21, 41, 43, 55] aims to find the corresponding image/video in a library. The common latent space based methods [1,55] have been proven their effectiveness, which first extract the visual and textual features and map them into a common latent space to directly measure vision-language similarity. Representatively, [16] applies CNN and RNN to encode images and sentences respectively, and learn image-caption matching based on ranking loss. [55] proposes a semantic graph to generate multi-level visual embeddings and aggregate results from the hierarchical levels for the overall cross-modal similarity. Recently, transformer [46] exhibits better performance in Natural Language Processing [12, 20], Computer Vision [4, 13, 25, 26, 28] and multi-modal area [23, 24, 27, 33, 48, 50–52] than previous architecture, especially for global information understanding. Unsuprisingly, there is an increasing effort on repurposing such powerful models [1, 17, 31, 57] for VR. They apply transformer to learn joint multi-mmodal representations and model detailed cross-modal relation, which achieves satisfactory performance.

2.2. Visual Grounding

The paradigm of Visual Grounding (VG) [29, 36, 40, 45], which aims to localize the objects on an image, is similar as Visual Retrieval (VR), they are both to search the best matching part in visual signals according to the text query. Compared to VR, modeling fine-grained internal relations of the image is more significant for VG. In early work, two-stage methods [6, 22, 53] were widely used, which first generate candidate object proposals, then leverage the language
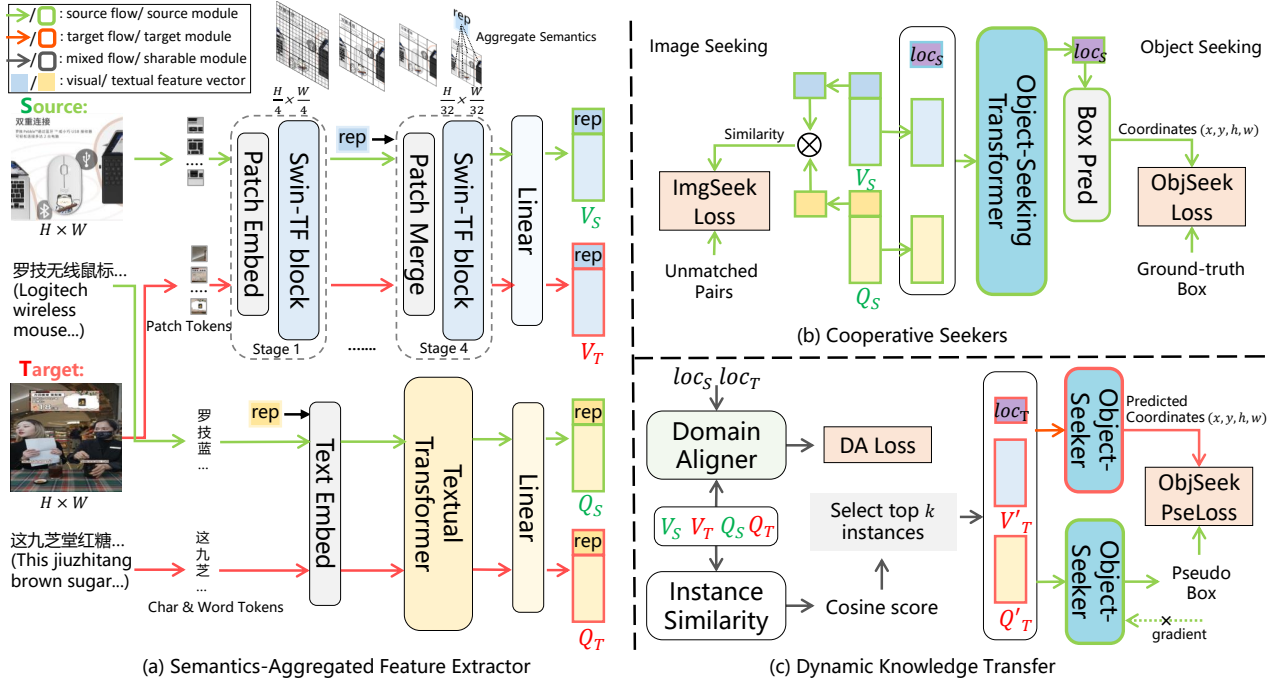
Figure 2. Overview of our DATE. (a) is the feature extractor, applying the semantics-aggregated transformers to obtain image and query features. (b) is the cooperative seekers, calculating the similarity to seek the image for PR and predicting coordinates to seek the object for PG. (c) includes a domain aligner to minimize distribution divergence between source and target domains and a pseudo box generator to select reliable instances and generate bounding boxes for knowledge transfer in PG-DA.

descriptions to select the most relevant object, by leveraging off-the-shelf detectors or proposal generators to ensure recall. However, the computation-intensive proposal generation is time-consuming and also limits the performance of these methods, one-stage methods [32, 49] concentrate on localizing the referred object directly. Concretely, [49] fuses the linguistic feature into visual feature maps and predict bounding box directly in a sliding-window manner. Recently, [10] re-formulates VG as a coordinates regression problem and applies transformer to solve it.

Generally, VR and VG are regarded as two separate problems. In this paper, we mine the commonalities of the two problems and design a unified architecture based on cooperative seeking to efficiently solve VR and VG effectively.

2.3. Un-supervised Domain Adaptation

Unsupervised domain adaptation (UDA) aims to transfer knowledge from the annotated source domain to the unlabelled target domain, and the challenge is how to overcome the influence of domain gap. In uni-modal tasks applications, several UDA techniques have been explored, including aligning the cross-domain feature distribution [18, 34], applying adversarial learning strategy [2, 38] or reconstruction method [8] to obtain domain-invariant features. And
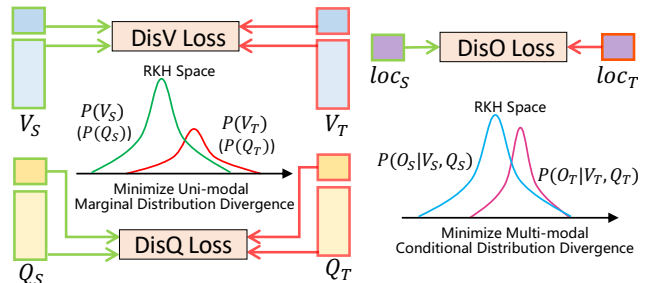

Figure 3. The multi-modal domain aligner.

[9] uses optimal transport to estimate the discrepancy between the two distributions and exploits labels from the source domain. Different from the works described above, our task is cross-modal in nature, which is more challenging due to the heterogeneous gap between different modalities. In multi-modal area, few works have considered UDA, [5] studies the cross-dataset adaptation for visual question answering, [7] studies the video-text retrieval with pseudo-labelling algorithm. To the best of our knowledge, this is the first work to consider un-supervised Visual Grounding in domain adaptation setting.

3. Proposed DATE

3.1. Problem Formulation

In this paper, we explore fully-supervised PR and PG, and un-supervised PG-DA in domain adaptation setting. In the next, we will formulate them.

PR and PG. We collect a fully-annotated dataset $\{V, Q, O\}$, given a textual query Q_i in query set Q, PR and PG aim to seek the image-level product V_{Q_i} from whole image gallery V, and object-level product O_{Q_i} from an matched image V_{Q_i}. The O is the bounding box annotation.

PG-DA. We have access to a fully-annotated source domain $\mathcal{S} = \{V^S, Q^S, O^S\}$, and an unannotated target domain $\mathcal{T} = \{V^T, Q^T\}$ without box annotation O^T. The goal of PG-DA is to transfer the knowledge from $\mathcal{S}$ to $\mathcal{T}$, and seek the object-level product on $\mathcal{T}$.

3.2. Semantics-Aggregated Feature Extractor

As Figure 2(a) shown, for both settings, we share the feature extractor, which can aggregate the global semantics of each modality for image seeking as well as capture comprehensive and context-aware features for object seeking.

Image Stream. Given a RGB image v, we first split it into non-overlapping patches, then we refer to Swin-TF [37] for hierarchical feature extraction. Swin is mainly through the stack of patch merging module and Swin Transformer block to achieve 4-stage encoding, and the resolution is halved at each stage to acquire hierarchical features. The original Swin-TF utilizes average pooling to obtain image representation vector, ignoring the difference in importance of each token for semantics extraction. For promotion, we append a learnable [REP] token in front of visual token sequence during 4th stage, which is involved in the computation of self-attention and absorbs the weighted global image features. After the 4th stage, we can acquire the semantics-aggregated visual feature, and we name this advanced visual encoder as SA-Swin. Next we apply a linear layer to project them into dimension d to obtain $\boldsymbol{V}_{SA} = [V_{rep}, \boldsymbol{V}] \in R^{d \times (1+N_v)}$, where N_v is the number of visual tokens, V_{rep} and $\boldsymbol{V}$ are concentrated and comprehensive features respectively.

Query Stream. Given a textual query q, we first split it into character-level sequence and convert each character into a one-hot vector. After that, we tokenize each one-hot vector into a dense language vector in the embedding layer. Similar to image stream, we append a [REP] token in front of the tokenized query sequence to aggregate the global semantics. Note that the visual and textual [REP] tokens are independent for respective aggregation. Next we take all tokens into a textual transformer to produce the semantics-aggregated query features. Then we project them into the common space dimension d as image stream, to obtain

$\boldsymbol{Q}_{SA} = [Q_{rep}, \boldsymbol{Q}] \in R^{d \times (1+N_q)}$, where N_q is the number of textual tokens.

3.3. Cooperative Seekers

After acquiring common space image feature $\boldsymbol{V}_{SA} = [V_{rep}, \boldsymbol{V}]$ and query feature $\boldsymbol{Q}_{SA} = [Q_{rep}, \boldsymbol{Q}]$, as Figure 2(b) shown, we design two cooperative seekers to search the matched image and localize the object on this image. Next we describe the responsibility of our two seekers.

Image Seekers for PR. The goal of the image seeker is to search the image corresponds to a query. we can directly compute the cosine distance between concentrated features V_{rep} and Q_{rep} to measure the simliarity between image and query, which is time-efficient to search the most similar item and ensures seeking inference is of trivial cost. Given a batch $\mathcal{B}$ with B image-text pairs during training, we calculate the text-to-vision similarity as

$$p^{q2v}(q) = \frac{\exp(l \cdot s(V_{rep}, Q_{rep}) \cdot m^{q2v})}{\sum_{v \in \mathcal{B}} \exp(l \cdot s(V_{rep}, Q_{rep}) \cdot m^{q2v})} \quad (1)$$

$$m^{q2v} = \frac{\exp(\tau \cdot s(V_{rep}, Q_{rep}))}{\sum_{q \in \mathcal{B}} \exp(\tau \cdot s(V_{rep}, Q_{rep}))} \quad (2)$$

where $p^{q2v}(q)$ is text-to-vision probability distribution, l is a learnable logit scaling parameter, $s(\cdot, \cdot)$ denotes cosine similarity, m denotes the prior matrix to refine the similarity distribution following [14], τ represents a temperature hyperparameter. For product retrieval on our datasets, the query (title or description of the product) can be also retrieved by the image, and the vision-to-text similarity is $p^{v2q}(v)$. Then, we treat matching pairs in the batch as positives, and all other pairwise combinations are treated as negatives, thus the image seeking loss can act as

$$\mathcal{L}_{ImgS} = \frac{1}{2} \mathbb{E}_{v,q \sim \mathcal{B}}[H\left(p^{q2v}(q), y^{q2v}(q)\right) \\ + H(p^{v2q}(v), y^{v2q}(v))], \quad (3)$$

where $H(\cdot, \cdot)$ is the cross-entropy formulation, $y(\cdot)$ is the ground-truth binary label that positive and negative pairs are 1 and 0 respectively.

Object Seeker for PG. Different from the image seeker, the ambition of object seeker is to localize the microscopic object-level product on an image, and more sufficient image-query interaction and fine-grained seeking are required. Thus, we leverage comprehensive image and query features $\boldsymbol{V}$ and $\boldsymbol{Q}$ for object seeking. We consider apply a transformer to fuse cross-modal tokens adequately, in order to learn how to localize the product during interaction, we frist append a learnable [LOC] token with visual and textual features as $\boldsymbol{T}_O = [T_{loc}, \boldsymbol{V}, \boldsymbol{Q}] \in R^{d \times (1+N_v+N_q)}$. Then we apply a cross-modal object-seeking transformer to embed $\boldsymbol{T}_O$ into a common space by performing intra- and

inter-modality semantic interaction. Besides, we add learnable modal-type embedding and position embedding to the input of each transformer encoder layer.

We leverage the output state of the [LOC] token f_{loc} from the object-seeking transformer and attach a regression module to it to predict 4-dim box coordinates. Further, to eliminate the influence of scale problem, we normalize the coordinates of the ground-truth box by the scale of the image and perform the object seeking loss as

$$\mathcal{L}_{ObjS} = \|b - \hat{b}\|_1 + G(b, \hat{b}), \tag{4}$$

where $G(\cdot, \cdot)$ is GIoU Loss [44], $b = (x, y, w, h)$ and $\hat{b} = (\hat{x}, \hat{y}, \hat{w}, \hat{h})$ are our prediction the normalized ground-truth box respectively.

So far, PR and PG can be solved simultaneously by the cooperation of two seekers, and our cooperative seeking loss is

$$\mathcal{L}_{coop} = \lambda_{co}\mathcal{L}_{ImgS} + \mathcal{L}_{ObjS}, \tag{5}$$

where $\lambda_{co} \in \mathbb{R}$ are hyperparameters to weigh two losses.

3.4. Dynamic Knowledge Transfer

As Figure 2(a) shown, we design a knowledge transfer method for PG-DA, including a domain aligner to alleviate feature distribution shift and a dynamic pseudo box generator to promote transfer.

Domain Aligner. As Sec 3.3, we extract visual feature $\mathbf{V}_{SA}^S = [V_{rep}^S, \mathbf{V}^S]$ and textual feature $\mathbf{Q}_{SA}^S = [Q_{rep}^S, \mathbf{Q}^S]$ from $\mathcal{S}$ domain, and we acquire $\mathbf{V}_{SA}^T = [V_{rep}^T, \mathbf{V}^T]$ and $\mathbf{Q}_{SA}^T = [Q_{rep}^T, \mathbf{Q}^T]$ from $\mathcal{T}$ domain in the same way. To alleviate the domain discrepancy, we design an alignment approach based on Maximum Mean Discrepancy (MMD), which compares two distributions by embedding each distribution in to Reproducing Kernel Hibert Space (RKHS) $\mathcal{H}$ with a kernel function ϕ. And we utilize multiple Gaussian Radial Basis Function kernels as ϕ. Given two marginal distributions P_{X^S} and P_{X^T} from uni-modal source and target domain respectively, MMD can be expressed as

$$\mathrm{MMD}_{uni}(P_{X^S}, P_{X^T}) = \left\| \mu_{P_{X^S}} - \mu_{P_{X^T}} \right\|_{\mathcal{H}}. \tag{6}$$

In order to compute the inner product of vectors using the kernel function ϕ in RKHS, we square MMD as

$$\mathrm{MMD}_{uni}^2(P_{X^S}, P_{X^T}) = \left\| \mu_{P_{X^S}} - \mu_{P_{X^T}} \right\|_{\mathcal{H}}^2$$
$$= \left\| \frac{1}{n_S^2} \sum_{i=1}^{n_S} \sum_{i'=1}^{n_S} \phi\left(x_i^S, x_{i'}^S\right) - \frac{2}{n_S n_T} \sum_{i=1}^{n_S} \sum_{j=1}^{n_T} \phi\left(x_i^S, x_j^T\right) \right.$$
$$\left. + \frac{1}{n_T^2} \sum_{j=1}^{n_T} \sum_{j'=1}^{n_T} \phi\left(x_j^T, x_{j'}^T\right) \right\|_{\mathcal{H}}. \tag{7}$$

Then, we can minimize the distance between visual feature distributions from different domains through MMD_{uni}^2 as

$$\mathcal{L}_{DisV} = \sum_{v \in \mathcal{B}} [\mathrm{MMD}_{uni}^2(V_{rep}^S, V_{rep}^T)$$
$$+ \mathrm{MMD}_{uni}^2(\mu(\mathbf{V}^S), \mu(\mathbf{V}^T))], \tag{8}$$

where $\mu(\cdot)$ is calculating the mean value of $\mathbf{V}$ on token dimension. In the same way, we compute $\mathcal{L}_{DisQ}$ for textual feature. After that, we can obtain domain-invariant features.

In addition to the discrepancy of uni-modal marginal distribution, we compute the multi-modal conditional distribution divergence to adjust the output distribution for better adaptation, and the form of MMD computation becomes

$$\mathrm{MMD}_{mul}[P(Y^S|X_V^S, X_Q^S), P(Y^T|X_V^T, X_Q^T)]. \tag{9}$$

Concretely, we take out the output of [LOC] token f_{loc}^S and f_{loc}^T in object seeking transformer from two domains and minimize MMD_{mul}^2 to reduce distance of output feature distribution from different domains as

$$\mathcal{L}_{DisO} = \sum_{f_{loc}^S, f_{loc}^T \in \mathcal{B}} \mathrm{MMD}_{mul}^2(f_{loc}^S, f_{loc}^T). \tag{10}$$

The total domain alignment loss function is as follows

$$\mathcal{L}_{DA} = \lambda_{Dv}\mathcal{L}_{DisV} + \lambda_{Dq}\mathcal{L}_{DisQ} + \mathcal{L}_{DisO}, \tag{11}$$

where $\lambda_{Dv}, \lambda_{Dq} \in \mathbb{R}$ are hyperparameters to weigh losses.
Dynamic Pseudo Box Generator. To further transfer the knowledge from $\mathcal{S}$ to $\mathcal{T}$, we attempt to generate pseudo bounding boxes by model on $\mathcal{S}$ to train the model on $\mathcal{T}$. However, it is unlikely that all data can be precisely boxed by source model, which may result in dissatisfactory performance. Therefore, the instances from $\mathcal{T}$ which are close to $\mathcal{S}$ are relatively reliable to be selected. For more precise selection, we compute the instance similarity between two datasets rather than batches. Thus, given the datasets $\{V^S, Q^S\}$ and $\{V^T, Q^T\}$, we calculate the cosine score of features encoded by semantics-aggregated extractor for every pair $\{V^S, V^T\}$ and $\{Q^S, Q^T\}$ in each modality to obtain similarity matrixs M_V and M_Q, and we add them to $M \in [-1, 1]^{N_S \times N_T}$, where N_S and N_T are lengths of source and target datasets respectively. Next, we rank the target instances based on the counts exceed the similarity threshold θ and select the top k percent high-score instances $\{V^{T'}, Q^{T'}\}$. Then, we generate the pseudo box $\widetilde{b}'$ by source object seeker and predict the coordinate b' by target object seeker. Like Eq. 4, we perform the pseudo object seeking loss as

$$\mathcal{L}_{PObjS} = \|b' - \widetilde{b}'\|_1 + G(b', \widetilde{b}'). \tag{12}$$

We compute M each epoch after executing box generation, and the selected instances are dynamically updated.

Table 1. Performance of Product Retrieval (text-to-vision) on our TMPS and TLPS datasets.

| Method | TMPS | | | |
	R@1	R@5	R@10	R@50
Random	0.00	0.04	0.09	0.43
VSEpp	10.23	29.24	34.42	69.73
ViLT	14.39	38.42	50.74	**83.23**
DATE	**16.32**	**40.54**	**51.23**	82.58
	TLPS			
Random	0.03	0.14	0.23	1.59
VSEpp	3.41	15.33	29.12	43.24
ViLT	5.38	19.29	35.95	57.48
DATE	**6.44**	**21.71**	**36.32**	**59.58**

Table 2. Performance of Product Grounding on our TMPS and TLPS datasets.

| Method | TMPS | | TLPS | |
	mIoU	Pr@1	mIoU	Pr@1
Random	29.51	18.22	23.91	10.09
MAttNet	80.71	85.33	62.12	73.24
FAOA	76.24	83.72	61.31	69.13
TransVG	84.52	89.50	67.11	77.93
DATE	**86.67**	**92.12**	**70.24**	**81.43**

With the constant knowledge transfer, more instances can be labeled correctly, and hyper-parameter ratio k will be increased. The total knowledge transfer loss function is as follows

$$\mathcal{L}_{KT} = \mathcal{L}_{DA} + \lambda_{PO}\mathcal{L}_{PObjS}, \qquad (13)$$

where $\lambda_{PO} \in \mathbb{R}$ are hyperparameters to weigh losses.

3.5. Training and Testing

Fully-supervised PR and PG. We perform $\mathcal{L}_{coop}$ for training, and we search the image of product by image-seeker for PR, and directly predict the coordinates of product on the image by object-seeker for PG during testing.

Un-supervised PG-DA. We train the model in three stages. First, we warm up our model under fully-supervised setting on $\mathcal{S}$ domain by $\mathcal{L}_{stage_1} = \mathcal{L}_{ObjS}$. Next, we perform $\mathcal{L}_{stage_2} = \lambda_O\mathcal{L}_{ObjS} + \mathcal{L}_{DA}$ on $\mathcal{S}$ and $\mathcal{T}$ to reduce domain gap. Then, we execute dynamic box genereating and add $\mathcal{L}_{PObjS}$ as $\mathcal{L}_{stage_3} = \lambda_O\mathcal{L}_{ObjS} + \mathcal{L}_{KT}$ to further transfer the knowledge. We test the model on $\mathcal{T}$ domain in the same approach as PG.

4. Experiments

4.1. Our Product Seeking Datasets

We collect two large-scale Product Seeking datasets from Taobao Mall (TMPS) and Taobao Live (TLPS) with

Table 3. Performance of Product Grounding-DA on our datasets. (L→M means we transfer the knowledge from TLPS to TMPS. And F, W, U stand for Fully-, Weakly-, Un-supervised respectively.)

| Method | Mode | TMPS | | TLPS | |
		mIoU	Pr@1	mIoU	Pr@1
Random	-	29.51	18.22	23.91	10.09
ARN	W	70.72	73.32	51.31	53.24
MAF	W	72.52	75.09	54.82	59.04
FAOA	F	76.24	83.72	61.31	69.13
DATE	F	**86.67**	**92.12**	**70.24**	**81.43**
		L→M		**M→L**	
Source-only	U	75.20	83.62	59.64	67.71
MMD-uni	U	76.93	84.87	60.74	69.01
Pseudo-label	U	77.02	86.23	62.87	71.48
DATE	U	**79.92**	**89.35**	**64.86**	**74.75**

about 474k image-title pairs and 101k frame-description pairs respectively. They are first two benchmark e-commerce datasets involving cross-modal grounding. For TMPS, each product item corresponds to a single title, three levels of categories and multiple displayed images with the manually annotated bounding box. For TLPS, we collect frames and descriptions from the livestreamer in live video streams, and annotate the location of described product. Note that the language in our datasets is mainly Chinese. The basic statistics about our datasets is in Appendix A.1. We can see the categories of our datasets are diverse and the number of images are tens of times larger than existing datasets. After the collection, we split each dataset into training/validation/testing sets in a 8:1:1 ratio, and we make sure each product is isolated within one set.

4.2. Evaluation Metrics

Product Grounding. Following [6], we measure the performance by mIoU (mean Intersection over Union) and precision (predicted object is true positive if its IoU with ground-truth box is greater than 0.5).

Product Retrieval. We use standard retrieval metrics (following [1, 57]) to evaluate text-to-vision (t2v) retrieval and vision-to-text (v2t) retrieval. We measure rank-based performance by R@K.

4.3. Performance Comparison and Analysis

To evaluate the effectiveness of DATE, we compare it with various related methods (More details of our methods are reported in Appendix A.2). For each task, we apply untrained model to predict results as *Random* method to perceive the difficulty of tasks.

Product Retrieval. We re-implement these representative cross-modal retrieval methods to compare with our DATE.

Table 4. Ablation study of Product Retrieval and Grounding on TMPS and TLPS datasets.

| Method | TMPS | | | | | | TLPS | | | | | |
| | Grounding | | T2V Retrieval | | | | Grounding | | T2V Retrieval | | | |
	mIoU	Pr@1	R@1	R@5	R@10	R@50	mIoU	Pr@1	R@1	R@5	R@10	R@50
Visual Feature Extractor												
ResNet	80.73	84.13	10.85	29.10	40.82	70.52	64.12	72.25	2.91	13.82	30.94	49.31
DETR	82.29	87.71	12.12	33.52	44.52	74.13	66.13	76.81	4.33	16.39	32.81	54.91
Swin	83.11	89.19	13.21	35.54	46.12	77.59	67.31	78.35	5.01	18.56	34.14	56.25
SA-DETR	84.21	90.03	14.81	36.84	47.21	78.23	68.62	79.11	5.43	19.39	35.81	57.28
SA-Swin (Ours)	**86.67**	**92.12**	**16.32**	**40.54**	**51.23**	**82.58**	**70.24**	**81.43**	**6.44**	**21.71**	**36.32**	**59.58**
Cooperative Seekers												
w/o Rep	83.11	89.19	13.21	35.54	46.12	76.59	67.31	78.35	5.01	18.56	34.14	55.25
w/o ObjS	82.25	87.59	12.85	36.12	45.24	75.23	65.82	75.47	4.93	18.39	35.33	54.12
w/o Rep&ObjS	80.45	85.31	11.78	31.17	43.23	72.23	63.21	71.91	4.13	16.53	31.82	51.10
Full (Ours)	**86.67**	**92.12**	**16.32**	**40.54**	**51.23**	**82.58**	**70.24**	**81.43**	**6.44**	**21.71**	**36.32**	**59.58**

1) *VSEpp* [16], a respectively encoding method based on CNN and RNN.

2) *ViLT* [31], a jointly encoding method based on transformer.

Product Grounding. In addition to cross-modal retrieval baselines above, we re-implement these classic visual grounding baselines to compare with our DATE.

1) *MAttNet* [53], a two-stage model.

2) *FAOA* [49], a one-stage model.

3) *TransVG* [10], a regression-based model under transformer architecture.

The PR and PG results are presented in Table 1 and Table 2 respectively. We can see that (1) the *Random* results in both tasks are pretty low, showing our PR and PG are challenging. (2) The proposed DATE outperforms all the baselines by a large margin, indicating the effectiveness of our method for both PR and PG. (3) Although the performance of *TransVG* and *ViLT* is little behind ours, they are two separate models, and our method under unified architecture is more time-efficient and memory-saving.

Un-supervised Product Grounding-DA. To validate the effectiveness of our DATE in DA setting, we further re-implement these typical weakly-supervised VG baselines for comparison.

1) *ARN* [35], a reconstruction-based model.

2) *MAF* [47], a contrast-based model.

For DA setting, we serve these methods as baselines for comparison.

1) *Source-only*, which applies the model trained on source domain to straightway test on the target dataset.

2) *MMD-uni*, which only utilizes MMD loss to minimize the uni-modal marginal distribution distance for visual and textual feature.

3) *Pseudo-label*, which trains the model on target domain entirely based on the pseudo box labels generated by the model trained on source domain.

The results are presented in Table 3, and we can distill the following observations: (1) our un-supervised DATE outperforms all weakly-supervised methods and fully-supervised methods *FAOA* significantly, demonstrating the knowledge has been transfered to target domain effectively. (2) *Source-only* method degenerates the performance severely due to the huge semantic gap between two domains, and *MMD-uni* only achieves slight improvement as the cross-domain discrepanciy fails to reduced sufficiently. (3) *Pseudo-label* enhances limited performance since a number of bad instances are incorrectly labeled which misleads the model, while our DATE can dynamically select instances and generate reliable bounding boxes for transfer and boosting performance.

4.4. Ablation Study

In this section, we study the effect of different visual feature extractors, text options (in Appendix A.3) and cooperative seeking strategies in Table 4.

Visual Feature Extractor. We compare our *SA-Swin* to *ResNet*, *DETR*, *Swin* and *SA-DETR* methods, where *ResNet*, *DETR* and *Swin* apply ResNet-50 [19], DETR-50 [4] Swin-base [37] to extract image features respectively, and leverage the average pooled feature for PR and feed the flattened last feature map as tokens into object-seeking transformer for PG. And *SA-DETR* executes the same way as the former methods for PG, but injects the semantics-aggregated token from beginning for PR as *SA-Swin* performs. From the results in Table 4, we can find following interesting points:

Figure 4. T-SNE visualization of visual and textual features.

(1) *Swin* surpasses *ResNet* and *DETR*, illustrating better visual features are extracted by hierarchical transformer. (2) *SA-DETR* performs better than *Swin* which has more powerful feature extraction ability during cooperative training, demonstrating our designed semantics-aggregated encoder can extract concentrated and comprehensive features for following cooperative seeking for both PR and PG.

Cooperative Seeking Strategies. We conduct ablative experiments as follows: **w/o Rep**: using the average pooling of two modal features for image seeking (PR) rather than [REP] token. **w/o ObjS**: removing object-seeking transformer, and applying an MLP to fuse visual and textual [REP] token for object seeking; **w/o Rep&ObjS**: using the average pooled feature for both image and object seeking. From Table 4, we observe that the performance decreases sharply after removing [REP] or ObjS. To analyse: (1) more discriminative representation of image and query can be extracted by weighted vector (i.e. [REP] token) than average pooling, confirming the effectiveness of our semantics-aggregated feature extractor. (2) As **w/o Rep** result shown, the performance of object seeking (PG) degenerates although [REP] is not involved in it, which demonstrates such disadvantageous image seeking (PR) approach drags down object seeking (PG) during multi-task learning. (3) Image and object levels seeking falls on the shoulder of [REP] tokens in **w/o ObjS** model, which is detrimental for both levels seeking. The above two points prove the reasonableness of our designed cooperative seeking strategy.

4.5. Feature Visualization

To help prove the validity of our DATE, we visualise visual and textual features by T-SNE for TMPS→TLPS in Figure 4, earned by *Source-only* baseline and our DATE method. We can observe the shift between source and target domains is apparent, meanwhile there are overlaps in two domains, which is reasonable since a few scenes in Taobao Mall and Live are similar. With our proposed method, the discrepancy in feature distribution of two domains becomes narrow significantly, suggesting our method has effectively aligned two domains.

Query: Farmacy木瓜植萃 补水保湿 氨基酸修复肌肤 绿胖子温和洁面乳
(Farmacy Papaya Plant-Extracted Moisturizing Amino-Acid Repairing Skin Green Fat Mild Cleanser)

Figure 5. Qualitative results of Product Retrieval sampled from TMPS dataset (green: correct, red: incorrect).

4.6. Qualitative Analysis

To qualitatively investigate the effectiveness of our DETA, we compare *ViLT* and our DATE for PR as Figure 5 shown. We can find that the image-level product can be sought precisely by our DATE while *ViLT* fails to find the correct image until Rank3. Further, the whole top4 results retrieved by DATE are more relevant to the text query than the results from *ViLT*, which illustrates the multi-modal semantic understanding and interaction are sufficient through our DATE. More examples and qualitative analysis are in Appendix A.4.

5. Conclusion

In this paper, we study the fully-supervised product retrieval (PR) and grounding (PG) and un-supervised PG-DA in domain adaptation setting. For research, we collect and manually annotate two large-scale benchmark datasets TMPS and TLPS for both PR and PG. And we propose a DATE framework with the semantics-aggregated feature extractor, efficient cooperative seekers, multi-modal domain aligner and a pseudo bounding box generator to solve the problems effectively on our datasets. We will release the desensitized datasets to promote investigations on product retrieval, product grounding and multi-modal domain adaptation. In the future, we will consider more specific techniques like Optical Character Recognition (OCR) and Human Object Interaction (HOI) to further improve the performance of PR and PG.

Acknowledgments

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 1, 2, 6

[2] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *NIPS*, 2016. 3

[3] Andrew Brown, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. Smooth-ap: Smoothing the path towards large-scale image retrieval. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, 2020. 1, 2

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, 2020. 2, 7

[5] Wei-Lun Chao, Hexiang Hu, and Fei Sha. Cross-dataset adaptation for visual question answering. In *CVPR*, 2018. 2, 3

[6] Kan Chen, Rama Kovvuri, and Ram Nevatia. Query-guided regression network with context policy for phrase grounding. In *ICCV*, 2017. 2, 6

[7] Qingchao Chen, Yang Liu, and Samuel Albanie. Mind-the-gap! unsupervised domain adaptation for text-video retrieval. In *AAAI*, 2021. 2, 3

[8] Stéphane Clinchant, Gabriela Csurka, and Boris Chidlovskii. A domain adaptation regularization for denoising autoencoders. In *ACL*, 2016. 2, 3

[9] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *TMAPI*, 2016. 3

[10] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. *CoRR*, abs/2104.08541, 2021. 3, 7

[11] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie-Francine Moens. Talk2car: Taking control of your self-driving car. *arXiv preprint arXiv:1909.10838*, 2019. 11

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 2, 11

[13] 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. 2

[14] Cheng Xing etc. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv:2109.04290*, 2021. 4

[15] Xiaopeng Lu etc. Visualsparta: An embarrassingly simple approach to large-scale text-to-image search with weighted bag-of-words. In *ACL*, 2021. 1

[16] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018. 2, 7

[17] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020. 2

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

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7

[20] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: decoding-enhanced bert with disentangled attention. In *ICLR*, 2021. 2

[21] Tuan Hoang, Thanh-Toan Do, Dang-Khoa Le Tan, and Ngai-Man Cheung. Selective deep convolutional features for image retrieval. In *ACM MM*, pages 1600–1608, 2017. 1, 2

[22] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *CVPR*, 2017. 2

[23] Rongjie Huang, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech. In *Advances in Neural Information Processing Systems*. 2

[24] Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2595–2605, 2022. 2

[25] Tao Jin, Siyu Huang, Ming Chen, Yingming Li, and Zhongfei Zhang. Sbat: Video captioning with sparse boundary-aware transformer. *arXiv preprint arXiv:2007.11888*, 2020. 2

[26] Tao Jin, Siyu Huang, Yingming Li, and Zhongfei Zhang. Low-rank hoca: Efficient high-order cross-modal attention for video captioning. *arXiv preprint arXiv:1911.00212*, 2019. 2

[27] Tao Jin, Siyu Huang, Yingming Li, and Zhongfei Zhang. Dual low-rank multimodal fusion. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 377–387, 2020. 2

[28] Tao Jin, Zhou Zhao, Peng Wang, Jun Yu, and Fei Wu. Interaction augmented transformer with decoupled decoding for video captioning. *Neurocomputing*, 492:496–507, 2022. 2

[29] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676, 2017. 1, 2

[30] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 11

[31] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 2, 7

[32] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *CVPR*, 2020. 3

[33] Zhijie Lin, Zhou Zhao, Haoyuan Li, Jinglin Liu, Meng Zhang, Xingshan Zeng, and Xiaofei He. Simullr: Simultaneous lip reading transducer with attention-guided adaptive memory. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1359–1367, 2021. 2

[34] Anan Liu, Shu Xiang, Wenhui Li, Weizhi Nie, and Yuting Su. Cross-domain 3d model retrieval via visual domain adaptation. In *IJCAI*, 2018. 2, 3

[35] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Dechao Meng, and Qingming Huang. Adaptive reconstruction network for weakly supervised referring expression grounding. In *ICCV*, 2019. 7

[36] Yongfei Liu, Bo Wan, Lin Ma, and Xuming He. Relation-aware instance refinement for weakly supervised visual grounding. In *CVPR*, pages 5612–5621, 2021. 1, 2

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030, 2021. 2, 4, 7, 11

[38] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NIPS*, 2018. 2, 3

[39] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 11

[40] Zongshen Mu, Siliang Tang, Jie Tan, Qiang Yu, and Yueting Zhuang. Disentangled motif-aware graph learning for phrase grounding. In *AAAI*, pages 13587–13594, 2021. 1, 2

[41] Tony Ng, Vassileios Balntas, Yurun Tian, and Krystian Mikolajczyk. SOLAR: second-order loss and attention for image retrieval. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, volume 12370, pages 253–270. Springer, 2020. 1, 2

[42] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 11

[43] Jérôme Revaud, Jon Almazán, Rafael S. Rezende, and César Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *ICCV*, pages 5106–5115, 2019. 1, 2

[44] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 5

[45] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, volume 9905, pages 817–834, 2016. 1, 2

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 2

[47] Qinxin Wang, Hao Tan, Sheng Shen, Michael W. Mahoney, and Zhewei Yao. MAF: multimodal alignment framework for weakly-supervised phrase grounding. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *EMNLP*, 2020. 7

[48] Yan Xia, Zhou Zhao, Shangwei Ye, Yang Zhao, Haoyuan Li, and Yi Ren. Video-guided curriculum learning for spoken video grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5191–5200, 2022. 2

[49] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *ICCV*, 2019. 3, 7

[50] Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. Mlslt: Towards multilingual sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5109–5119, 2022. 2

[51] Aoxiong Yin, Zhou Zhao, Jinglin Liu, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. Simulslt: End-to-end simultaneous sign language translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4118–4127, 2021. 2

[52] Aoxiong Yin, Tianyun Zhong, Li Tang, Weike Jin, Tao Jin, and Zhou Zhao. Gloss attention for gloss-free sign language translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, Canada, June 17-23, 2023*. IEEE, 2023. 2

[53] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, 2018. 2, 7

[54] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *ECCV*, 2016. 11

[55] Yawen Zeng, Da Cao, Xiaochi Wei, Meng Liu, Zhou Zhao, and Zheng Qin. Multi-modal relational graph for cross-modal video moment retrieval. In *CVPR*, 2020. 1, 2

[56] Yang Zhao, Chen Zhang, Haifeng Huang, Haoyuan Li, and Zhou Zhao. Towards effective multi-modal interchanges in zero-resource sounding object localization. In *NIPS*, 2022. 1

[57] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, 2020. 2, 6