DUET: Cross-Modal Semantic Grounding for Contrastive Zero-Shot Learning

Zhuo Chen[1, 2, 6], **Yufeng Huang**[3, 6], **Jiaoyan Chen**[4], **Yuxia Geng**[1, 6], **Wen Zhang**[3, 6], **Yin Fang**[1, 6], **Jeff Z. Pan**[5], **Huajun Chen**[1, 2, 6*]

[1]College of Computer Science and Technology, Zhejiang University
[2]Donghai Laboratory, Zhoushan 316021, China
[3]School of Software Technology, Zhejiang University
[4]Department of Computer Science, The University of Manchester
[5]School of Informatics, The University of Edinburgh
[6]Alibaba-Zhejiang University Joint Institute of Frontier Technologies
{zhuo.chen, huangyufeng, gengyx, wenzhang2015, fangyin, huajunsir}@zju.edu.cn,
jiaoyan.chen@manchester.ac.uk, j.z.pan@ed.ac.uk

Abstract

Zero-shot learning (ZSL) aims to predict unseen classes whose samples have never appeared during training. As annotations for class-level visual characteristics, attributes are widely used semantic information for zero-shot image classification. However, the current methods often fail to discriminate those subtle visual distinctions between images due to not only the lack of fine-grained annotations, but also the issues of attribute imbalance and co-occurrence. In this paper, we present a transformer-based end-to-end ZSL method named DUET, which integrates latent semantic knowledge from the pre-trained language models (PLMs) via a self-supervised multi-modal learning paradigm. Specifically, we (1) developed a cross-modal semantic grounding network to investigate the model's capability of disentangling semantic attributes from the images; (2) applied an attribute-level contrastive learning strategy to further enhance the model's discrimination on fine-grained visual characteristics against the attribute co-occurrence and imbalance; (3) proposed a multi-task learning policy for considering multi-model objectives. We find that DUET can achieve state-of-the-art performance on three standard ZSL benchmarks and a knowledge graph equipped ZSL benchmark, and that its components are effective and its predictions are interpretable.

Introduction

Zero-shot learning (ZSL) aims to mimic human's inference ability to learn novel concepts based on prior experience without seeing them beforehand. Early embedding-based ZSL methods project the input into a common vector space where the unseen class prediction can be implemented by searching the nearest class. Generative ZSL methods create synthetic data via the side information of unseen classes, which transforms ZSL into a standard supervised learning problem with less bias toward seen or unseen classes.

As annotations for image visual characteristics, attributes are among the most popular semantic information for ZSL. However, the attributes in real world are typically not annotated to image regions but to a whole class. Recently, some attention-based ZSL methods (Chen et al. 2022b,a) emerge to distinguish the discriminative regions in image classification under the guidance of attentive attribute information. As pointed out by (Wang et al. 2021), these systems suffer from the *imbalanced attribute distribution* (i.e., some attributes are highly frequent while some are rare), as well as the *attribute co-occurrence* which impacts attributes' discrimination capability. For example, in a zero-shot scene classification dataset SUN (Patterson and Hays

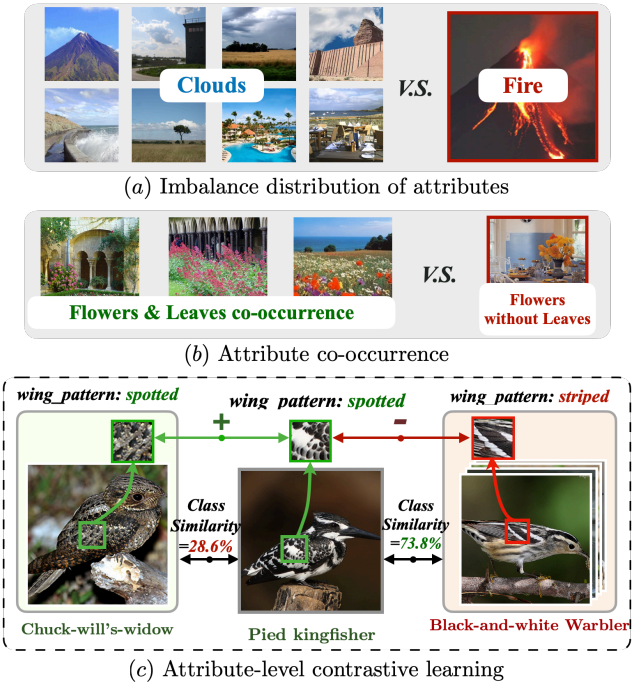


Figure 1: (a) Attribute imbalance. (b) Attribute co-occurrence. (c) Our attribute-level contrastive learning strategy which chooses those distinctive classes as positive references when they are associated with one common attribute (e.g., "spotted") and those similar classes as negative references when they have mutually exclusive attributes (e.g., "striped") toward the same aspect (e.g., "wing_pattern").

*Corresponding Author.

2012), the attributes "trees" and "clouds" are associated with 301 and 318 classes, respectively, while "railroad" and "fire" only appear in 15 and 10 classes. Also, "flowers" appears with "leaves" 39 times, but "flowers" alone only appears 10 times; Such distribution bias may influence the model's judgment on those unseen classes which contain rare attributes or new attribute combinations.

To address these issues, we propose a novel end-to-end ZSL framework named **DUET** (Cross-mo**D**al Semantic Gro**U**nding for Contrastiv**E** Zero-sho**T** Learning). Unlike previous ZSL methods in Figure 2(a) that emphasize utilizing more external class knowledge, augmenting data, or developing better vision encoders, we focus on transferring knowledge from PLMs to vision encoder in a self-supervised manner, as shown in Figure 2(b), giving model the ability for *fine-grained semantic grounding* (i.e., the ability for locating relevant visual characteristics in an image given a textual attribute). Specifically, a prompt-based Feature-to-Sequence Transformation (FST) proxy is utilized to transform different types of attributes into a textual sequence, making our model compatible to multiple ZSL tasks with diverse side information. A Cross-modal Semantic Grounding (CSG) network is developed to leverage the semantics in a PLM via a multi-task learning procedure. Moreover, we propose an attribute-level contrastive learning (ACL) mechanism as shown in Figure 1(c), where distinctive classes are selected as positive references when they are associated with one common attribute (e.g., "spotted") and those similar classes as negative references when they have mutually exclusive attributes (e.g., "striped") toward the same aspect (e.g., "wing_pattern") of the image. This mechanism enables the model to distinguish subtle attribute differences between closed images, and find out the overlapped features between different images. The contributions can be summarized as:

- To the best of our knowledge, DUET is the first to investigate PLMs for zero-shot image classification. It includes a novel end-to-end multi-modal learning paradigm.
- A cross-modal semantic grounding network is developed for effective knowledge transfer from the PLM to the vision transformer encoder. An attribute-level contrastive learning mechanism is proposed to address the attribute imbalance and co-occurrence issues, which further enhances the model's ability for distinguishing fine-grained vision characteristics in both seen/unseen images.
- Experiments is conducted on various ZSL benchmarks equipped with attributes and knowledge graphs. Our code is available at https://github.com/zjukg/DUET.

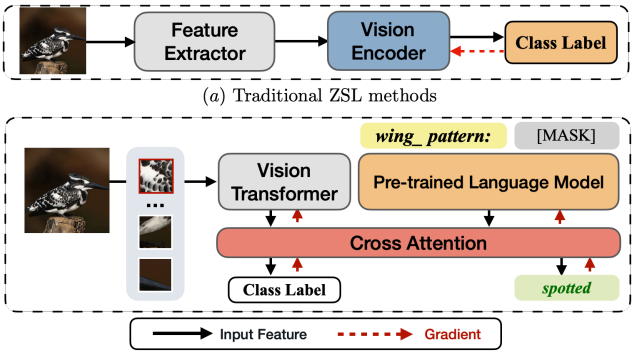
Related Work

Zero-shot Image Classification

The core idea of zero-shot image classification is to transfer semantic knowledge from seen classes to unseen classes based on their semantic information (Chen et al. 2021a,b).

Embedding-based ZSL methods (Frome et al. 2013) intend to build mapping functions toward the images and/or the classes, and whether a class is the label of a sample can be determined by matching their vectors in the same space using similarity metrics. The **Generative** ZSL methods (Chen et al. 2021d,c; Geng et al. 2021a) are introduced to use various generative models (e.g., VAEs and GANs) for creating synthetic data based on semantic features, which can compensate for the shortage of unseen classes and convert ZSL into a supervised classification task. Recently, some **Attention-based** methods begin to explore the discriminative region features guided by attentive semantic information. Specifically, RGEN (Xie et al. 2020) devises the attention technique to construct a region graph for transferring knowledge among different classes. GEM-ZSL (Liu et al. 2021) utilizes gaze embedding to improve the localization of discriminative attributes. MSDN (Chen et al. 2022b) incorporates mutually visual-attribute attention sub-net for semantic distillation, while TransZero (Chen et al. 2022a) further extends MSDN via improving the attention layers by transformers. However, they are still confused by the universal phenomena of the attribute imbalance and co-occurrence (Zhao et al. 2019; Wang et al. 2021), as shown in Figure 1.

In contrast to these methods, we leverage the semantic knowledge in PLMs, and design a cross-modal semantic grounding network to encourage the model to separate those attributes from images. Furthermore, we develop an attribute-level contrastive learning mechanism to address the attribute imbalance and co-occurrence issues, which further enhances the model's discrimination of different independent characteristics in a self-supervised manner.



(a) Traditional ZSL methods

(b) Cross-modal semantic grounding

Figure 2: (a) The paradigm of previous ZSL methods. (b) The paradigm of our method DUET which exploits the semantics of PLMs to augment the transformer-based vision encoder via reconstructing masked attributes (e.g., "spotted") with a cross-model attention mechanism.

Methodology

Let $\mathcal{D}_s = \{(x^s, y^s)|x^s \in \mathcal{X}^s, y^s \in \mathcal{Y}^s\}$ be the training set, where x^s is an image with label y^s attached, and $\mathcal{D}_u = \{(x^u, y^u)|x^u \in \mathcal{X}^u, y^u \in \mathcal{Y}^u\}$ be the unseen dataset, where $\mathcal{Y}^u$ and $\mathcal{Y}^s$ are disjoint. Each label y corresponds to a class $c \in \mathcal{C} = \mathcal{C}^s \cup \mathcal{C}^u$. Specifically, ZSL aims to recognize images of unseen classes ($\mathcal{C}^u$) by transferring learned knowledge from seen classes ($\mathcal{C}^s$) using their side information (e.g., attributes). In this study, we assume attributes annotated to classes are given, where each attribute is some-

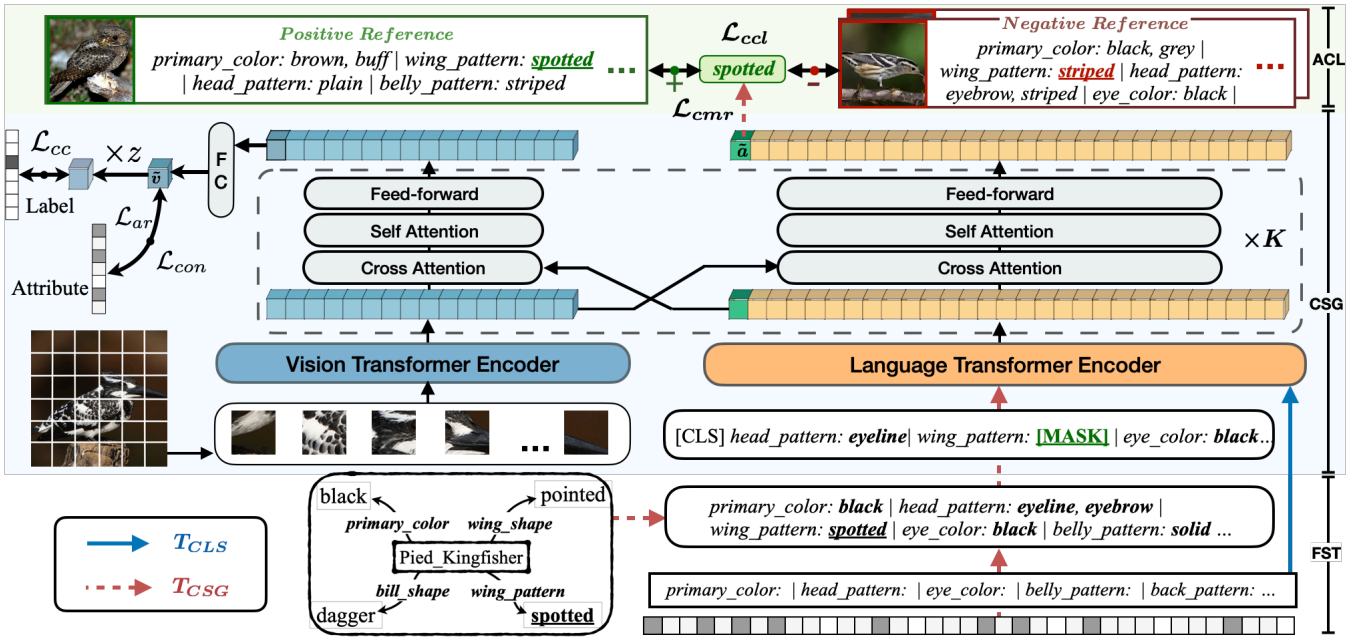


Figure 3: DUET consists of three parts: (1) a Feature-to-sequence transformation (FST) module which unifies attributes of each class into a textual format; (2) a Cross-modal semantic grounding (CSG) module which enables the knowledge transfer from PLM to vision transformer encoder via cross-modal mask reconstruction (CMR); and (3) a Attribute-level contrastive learning (ACL) module which enhances the signal in CSG in a self-supervised manner.

times associated with a real or binary value for indicating its degree. All the attributes of a dataset are denoted as $\mathcal{A} = \{a_1, \ldots, a_{|\mathcal{A}|}\}$, and the attributes of a class c is denoted as $z^c = \left[z_1^c, \ldots, z_{|\mathcal{A}|}^c\right]^\top$.

Feature-to-Sequence Transformation

For ZSL datasets with binary format attributes, we assume $a_i \in \mathcal{A}$ is in the attribute set $\mathcal{A}^c$ of class c if $z_i^c = 1$. Specifically, we propose a prompt-based policy to semi-serialize these discrete attributes to accommodate the sequential inputs of PLMs, inspired by the structured tabular data pre-training (Yin et al. 2020). Concretely, we cluster fine-grained attributes to define k class-specific prompt set $\mathcal{P}$ (i.e., abstract attribute set) where $\mathcal{A} = \mathcal{P}_1 \cup \ldots \cup \mathcal{P}_k$. Then, given a class c, we semi-serialize its attributes with prompt (name) put ahead of each $\mathcal{P}^c$, and take special symbol "|" for prompt set separation. Taking the encoded attribute sentence $\widehat{\mathcal{A}}^c$ for the class "*Otter*" in AWA2 (Xian et al. 2019) dataset as an example:

$$..|\underbrace{\texttt{color}}_{\text{Prompt}}:\ \underbrace{\texttt{brown}}_{\text{Attribute}}|\underbrace{\texttt{haspart}}_{\text{Prompt}}:\ \underbrace{\texttt{tail},\texttt{flippers},\ldots}_{\text{Attributes}}|\ldots \quad (1)$$

Obviously, compared to annotating large-scale fine-grained attributes for each image, it is easier to cluster limited attribute names. Since many ZSL datasets already have their incipient attribute divisions such as SUN (Patterson and Hays 2012), we just need to make little adjustments like removing the repeated prefixes (e.g., "has") and revising some ambiguous $\mathcal{P}$. For knowledge-based ZSL (a.k.a. K-ZSL) datasets such as AWA2-KG in OntoZSL (Geng et al. 2021a), given a triple (c, rel, a), e.g., (*Zebra*, *hasPart*, *Four_legs*), we simply take the relation rel as the prompt of attribute a.

Cross-modal Semantic Grounding

Attribute Phrase Masking (APM). We apply an APM strategy to mask a complete attribute phrase at each step and then urge the model to recover it. We think *discriminative attributes with low frequency within the attribute collection are more important*. Therefore, we sample the target attribute a_t to be masked via a linear weighted random sampling (**LWRS**) strategy: $a_t = LWRS(\mathcal{A})$. Given a class c, the probability $P(a_t = a_j|\mathcal{A}^c)$ for sampling attribute a_j as the a_t^c is:

$$P = \frac{w_j}{\sum_{a_i \in \mathcal{A}^c} w_i}, \ w_j = \frac{1}{\sum_{c' \in \mathcal{C}^s} \mathbb{I}[a_j \in \mathcal{A}^{c'}]}, \quad (2)$$

where $\mathbb{I}[a_j \in \mathcal{A}^{c'}]$ is an indicator function (i.e., it is 1 when $a_j \in \mathcal{A}^{c'}$, otherwise 0).

Since the scale of non-repetitive attribute sentence $\widehat{\mathcal{A}}$ is normally much smaller than $\mathcal{X}^s$ (i.e., $|\mathcal{C}^s| \ll |\mathcal{X}^s|$), we propose random attributes pruning (**RAP**) over the $\mathcal{A}^c$ to remove part of the attributes (except a_t) toward a class within each training step. Specifically, we denote $\mathcal{A}_{rap} = RAP(r_{rap}, \mathcal{A})$ with hyperparameter r_{rap} as the pruning ratio. This will make the model to recover the attribute based on relevant visual information rather than *trickily utilizing attribute co-occurrence*. Thus, the masked attribute sentence constructed based on $\mathcal{A}_{rap}$, denoted as $\widehat{\mathcal{A}}_{rap\setminus t}$, is the input of PLM encoder.

Cross-modal Mask Reconstruction (CMR). We leverage the transformer architecture to encode both the visual features and textual attributes. Specifically, we split an image x (in class c) into patches sequence and feed them into the vision transformer encoder with 1-D position embedding attached. Meawhile, a learnable embedding v_{cls} (marked with `[CLS]`) is prepended whose state at the output serves as the representation of the image. Subsequently, as shown in Figure 3, K cross attention layers are stacked behind the parallel encoders for cross-modal information transfer. Each of them consists of one bi-directional cross-attention block, two self-attention blocks and two feed-forward blocks. A residual connection and layer normalization are added behind each block. The keys and values (Vaswani et al. 2017) from each modality are passed as the input to other modality's multi-headed attention blocks. Let $\tilde{v}$ and $\tilde{a}$ be the output representation of image x and the masked target attribute, respectively. The objective $\mathcal{L}_{cmr}$ for CMR is

$$\mathbb{E}_{x\sim\mathcal{X}^s}\left[-z_{a_t}\sum\nolimits_{i=1}^{Len(w)}\log P(w_i\,|\,\widehat{\mathcal{A}}_{rap\backslash t},x)\right], \quad (3)$$

where w represents the token sequence of target attribute a_t in PLM's vocabulary $\mathcal{V}$, Specifically, z_{a_t} is the expressive degree score for attribute a_t in class c, which *adaptively gives more weights to those highly confident attributes* (i.e., conspicuous characteristics in a class). Moreover, we denote

$$P(w_i\,|\,\widehat{\mathcal{A}}_{rap\backslash t},x)=\exp\left(\tilde{a}_i\cdot e_{w_i}\right)/\sum\nolimits_{w'\in\mathcal{V}}\exp\left(\tilde{a}_i\cdot e_{w'}\right), \quad (4)$$

where e_w refers to the token embedding of w.

Basic ZSL Classification. Following (Chen et al. 2022a; Xu et al. 2020), we present the attribute regression loss $\mathcal{L}_{ar}$ to encourage DUET to accurately map the image representation into corresponding attribute embedding:

$$\mathcal{L}_{ar}=\mathbb{E}_{x\sim\mathcal{X}^s}\|\tilde{v}-z\|_2^2\,, \quad (5)$$

where z is the class-level attribute vector for image x. Meanwhile, we utilize the cross-entropy loss to enforce the image to have the highest compatibility score with its corresponding class semantic vector:

$$\mathcal{L}_{cc}=\mathbb{E}_{x\sim\mathcal{X}^s}[-\log\frac{\exp\left(\tilde{v}\cdot z\right)}{\sum_{\hat{c}\in\mathcal{C}^s}\exp\left(\tilde{v}\cdot z^{\hat{c}}\right)}]\,. \quad (6)$$

To further strengthen DUET's discriminative ability towards different classes with limited samples, we define a class-level supervised contrastive loss:

$$\mathcal{L}_{con}=\mathbb{E}_{x\sim\mathcal{X}^s}[-\log f_\theta(\tilde{v}\,|\,s,x)]\,, \quad (7)$$

where s is the input sentence on language side. Specifically, let $\tilde{v}^+$ be the representation of positive image which has the same class label as $\tilde{v}$, those features with distinct label inside the mini-batch make up a negative samples $\mathcal{N}(\tilde{v})$. We define $f_\theta(\tilde{v}|s,x)$ as:

$$f_\theta=\frac{\exp(D(\tilde{v},\tilde{v}^+)/\tau)}{\exp(D(\tilde{v},\tilde{v}^+)/\tau)+\sum_{\tilde{v}'\in\mathcal{N}(\tilde{v})}\exp(D(\tilde{v},\tilde{v}')/\tau)}\,, \quad (8)$$

where τ is the temperature hyper-parameter. $D(\tilde{v},\tilde{v}^+)$ denotes the cosine similarity between $H(\tilde{v})$ and $H(\tilde{v}^+)$, where $H(.)$ is a non-linear projection head (Chen et al. 2020).

Multi-task Learning. As the core part of cross-modal semantic grounding (CSG), the multi-task learning procedure (Sener and Koltun 2018; Whitehead et al. 2021) forces the model to spread attribute information between the vision side and the language side via a task switching strategy. Namely, DUET conducts CMR at stage T_{CSG} by accessing the visual patches and $\widehat{\mathcal{A}}_{rap}$, and conducts simple image classification task at stage T_{CLS} without seeing the textual attributes. Specifically, the input sequence s_{tmp} at T_{CLS} is fixed as the prompt template to mimic the single-modal testing phase:

$$..|\underbrace{\texttt{color}:}_{\text{Prompt}}|\underbrace{\texttt{haspart}:}_{\text{Prompt}}|\underbrace{\texttt{pattern}:}_{\text{Prompt}}|\underbrace{\texttt{shape}:}_{\text{Prompt}}|... \quad (9)$$

Let L_{CLS}, L_{CSG} be the loss function of basic ZSL classification and CSG, respectively. At each step, we apply the objective L_{CLS} (for T_{CLS}) with probability $1-\rho$, or L_{CSG} (for T_{CSG}) with probability ρ:

$$L_{CLS}=\mathcal{L}_{zsl}+\lambda_{con}\mathcal{L}_{con}\,, \quad (10)$$

$$L_{CSG}=\mathcal{L}_{zsl}+\lambda_{cmr}\mathcal{L}_{cmr}\,, \quad (11)$$

where we denote $\mathcal{L}_{zsl}=\mathcal{L}_{cc}+\lambda_{ar}\mathcal{L}_{ar}$ as the "ZSL loss".

Attribute-level Contrastive Learning

To further strengthen the model's sensitivity on subtle visual differences against the attribute co-occurrences, we introduce an attribute-level contrastive learning (ACL) module with the adaptive loss function:

$$\mathcal{L}_{acl}=\mathbb{E}_{x\sim\mathcal{X}^s}[-Min(z_a,z_{a^+})\log f_\phi(\tilde{a}\,|\,s,x)]\,, \quad (12)$$

where $f_\phi(\tilde{a}\,|\,s,x)$ follows the base formulation of Eq. (8), but there are 3 main differences between f_ϕ and f_θ:

(i) Target Object and Stage. f_ϕ targets at the mean-pooling representation of $\tilde{a}$ on language side of stage T_{CSG}, where the input sentence s is $\widehat{\mathcal{A}}_{rap\backslash t}$. While f_θ targets at the feature ($\tilde{v}$) on vision side, which is applied at stage T_{CLS} with a fixed prompt template (9) as the s_{tmp}.

(ii) Sampling Strategy. For class-level f_θ, we simply pick those images, which share the same class label with original sample as positive, and then define the rest as in-batch negative. While for attribute-level f_ϕ, we design an attribute-based sampling strategy: Given a class c and its target attribute a_t^c, we assume a_t^{c-} as the negative attribute from seen class c^-, and a_t^{c+} as the positive attribute from seen class c^+. We claim the precondition as:

$$c\neq c^+\neq c^-,\ a_t^c=a_t^{c+}\neq a_t^{c-}\,, \quad (13)$$

$$a_t^c\,,\,a_t^{c-}\in\mathcal{P}_t,\ \ a_t^{c-}\notin\mathcal{P}_t^c,\ a_t^c\notin\mathcal{P}_t^{c-}\,, \quad (14)$$

where $\mathcal{P}_t$ is the original class-agnostic prompt set that a_t belongs to, and $\mathcal{P}_t^c$, $\mathcal{P}_t^{c-}$ is the class-specific prompt set in class c, c^-. All c^+, c^- that satisfies this precondition make up the candidate class set $\mathcal{C}^+$ and $\mathcal{C}^-$, respectively.

(iii) Sampling Probability. We employ a heuristic process to let the model select those c^+ whose $\mathcal{A}^{c+}$ are more *inconsistent*, and c^- whose $\mathcal{A}^{c-}$ are more *similar*, compared with $\mathcal{A}^c$. Then, we non-repetitively choose instances (i.e. x^{c-} and x^{c+}) from these classes, and encode them by DUET to get

	Methods	CUB				SUN				AWA2			
		CZSL	GZSL			CZSL	GZSL			CZSL	GZSL		
		T1	U	S	H	T1	U	S	H	T1	U	S	H
†	TF-VAEGAN (ECCV) (2020)	64.9	52.8	64.7	58.1	66.0	45.6	40.7	43.0	72.2	59.8	75.1	66.6
†	Composer (NeurIPS) (2020a)	69.4	56.4	63.8	59.9	62.6	55.1	22.0	31.4	71.5	62.1	77.3	68.8
†	CE-GZSL (CVPR) (2021)	77.5	63.1	66.8	65.3	63.3	48.8	38.6	43.1	70.4	63.1	78.6	70.0
†	GCM-CF (CVPR) (2021)	–	61.0	59.7	60.3	–	47.9	37.8	42.2	–	60.4	75.1	67.0
†	FREE (ICCV) (2021c)	–	55.7	59.9	57.7	–	47.4	37.2	41.7	–	60.4	75.4	67.1
†	HSVA (NeurIPS) (2021d)	62.8	52.7	58.3	55.3	63.8	48.6	39.0	43.3	–	59.3	76.6	66.8
†	AGZSL (ICLR) (2021)	57.2	41.4	49.7	45.2	63.3	29.9	40.2	34.3	73.8	65.1	78.9	71.3
⋆	APN (NeurIPS) (2020)	72.0	65.3	69.3	67.2	61.6	41.9	34.0	37.6	68.4	57.1	72.4	63.9
⋆	DVBE (CVPR) (2020)	–	53.2	60.2	56.5	–	45.0	37.2	40.7	–	63.6	70.8	67.0
⋆	DAZLE (CVPR) (2020b)	66.0	56.7	59.6	58.1	59.4	52.3	24.3	33.2	67.9	60.3	75.7	67.1
⋆	RGEN (ECCV) (2020)	76.1	60.0	73.5	66.1	63.8	44.0	31.7	36.8	73.6	67.1	76.5	71.5
⋆	GEM-ZSL (CVPR) (2021)	77.8	64.8	69.3	67.2	62.8	38.1	35.7	36.9	67.3	64.8	77.5	70.6
⋆	MSDN (CVPR) (2022b)	76.1	68.7	67.5	68.1	65.8	52.2	34.2	41.3	70.1	62.0	74.5	67.7
⋆	TransZero (AAAI) (2022a)	76.8	69.3	68.3	68.8	65.6	52.6	33.4	40.8	70.1	61.3	82.3	70.2
	DUET (Ours)	72.3	62.9	72.8	67.5	64.4	45.7	**45.8**	**45.8**	69.9	63.7	**84.7**	**72.7**

Table 1: Results (%) of our method and the baselines. † and ⋆ indicate generative methods and non-generative methods, respectively. The best results in baselines are marked with underline, and we highlight our results with bold when we achieve new SOTA. For CZSL, results are reported with the top-1 classification accuracy (T1). For GZSL, results are reported in terms of T1 accuracy of unseen (U) and seen (S) classes, together with their harmonic mean (H) where $H = (2 \times S \times U)/(S + U)$.

the final $\tilde{a}^-$ and $\tilde{a}^+$. Note that a_t^{c-} and a_t^{c+} are not masked to accelerate the convergence.

Finally, we stack this pluggable ACL module into CSG:

$$L_{CSG} \longleftarrow L_{CSG} + \mathcal{L}_{acl}\,. \tag{15}$$

Remark 1 *Considering the example in Figure 1, we assume "Pied Kingfisher" as the original bird class (c) with target attribute "spotted" (a_t) in the prompt set "wing pattern" ($\mathcal{P}_t^c$). We are likely to sample "Chuck will's widow" as the positive class c^+ which contains spotted wing pattern, but has a low class similarity (28.6% after normlization) compared with "Pied Kingfisher". Besides, we prefer to sample "Black-and-white Warbler" as the negative class c^- whose wing pattern is striped (not "spotted") but the class characteristic is pretty closed (73.8%) to "Pied Kingfisher".*

Zero-Shot Prediction

We use the cosine metric space for zero-shot recognition with two evaluation settings: conventional ZSL (CZSL) classifies the testing samples with candidate classes from $\mathcal{C}^u$; generalized ZSL (GZSL) extends the candidate classes to $\mathcal{C}^s \cup \mathcal{C}^u$. Specifically, we take the prompt template s_{tmp} together with an test image x as the input. Following (Liu et al. 2021; Chen et al. 2022b), we predict the label c^* via:

$$c^* = \underset{c \in \mathcal{C}^u/\mathcal{C}}{\arg\max} \left(\tilde{v} \cdot z^c\right) - \gamma \mathbb{I}\left[c \in \mathcal{C}^s\right]\,, \tag{16}$$

where $\mathbb{I} = 1$ if c is a seen class and 0 otherwise. γ is the calibration factor tuned on a held-out validation set, and $\mathcal{C}^u/\mathcal{C}$ corresponds to the CZSL/GZSL setting respectively.

Experiments

Datasets

We select three standard attribute equipped ZSL benchmarks **AWA2** (Xian et al. 2019), **CUB** (Welinder et al. 2010), **SUN** (Patterson and Hays 2012) with their splits proposed in (Xian et al. 2019), as well as a knowledge graph (KG) equipped benchmark **AWA2-KG** which has the same split as AWA2 but includes semantic information about hierarchical classes and attributes, for evaluation. In AWA2-KG, we assume that the class c has the attribute a_i when the length of the shortest same-direction relation path between them in KG is h, where h is a hyperparameter. For example, given two triples (*Zebra*, *hasPart*, *Four_legs*) and (*Four_legs*, *subClassOf*, *Leg*), the attribute of class "Zebra" is "Four_leg" when h=1 and "Leg" when h=2. Since we observe that the attribute a become more coarse-grained when they are far away from the class c in KG, we simply define h as 1.

Experimental Settings

Unlike previous ZSL studies which pre-extract the image features using a pre-trained CNN model e.g., ResNet (He et al. 2016), we take as input the raw images and apply vision transformer to interact with the PLM for knowledge transfer. For those coefficients in AWA2, we set λ_{ar} to 0.01, λ_{con} to 0.05, λ_{cmr} to 1, λ_{acl} to 0.01, r_{rap} to 0.5, ρ to 0.4 and γ to 0.8. We report the class-averaged (macro) accuracy as the basic metric, following the current literature (Xu et al. 2020; Chen et al. 2022a).

Overall Results

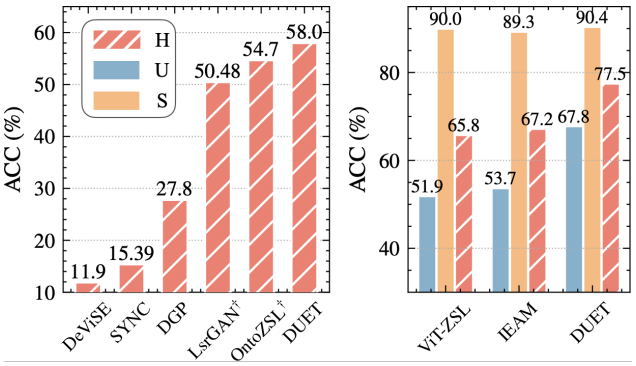
Standard ZSL Datasets. We compare our method with **14** representative or state-of-the-art (SOTA) methods proposed in recent three years. These baselines are divided into two categories: non-generative (Xu et al. 2020; Min et al. 2020; Huynh and Elhamifar 2020b; Xie et al. 2020; Liu et al. 2021; Chen et al. 2022b,a) and generative (Narayan et al. 2020; Huynh and Elhamifar 2020a; Han et al. 2021; Yue et al. 2021; Chen et al. 2021c,d; Chou, Lin, and Liu 2021). All those non-generative methods are attention-based except for DVBE (Min et al. 2020).

SUN contains more than 700 scene classes but each class has only 10-20 images instances, where the attribute imbalance and co-occurrence problem are universal. We find that DUET achieves the best accuracy (45.8%) on H with a large margin (2.5%) compared with those SOTA methods, and surpass MSDN by 4.5% on H, which is the SOTA no-generative methods on SUN. On **AWA2**, DUET gains 1.2% improvements over the SOTA performance, and outperforms the transformer-based method TransZero on all the GZSL metrics. On **CUB**, DUET achieves competitive performance, surpassing all generative methods on H, except for the attention-based methods TransZero and MSDN. We own this to the fact that the prompts in CUB are mostly region-related (e.g., ""breast_color" and "wing_color"), but DUET simply attaches the image patches with sequential 1-dimensional positional embedding as the input, making it hard to capture the fine-grained positional relationship. Instead, TransZero takes 2D center coordinates to construct learnable relative region geometric embeddings for feature augmentation, which gets accurate position representations and helps the model achieve good performance. Notably, when it does not use feature augmentation from relative geometry relationships, the H on CUB dramatically decreases to 66.5% (Chen et al. 2022a).

Moreover, DUET also achieves great performance on seen classes (S) on all three datasets, outperforming all baselines on SUN and AWA2 by at least 5.1% and 2.4% respectively. This proves that DUET well preserves the predictive ability on seen classes in addressing unseen classes.

K-ZSL Dataset. We evaluate on AWA2-KG (Geng et al. 2021a) to validate DUET's flexibility on various ZSL attribute formats. Specifically, we pick the KG from (Geng et al. 2021b) as the knowledge resource, and compare with baselines including DeViSE (Frome et al. 2013), SYNC (Changpinyo et al. 2016), DGP (Kampffmeyer et al. 2019), LsrGAN† (Vyas, Venkateswara, and Panchanathan 2020). We abandon the real-value attributes for fairness, and follow (Geng et al. 2021a, 2022; Chen et al. 2021e) to take the KG embedding (Bordes et al. 2013) for entity class representation toward $\mathcal{L}_{ar}$ and $\mathcal{L}_{cc}$. As shown in Figure 4(a), DUET achieves higher performance among all other methods. In particular, it achieves a 30.2% improvement on metric H compared to the non-generative method DGP.

ViT-based DUET. To get further insights into our model, we report the results of DUET with ViT-base (Dosovitskiy et al. 2021) as the vision encoder. Remarkably, since the released ViT-base is pre-trained on ImageNet-21K which may contain unseen objects, we only select 2 recent ViT-based ZSL methods, ViT-ZSL (Alamri and Dutta 2021) and IEAM-ZSL (Narayan et al. 2020), for comparison. As shown in Figure 4(b), DUET surpasses these two methods by a large margin (14.1% improvement on U and 10.3% improvement on H) and also exceeds our SOTA performance (H) by 4.8%. This supports that our DUET greatly ameliorates the ZSL ability where the original vision transformer is poor. We believe that the performance will be further improved by plugging in a better vision transformer encoder.



(a) K-ZSL (on AWA2-KG) (b) ViT-based (on AWA2)

Figure 4: (a) Results (%) on AWA2-KG from OntoZSL. The attribute values in this dataset are all represent in 0/1 binary form. We marks those generative methods with "†". (b) Results (%) on AWA2 with ViT-base as the vision encoder.

Ablation Studies

Methods	H	△
Only ENC_{vis}	64.1	8.6↓
1) CSG$_{freeze\ ENC_{lan}}$	66.5	6.2↓
2) CSG$_{w/\ only\ Prompt}$	61.7	11.0↓
3) CSG$_{w/o\ Prompt}$	64.9	7.8↓
4) CSG$_{w/o\ LWRS}$	66.9	5.8↓
5) CSG$_{w/o\ RAP}$	67.4	5.3↓
6) CSG$_{w/o\ \mathcal{L}_{con}}$	68.4	4.3↓
7) CSG	69.2	3.5↓
DUET (Full model)	**72.7**	-

Table 2: Results (%) of ablation studies on AWA2 by GZSL. The metric is harmonic mean (H) accuracy. △ indicates the performance drop compared with our full model.

Component Analysis. We evaluate various stripped-down versions of our model to compare the (H) performance gain brought by different components on AWA2. Concretely, we observe that the performance drops sharply when **(1)** freezing the language transformer encoder ENC_{lan}. Although it can reduce the overall learnable parameters, it makes the model harder to understand the special relationship among prompts, textual attributes, and visual features. From the results of taking **(2)** only the prompt and **(3)** only concatenating attribute as sequence input without the prompt, we observe that employing our FST strategy for semi-serializing attributes indeed benefits our model with 4.3% improvement. We also exploit the influence of **(4)** randomly masking attributes, **(5)** not conducting attribute pruning, which leads to 2.3%, 1.8% falls compared with **(7)** applying the full CSG, proofing the necessity of both sampling target attribute with adaptive weight and pruning part of the attribute. Besides, **(6)** abandoning class-level contrastive learning leads to 0.8% decrease. We own this to the fact that contrastive learning can help model learn better visual representations by narrowing the distance within a class in the latent space. Most importantly, our pluggable ACL module further boosts

the performance by 3.5% based on CSG, which illustrates that both of these modules are beneficial.

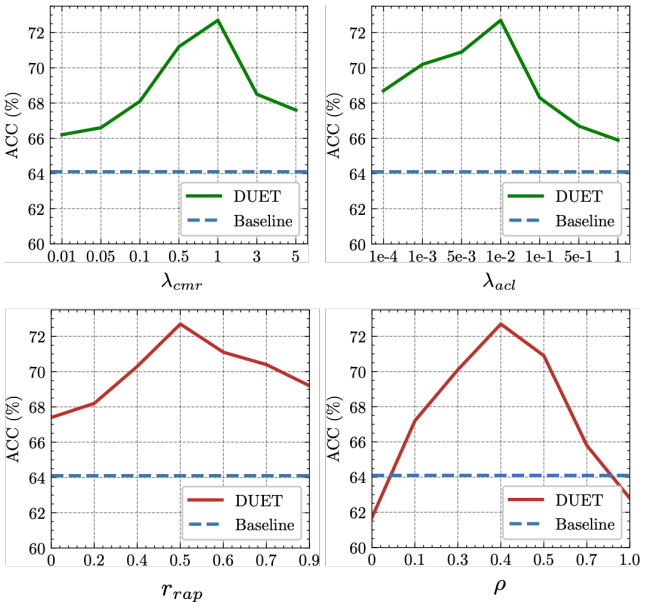


Figure 5: A parameter analysis of coefficients λ_{cmr}, λ_{acl}, r_{rap} (the pruning ratio for $RAP(.)$), ρ (the probability for employing L_{CSG}). Baseline denotes the pure vision trasnformer encoder. H accuracy on AWA2 is reported.

Hyperparameter Analysis. By comparing DUET's performance in Figure 5, we conclude that: *(i)* The performance decrease when λ_{cmr} and λ_{acl} are extreme large, since the weak signal from the self-supervised objectives (i.e., $\mathcal{L}_{cmr}$ and $\mathcal{L}_{acl}$) will gradually overwhelm the signal from supervised class label (i.e., $\mathcal{L}_{ar}$ and $\mathcal{L}_{cc}$). *(ii)* When ρ is close to 1 or 0, the protocol all drops below the baseline. This is because the model turns into a single-modal task without multi-task learning when $\rho = 0$. While when $\rho = 1$, DUET is forced to classify the image with attribute attached throughout the training, leading to model's poor generalization capability at test stage. *(iii)* Furthermore, we try a wide range of r_{rap}, i.e. r_{rap} = {0, 0.2, 0.4, 0.5, 0.6, 0.7, 0.8}, and find that DUET works best when r_{rap} is set to 0.5.

Interpretation

To proofs DUET's capability on disentangling image-specific semantic attributes, we feed an image into the well-trained DUET together with a crafted template: "...|$\widehat{\mathcal{P}}$: $[MASK]$|...", where each prompt name $\widehat{\mathcal{P}}$ is involved with the `[MASK]` followed to recover attribute tokens Figure 6 shows the prediction results from the cases in SUN's testing set. We observe that DUET can successfully retrieve most relevant attributes from the image given a concrete prompt, e.g., the "sport" function from a "basketball arena" and the "natural" light from an open-air "bus depot".

Sometimes, there also exist unreasonable attribute pairs within a class in GT attribute sets. For example, *(i)* "oilrig" has both "still water" and "waves" as its attributes,

Image Case	Prompt	Attribute
oilrig (S) 	transportation function	DUET : boating (46.6%); swimming (34.4%); digging (19.0%)
		GT : boating (52.0%); swimming (12.7%)
	coarse meterial	DUET : metal (39.9%); ocean (37.6%); wire (22.5%)
		GT : ocean (48.5%); metal (43.9%); wire (6.9%)
	specific meterial	DUET : fire (40.8%); waves (37.8%); smoke (21.5%)
		GT : still water (24.3%); waves (16.2%); fire (15.1%)
living room (S) 	environment function	DUET : reading (55.4%); eating (31.1%); working (13.5%)
		GT : reading (46.6%); eating (12.3%); socializing (6.9%)
	specific meterial	DUET : cloth (39.6%); flowers (38.5%); tiles (22.0%)
		GT : carpet (28.8%); cloth (23.3%); flowers (11.0%)
	feeling	DUET : soothing (60.0%); symmetrical (32.9%); cluttered (7.1%)
		GT : soothing (49.3%)
basketball arena (U) 	environment function	DUET : sports (49.2%); playing (32.8%); socializing (18.1%)
		GT : congregating (47.3%); sports (43.5%); playing (11.5%)
	technical function	DUET : competing (40.0%); audience (36.2%); exercise (23.8%)
		GT : competing (52.4%); audience (52.4%); exercise (17.9%)
	surface	DUET : glossy (41.6%); rusty (32.7%); sterile (25.7%)
		GT : glossy (17.9%); dry (15.3%)
bus depot (U) 	transportation function	DUET : driving (43.0%); biking (38.0%); climbing (19.1%)
		GT : driving (41.5%); biking (18.0%)
	coarse meterial	DUET : asphalt (34.3%); metal (33.7%); concrete (32.0%)
		GT : asphalt (38.9%); metal (22.1%); pavement (20.8%)
	light	DUET : natural (57.3%); indoor (33.6%); direct sunny (9.2%)
		GT : natural (54.5%); direct sun/sunny (20.4%)

Figure 6: Attribute prediction for interpretation.

which is *contradictory*; *(ii)* there is no "carpet" in this "living room" image, but it has high confidence. These situations occur when mutually exclusive attributes are independently contained in different images within the same class, since the class-level attribute values in SUN are collected by averaging the binary labels from annotators. In contrast, our DUET could achieve *instance-level semantic grounding*, correctly giving "waves" high score in this "oilrig", and ignoreing "carpet" in this "living room". Besides, the scarce attribute "fire" is confidently predicted in the "oilrig" image, while in "living room", the "flowers" are recovered without "leaves" bound together, which demonstrate the capability of DUET in *addressing the attribute imbalance and attribute co-occurrence issues* shown in Figure 1. Moreover, DUET can ground not only obvious visual attributes (e.g., "fire") but also those abstract properties (e.g., "soothing" for "feeling" and "competting" for "technical function"), which shows its potential capability for knowledge inference.

Conclusion

In this paper, we propose an end-to-end ZSL framework named DUET to address the well known issues of attribute imbalance and co-occurrence in zero-shot image classification. We design a cross-modal semantic grounding network with a novel attribute-level contrastive learning mechanism to enhance the model's discriminative ability towards novel classes, which could well address the issues of attribute imbalance and co-occurrence in zero-shot learning. With extensive ablation studies and the comparison with quite a few state-of-the-art methods on four ZSL benchmarks with real-valued and binary-valued attributes, we demonstrate the effectiveness of DUET as well as its support for interpretation.

Acknowledgements

We want to express gratitude to the anonymous reviewers for their hard work and kind comments. This work is partially funded by NSFCU19B2027/91846204, the EPSRC project ConCur (EP/V050869/1) and the Chang Jiang Scholars Program (J2019032).

References

Alamri, F.; and Dutta, A. 2021. Multi-Head Self-Attention via Vision Transformer for Zero-Shot Learning. *CoRR*, abs/2108.00045.

Bordes, A.; Usunier, N.; García-Durán, A.; Weston, J.; and Yakhnenko, O. 2013. Translating Embeddings for Modeling Multi-relational Data. In *NIPS*, 2787–2795.

Changpinyo, S.; Chao, W.-L.; Gong, B.; and Sha, F. 2016. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5327–5336.

Chen, J.; Geng, Y.; Chen, Z.; Horrocks, I.; Pan, J. Z.; and Chen, H. 2021a. Knowledge-aware Zero-Shot Learning: Survey and Perspective. In *IJCAI*, 4366–4373. ijcai.org.

Chen, J.; Geng, Y.; Chen, Z.; Pan, J. Z.; He, Y.; Zhang, W.; Horrocks, I.; and Chen, H. 2021b. Low-resource Learning with Knowledge Graphs: A Comprehensive Survey. *CoRR*, abs/2112.10006.

Chen, S.; Hong, Z.; Liu, Y.; Xie, G.-S.; Sun, B.; Li, H.; Peng, Q.; Lu, K.; and You, X. 2022a. TransZero: Attribute-guided Transformer for Zero-Shot Learning. In *AAAI*.

Chen, S.; Hong, Z.; Xie, G.-S.; Yang, W.; Peng, Q.; Wang, K.; Zhao, J.; and You, X. 2022b. MSDN: Mutually Semantic Distillation Network for Zero-Shot Learning. In *CVPR*.

Chen, S.; Wang, W.; Xia, B.; Peng, Q.; You, X.; Zheng, F.; and Shao, L. 2021c. FREE: Feature Refinement for Generalized Zero-Shot Learning. In *ICCV*, 122–131.

Chen, S.; Xie, G.-S.; Peng, Q.; Liu, Y.; Sun, B.; Li, H.; You, X.; and Shao, L. 2021d. HSVA: Hierarchical Semantic-Visual Adaptation for Zero-Shot Learning. In *35th Conference on Neural Information Processing Systems (NeurIPS)*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*, volume 119, 1597–1607.

Chen, Z.; Chen, J.; Geng, Y.; Pan, J. Z.; Yuan, Z.; and Chen, H. 2021e. Zero-Shot Visual Question Answering Using Knowledge Graph. In *ISWC*, volume 12922 of *Lecture Notes in Computer Science*, 146–162.

Chou, Y.; Lin, H.; and Liu, T. 2021. Adaptive and Generative Zero-Shot Learning. In *ICLR*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; and Mikolov, T. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *NIPS*, 2121–2129.

Geng, Y.; Chen, J.; Chen, Z.; Pan, J. Z.; Ye, Z.; Yuan, Z.; Jia, Y.; and Chen, H. 2021a. OntoZSL: Ontology-enhanced Zero-shot Learning. In *WWW*, 3325–3336.

Geng, Y.; Chen, J.; Chen, Z.; Pan, J. Z.; Yuan, Z.; and Chen, H. 2021b. K-ZSL: resources for knowledge-driven zero-shot learning. *arXiv preprint arXiv:2106.15047*.

Geng, Y.; Chen, J.; Zhang, W.; Xu, Y.; Chen, Z.; Pan, J. Z.; Huang, Y.; Xiong, F.; and Chen, H. 2022. Disentangled Ontology Embedding for Zero-shot Learning. *CoRR*, abs/2206.03739.

Han, Z.; Fu, Z.; Chen, S.; and Yang, J. 2021. Contrastive Embedding for Generalized Zero-Shot Learning. In *CVPR*, 2371–2381.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778. IEEE Computer Society.

Huynh, D.; and Elhamifar, E. 2020a. Compositional Zero-Shot Learning via Fine-Grained Dense Feature Composition. In *NeurIPS*.

Huynh, D.; and Elhamifar, E. 2020b. Fine-Grained Generalized Zero-Shot Learning via Dense Attribute-Based Attention. In *CVPR*, 4482–4492. Computer Vision Foundation / IEEE.

Kampffmeyer, M.; Chen, Y.; Liang, X.; Wang, H.; Zhang, Y.; and Xing, E. P. 2019. Rethinking Knowledge Graph Propagation for Zero-Shot Learning. In *CVPR*, 11487–11496.

Liu, Y.; Zhou, L.; Bai, X.; Huang, Y.; Gu, L.; Zhou, J.; and Harada, T. 2021. Goal-Oriented Gaze Estimation for Zero-Shot Learning. In *CVPR*, 3794–3803.

Min, S.; Yao, H.; Xie, H.; Wang, C.; Zha, Z.; and Zhang, Y. 2020. Domain-Aware Visual Bias Eliminating for Generalized Zero-Shot Learning. In *CVPR*, 12661–12670.

Narayan, S.; Gupta, A.; Khan, F. S.; Snoek, C. G. M.; and Shao, L. 2020. Latent Embedding Feedback and Discriminative Features for Zero-Shot Classification. In *ECCV (22)*, 479–495.

Patterson, G.; and Hays, J. 2012. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, 2751–2758.

Sener, O.; and Koltun, V. 2018. Multi-Task Learning as Multi-Objective Optimization. In *NeurIPS*, 525–536.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NIPS*, 5998–6008.

Vyas, M. R.; Venkateswara, H.; and Panchanathan, S. 2020. Leveraging seen and unseen semantic relationships for generative zero-shot learning. In *European Conference on Computer Vision*, 70–86. Springer.

Wang, L.; Huang, J.; Li, Y.; Xu, K.; Yang, Z.; and Yu, D. 2021. Improving Weakly Supervised Visual Grounding by Contrastive Knowledge Distillation. In *CVPR*, 14090–14100. Computer Vision Foundation / IEEE.

Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Schroff, F.; Belongie, S. J.; and Perona, P. 2010. Caltech-UCSD Birds 200. *Technical Report CNS-TR-2010-001, Caltech,*.

Whitehead, S.; Wu, H.; Ji, H.; Feris, R.; and Saenko, K. 2021. Separating Skills and Concepts for Novel Visual Question Answering. In *CVPR*, 5632–5641.

Xian, Y.; Lampert, C. H.; Schiele, B.; and Akata, Z. 2019. Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(9): 2251–2265.

Xie, G.; Liu, L.; Zhu, F.; Zhao, F.; Zhang, Z.; Yao, Y.; Qin, J.; and Shao, L. 2020. Region Graph Embedding Network for Zero-Shot Learning. In *ECCV (4)*, 562–580.

Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; and Akata, Z. 2020. Attribute Prototype Network for Zero-Shot Learning. In *NeurIPS*.

Yin, P.; Neubig, G.; Yih, W.; and Riedel, S. 2020. TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data. In *ACL*, 8413–8426.

Yue, Z.; Wang, T.; Sun, Q.; Hua, X.; and Zhang, H. 2021. Counterfactual Zero-Shot and Open-Set Visual Recognition. In *CVPR*, 15404–15414.

Zhao, B.; Fu, Y.; Liang, R.; Wu, J.; Wang, Y.; and Wang, Y. 2019. A Large-Scale Attribute Dataset for Zero-Shot Learning. In *CVPR Workshops*, 398–407. Computer Vision Foundation / IEEE.