Counterfactual-Driven Zero-Shot Classifier Expansion

Xiangyu Wang, Yanze Gao, Changxin Rong, Lyuzhou Chen
Derui Lyu, Xiren Zhou, Taiyu Ban, Huanhuan Chen*
School of Computer Science and Technology
University of Science and Technology of China, 96 Jinzhai Rd
Hefei, 230026, China
sa312@ustc.edu.cn, {gaoyz,rcx,clz31415,drlv}@mail.ustc.edu.cn
zhou0612@ustc.edu.cn, banty@mail.ustc.edu.cn, hchen@ustc.edu.cn

Abstract

Zero-shot classifier expansion aims to adapt existing model to new, unseen classes. It utilizes class attributes or textual descriptions to learn a mapping from the semantic space to the classifier's weight space, without requiring new visual training data. However, the learning process for this mapping relies solely on correlating semantic patterns with their corresponding classifier weights and lacks explicit modeling of inter-class differences. This makes it difficult for the model to capture the critical discriminative features required to define classification boundaries. To overcome this limitation, we reframe the problem from a causal perspective and introduce a novel framework driven by counterfactuals. Our method first generates factual descriptions alongside corresponding inter-class counterfactuals to pinpoint the causal attributes essential for classification, then refines these representations via a mutual purification process, and finally leverages a novel separation loss to explicitly push the factual and counterfactual classifier weights apart. This strategy forces the model to forge clearer and more discriminative classification boundaries, achieving more accurate and robust classification. Extensive experiments demonstrate that our approach significantly outperforms existing state-of-the-art methods.

1 Introduction

With the rapid evolution of the digital world, machine learning classifiers need to frequently adapt to newly emerging or redefined categories (Zhou et al. 2024). This requirement, known as *classifier expansion*, is crucial for maintaining the model's practicality (De Lange et al. 2021). Conventional approaches such as full retraining or transfer learning can meet this need, but the effort of gathering fresh data and the associated computational cost often makes them prohibitively expensive (Zhuang et al. 2020). To overcome these challenges, researchers have proposed *zero-shot learning* (ZSL)–based classifier expansion strategies, establishing a paradigm that relies only on images from previously seen classes (Xian et al. 2018a,b). To further improve the generalization of ZSL, recent studies push the concept into a fully image-free setting (Christensen et al. 2023). In this paradigm, a model learns to map semantic information into the existing visual weight space using only attribute lists or textual descriptions of each class, thereby enabling the classifier to recognize unseen classes without any images.

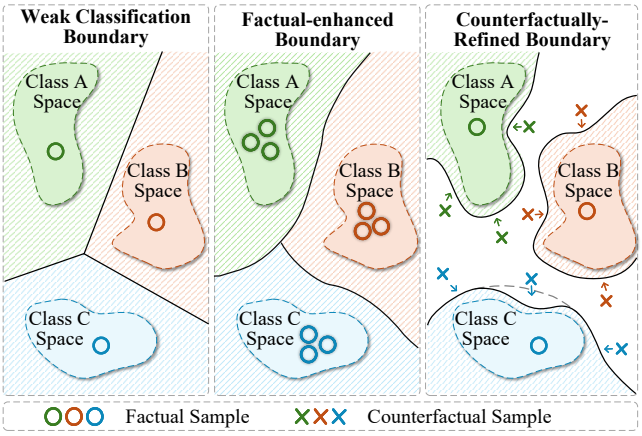

Figure 1: Conceptual illustration of classification boundary refinement. Counterfactuals (right) forge a clearer and more robust separation between classes compared to general classifiers (left and middle).

Although such approaches greatly reduce the need for images, the absence of visual calibration often amplifies semantic noise and leads to class confusion (Chen et al. 2024). Specifically, the mapping from semantics to weights in these methods is solely learned from attribute correlations and lacks explicit modeling of inter-class differences. These methods describe each class in isolation and then attempt to learn the co-occurrence between patterns from those descriptions and the corresponding weight patterns, which is essentially a form of shallow correlation fitting. When different classes have similar attributes or textual descriptions due to inherent similarity, this approach is particularly ineffective, as it struggles to capture the necessary fine-grained or discriminative features from co-occurrence patterns to distinguish between class differences.

The root of this issue is that by learning only from factual descriptions of classes, the model struggles to independently infer the critical discriminative information required to define classification boundaries. Even augmenting the model with more factual samples is often insuffi-

*Corresponding author.

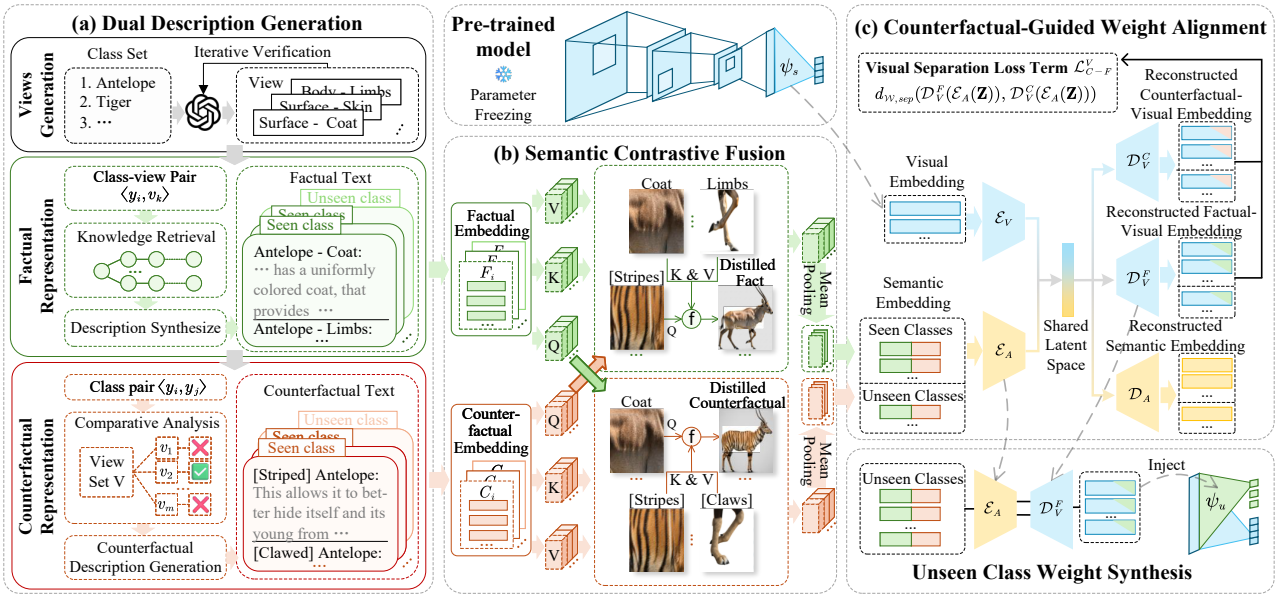

Figure 2: The overall architecture of our proposed framework. The figure illustrates the data flow across its three core modules: (a) Factual-Counterfactual Dual Description Generation (DDG), (b) Semantic Contrastive Fusion (SCF), and (c) Counterfactual-Guided Weight Alignment (CWA).

cient to refine these ambiguous boundaries (as shown in Figure 1). *Counterfactual reasoning*, a key concept in causal learning, is applied in many fields. It identifies the crucial drivers of an outcome by applying hypothetical interventions to certain variables (Pearl and Mackenzie 2018). Applying this causal-exploratory ability to the textual descriptions of classes can offer a viable path for discovering key discriminative features and support the precise characterization of inter-class differences. However, applying it presents several challenges. First, an effective intervention model amenable to counterfactual reasoning must be established, enabling reasonable interventions on semantic descriptions to ultimately discover the class's discriminative features. Second is the challenge of how to effectively characterize inter-class differences in order to efficiently construct informative counterfactual interventions that model these distinctions. Furthermore, how to better align the semantic and classifier weight space with the support of counterfactual information to achieve classifier expansion for unseen classes remains a critical challenge.

To address the above issues, this paper proposes a counterfactual-driven zero-shot classifier expansion framework, whose overall architecture is depicted in Figure 2. First, we design a Factual-Counterfactual Dual Description Generation (DDG) module. It automatically discovers descriptive perspectives that define a class's identity, thereby effectively constraining and narrowing the space for semantic intervention. Furthermore, by simulating attribute replacement to perform semantic intervention, it constructs the factual and corresponding counterfactual descriptions for a class. Next, a Semantic Contrastive Fusion Module (SCF) is designed. It employs a symmetric attention mechanism that allows the factual and counterfactual representations to mutually query and purify each other, jointly refining the most discriminative information relevant to the classification boundary. Finally, a Counterfactual-Guided Weight Alignment module (CWA) is constructed, which leverages the refined features within a counterfactual-assisted prediction strategy. It uses a novel separation loss to explicitly push the factual and counterfactual representations apart in the weight space, thereby forging more robust classification boundaries. The contributions of this paper are:

- We are the first to introduce counterfactuals into image-free classifier expansion, establishing a causal-view mapping between class semantics and classifier weights.
- A counterfactual-assisted prediction strategy is designed that aligns the semantic and weight space while jointly leveraging factual and counterfactual information, yielding more precise class-space modeling.
- We propose an automated pipeline to generate and verify factual-counterfactual descriptions, offering causal information with discriminative power beyond other methods and establishing a reusable paradigm for related tasks.

2 Related Work

2.1 ZSL-based Classifier Expansion

Zero-Shot Learning (ZSL) aims to recognize new classes without using any new samples but only using knowledge (Sun, Gu, and Sun 2021; Xian et al. 2018a). One common direction directly learns the classifiers for new classes, using techniques like generating weights with graph neural networks (Gidaris and Komodakis 2019) or applying subspace regularization (Akyürek et al. 2021). Another direction seeks to improve the alignment between visual and semantic features, for example, employing architectures like

Vision Mamba (Hou et al. 2025) or mechanisms such as cross-attention (Lai et al. 2024). A third line of work focuses on enriching the semantic space, for instance, by using Large Language Models (LLMs) to generate multi-view descriptions (Naeem et al. 2023) or by dynamically evolving semantic prototypes to better match visual data (Chen et al. 2023). However, despite these diverse strategies, a fundamental challenge persists: models often tend to learn spurious correlations from biased training data (Atzmon et al. 2020), especially when contextual cues are critical for distinguishing classes (Misra, Gupta, and Hebert 2017).

This problem is exacerbated in the Image-free ZSL (I-ZSL) setting, which operates without any visual data and relies solely on pre-trained model weights and semantic descriptions of original and new classes. (Christensen et al. 2023). In this paradigm, the reliance on semantic descriptions amplifies inherent textual biases (Chen et al. 2024). Methods applicable to this setting, such as those that combine classifiers based on co-occurrence statistics (Mensink, Gavves, and Snoek 2014), generate weights via a convex combination of semantic embeddings (Norouzi et al. 2014), or synthesize weights with dual-autoencoders (Christensen et al. 2023), still fundamentally rely on these potentially biased co-occurrence patterns. Relying only on semantic information, these models struggle to differentiate fine-grained classes that share similarities, leading to significant class confusion. Our work addresses this limitation by moving beyond co-occurrence to model the causal distinctions that define these classes.

2.2 Causal Reasoning in ZSL

Counterfactual reasoning is a core instrument of causal inference that can break spurious correlations (Pearl 2009). The central idea is to isolate the true causal effect of a variable by asking "what if" questions and contrasting factual outcomes with their hypothetical, counterfactual counterparts. Inspired by this, counterfactual reasoning has been widely adopted in computer vision, enabling unbiased prediction and generalization in tasks such as visual question answering and scene graph generation (Kolling et al. 2022; Song et al. 2024; Liu et al. 2025).

In the field of ZSL, some pioneering works have also drawn on causal principles. For instance, some methods approach the problem by learning disentangled representations for classes' attributes (Atzmon et al. 2020), or decomposing classes into their causal effects (Yang et al. 2022). More directly, some methods introduce counterfactual intervention by generating synthetic samples to rebalance the classifier between seen and unseen classes (Yue et al. 2021). However, these approaches have notable limitations. Their methods for constructing counterfactual scenarios, such as unconstrained feature modification, can be uncontrollable. In contrast, our work involves fine-grained interventions at semantic level, introducing a systematic and automatic pipeline for generating high-quality inter-class counterfactuals. By leveraging these explicit counterfactual descriptions, we guide the model to focus on causally-relevant discriminative features and enforce a clear separation margin in the weight space, thereby forging more robust classification boundaries.

3 Method

This section first formalizes the goal of our task, then details three core modules: the Factual-Counterfactual Dual Description Generation (DDG) module for establishing a semantic foundation, the Semantic Contrastive Fusion (SCF) module for refining the representations, and the Counterfactual-Guided Weight Alignment (CWA) module for the final weight synthesis.

3.1 Task Formulation

Let $\Phi : \mathcal{X} \rightarrow Y_s$ be a pre-trained classifier, where $\mathcal{X}$ represents the image space and Y_s is the set of seen classes. Specifically, the classifier Φ can be decomposed into a feature extractor ω and a classification layer ψ. This layer is parameterized by a weight matrix $\psi_s \in \mathbb{R}^{h_v \times |Y_s|}$, where h_v is the dimension of the features produced by ω. The objective of ZSL is to expand the classifier's capability to recognize a new set of unseen classes Y_u ($Y_u \cap Y_s = \emptyset$) by predicting their corresponding weight matrix $\psi_u \in \mathbb{R}^{h_v \times |Y_u|}$.

To effectively learn the key discriminative features, our work introduces a counterfactual intervention strategy. Our framework constructs a set of counterfactual auxiliary classes, each corresponding to a factual class $Y_f = Y_s \cup Y_u$. We then learn to synthesize not only the factual classifier weights but also the weights for these counterfactual counterparts, denoted as ψ_c. These counterfactual weights serve as a negative reference for a separation loss, which enforces a clear margin between factual and counterfactual representations in the weight space. This strategy guides the model to forge more robust classification boundaries. The following sections detail the components of our method, beginning with the generation of the factual and counterfactual dual descriptions that serve as its input.

3.2 Dual Description Generation

To forge a semantic foundation that isolates key discriminative features for each class, our framework generates a dual set of descriptions through the Factual-Counterfactual Dual Description Generation (DDG) module. The factual representation, $\mathbf{F}$, captures the inherent attributes of a class. The counterfactual representation, $\mathbf{C}$, is generated by simulating attribute interventions to pinpoint the causal views critical for distinguishing between classes. Together, this dual-representation approach provides a rich semantic input that explicitly encodes not only a class's characteristics but also the pivotal attributes that define its classification boundary.

Factual Description The generation process begins with creating factual descriptions for all classes in Y_f. These descriptions are constructed around a shared set of m descriptive "views", $V = \{v_1, v_2, \ldots, v_m\}$. Each view v_k represents a high-level perspective that encompasses a set of related attributes. For each class-view pair (y_i, v_k), we design an efficient and robust pipeline to generate the corresponding textual description, $\tilde{d}_{i,k}$, which consists of a textual account of the class's specific attributes pertinent to that view.

The above process is fully automated through a novel framework using a Large Language Model (LLM) and

Chain-of-Thought principles to ensure comprehensiveness[1]. This process yields a set of detailed textual descriptions. To prepare these for downstream modules, a pre-trained text encoder $\mathcal{E}(\cdot)$ transforms each description $\tilde{d}_{i,k}$ into a dense vector $\mathbf{f}_{i,k} \in \mathbb{R}^{d_e}$, where d_e is the dimension of the representation. These vectors are subsequently stacked to form the complete factual representation $F_i \in \mathbb{R}^{m \times d_e}$ for class y_i:

$$\mathbf{F}_i = [\mathbf{f}_{i,1}, \mathbf{f}_{i,2}, \ldots, \mathbf{f}_{i,m}]^T \quad (1)$$

Counterfactual Description While factual descriptions capture a class's attributes, they may not explicitly highlight the characteristics that are causally essential for classification. To isolate these causal drivers of distinction, our method employs counterfactual intervention. The goal is to generate specific descriptions that answer the question "what would happen if a key attribute were different?", thereby pinpointing the minimal changes that alter a class's identity and revealing the characteristics most critical for classification. Formally, a counterfactual statement of the outcome y of intervening on an attribute x can be expressed as[2]:

$$y(\text{do}(x = x')) = y' \quad (2)$$

However, not all interventions on descriptive views are meaningful for a classification task[3]. To focus our analysis on relevant and discriminative interventions, we introduce the concept of "Inter-Class Counterfactuals". This approach tailors the counterfactual generation process specifically to the task of distinguishing between classes. For a pair of classes (y_i, y_j), we find suitable inter-class counterfactuals by first identifying a set of discriminative views $V_{i,j} \subset V$. A view v_k is considered discriminative if an intervention that replaces the attribute of the target class y_i with that of the source class y_j is determined to be sufficient to change the identity of y_i. For each discriminative view $v_k \in V_{i,j}$, we then generate a coherent counterfactual text $\tilde{d}_k^{i,j}$. This is achieved by performing the causal intervention on the target class y_i, formalized as:

$$\tilde{d}_k^{i,j} \leftarrow y_i(\text{do}(\tilde{d}_{i,k} = \tilde{d}_{j,k})) \quad (3)$$

$\tilde{d}_k^{i,j}$ is a counterfactual description of target class y_i, which portrays the hypothetical state resulting from replacing its attribute under view v_k with that of the source class y_j. The resulting counterfactual texts, $\{\tilde{d}_k^{i,j}\}$, are then encoded into vectors using the same text encoder $\mathcal{E}(\cdot)$ and assembled to form the representation $\mathbf{C}_i$ for each target class y_i. If we denote the encoded vectors as $\{\mathbf{c}_{i,k}\}$ for each of the p discriminative views in $V_{i,j}$, the counterfactual representation $\mathbf{C}_i \in \mathbb{R}^{p \times d_e}$ is constructed as:

$$\mathbf{C}_i = [\mathbf{c}_{i,1}, \mathbf{c}_{i,2}, \ldots, \mathbf{c}_{i,p}]^T \quad (4)$$

Notably, the entire sophisticated process described here, including identifying discriminative views and generating and verifying counterfactual descriptions, is also handled by the efficient and robust LLM-driven pipeline, similar to the one introduced for factual descriptions[1].

[1]Detailed in Appendix C.

[2]Here, x is the attribute being intervened upon, and $y(\text{do}(x = x'))$ denotes the outcome of y after setting x to a new state x'.

[3]For instance, the view "has fur" is not useful for classifying a "lion" from a "tiger", while their "brindle" is.

3.3 Semantic Contrastive Fusion Module

Although the generated factual and counterfactual representations are comprehensive, they may contain semantic features that are either noisy or irrelevant to the classification task. Therefore, this module is designed to refine these initial representations into more discriminative embeddings. Traditional static fusion strategies, such as concatenation, tend to directly mix features of two representations, making it difficult to achieve effective refinement. In contrast, this module employs a symmetric attention mechanism where the factual and counterfactual representations for a class mutually query and purify each other, distilling discriminative information.

Specifically, the mechanism operates through two symmetric branches, shown in the Figure 2 (b). In the first, which we term the counterfactual refinement branch, a class's factual representation $\mathbf{F}_i$ queries its counterfactual representations $\mathbf{C}_i$ to produce a refined counterfactual vector $\mathbf{c}_i^r$:

$$\mathbf{c}_i^r = \text{MeanPool}(\text{Attn}(W_{Q_1}\mathbf{F}_i, W_{K_1}\mathbf{C}_i, W_{V_1}\mathbf{C}_i)) \quad (5)$$

Here, $W_{Q_1}, W_{K_1}, W_{V_1}$ and their counterparts in the symmetric branch, $W_{Q_2}, W_{K_2}, W_{V_2}$, are learnable weight matrices that project the inputs into query, key, and value spaces, respectively. Symmetrically, the second branch, the factual refinement branch, uses the counterfactual representations $\mathbf{C}_i$ as queries to attend to the factual matrix $\mathbf{F}_i$, producing a boundary-aware factual vector $\mathbf{f}_i^r$:

$$\mathbf{f}_i^r = \text{MeanPool}(\text{Attn}(W_{Q_2}\mathbf{C}_i, W_{K_2}\mathbf{F}_i, W_{V_2}\mathbf{F}_i)) \quad (6)$$

This process yields the final refined representations $\mathbf{F}^r$ and $\mathbf{C}^r$ by stacking the vectors $\mathbf{f}_i^r$ and $\mathbf{c}_i^r$ for all classes. These are then passed to the classifier weight alignment module.

In essence, this module functions as a mutual purification process. By compelling the factual (what a class is) and counterfactual (what a class is not) representations to interrogate each other, it ensures the resulting embeddings are distilled to contain the most causally relevant information, thus preparing a cleaner and more potent signal for the final weight alignment stage.

3.4 Counterfactual-Guided Weight Alignment

The final module of our framework maps the refined semantic representations from the SCF module to the visual weight space. Unlike standard alignment methods that risk learning from spurious correlations, our approach employs a counterfactual-assisted prediction strategy. It is explicitly guided by the contrast between factual and counterfactual representations to forge robust classification boundaries. This is operationalized through an enhanced dual-autoencoder network.

To contextualize our contributions, we first describe the standard dual-autoencoder framework (Liu, Li, and Gao 2020; Christensen et al. 2023). This framework consists of a semantic autoencoder $(\mathcal{E}_\mathcal{A}, \mathcal{D}_\mathcal{A})$ and a vision autoencoder $(\mathcal{E}_\mathcal{V}, \mathcal{D}_\mathcal{V})$. It is trained on the seen classes to align a semantic representation with the classifier weights. The training objective typically minimizes a weighted sum of four loss terms: a semantic reconstruction loss, a visual reconstruction loss, and two cross-modal alignment losses. For detailed formulations of these baseline losses, see Appendix A.1.

Our method enhances this architecture to explicitly model the structure of factual and counterfactual information. Specifically, we first concatente the refined factual and counterfactual representations for a comprehensive semantic representation, $\mathbf{Z} = [\mathbf{F}^r||\mathbf{C}^r]$, where $||$ denotes the concatenation operation. Building on this, we desgin two specialized decoders rather than a single one in the standard setting: a factual vision decoder ($\mathcal{D}_{\mathcal{V}}^{F}$) responsible for predicting the target classifier weights, and a counterfactual vision decoder ($\mathcal{D}_{\mathcal{V}}^{C}$) for predicting the auxiliary counterfactual weights.

The mechanism underpinning this counterfactual-assisted prediction is a separation loss, $\mathcal{L}_{C-F}^{\mathcal{V}}$. This loss, applied to the outputs of the two vision decoders, forges a clear classification boundary by maximizing the separation between the factual and their corresponding counterfactual weight vectors. It is defined over the entire set of classes as:

$$\mathcal{L}_{C-F}^{\mathcal{V}} = d(\mathcal{D}_{\mathcal{V}}^{F}(\mathcal{E}_{\mathcal{A}}(\mathbf{Z})), \mathcal{D}_{\mathcal{V}}^{C}(\mathcal{E}_{\mathcal{A}}(\mathbf{Z}))) \tag{7}$$

Here, the metric $d(\cdot, \cdot)$ is cosine similarity. By explicitly pushing the representation of what a class is away from what it could be, this loss forces the model to create a distinct margin, thereby helping to resolve confusion between semantically similar categories. The total loss is the sum of adapted reconstruction, cross-modal, and separation loss terms.

Weight Synthesis for Unseen Classes The synthesis of unseen classes' weight, ψ_u, is achieved by passing their refined semantic representations, $\mathbf{Z}_u$, through the trained semantic encoder and the factual vision decoder:

$$\psi_u = \mathcal{D}_{\mathcal{V}}^{F}(\mathcal{E}_{\mathcal{A}}(\mathbf{Z}_u)) \tag{8}$$

Thus, the counterfactual vision decoder and its auxiliary weights ψ_c serve their assistive role exclusively during training to create the discriminative margin, and are not used during the inference phase. The newly synthesized factual weights ψ_u are then injected into the original classifier, expanding its capabilities to recognize the new classes.

Core Theoretical Insights The effectiveness of separation loss $\mathcal{L}_{C-F}^{\mathcal{V}}$ is supported by a solid theoretical justification. The core geometric intuition is that the "inter-class counterfactuals" generated by the DDG module are constructed to be geometrically intermediate between the representations of two factual classes in the semantic space $\mathcal{A}$ (Appendix A.2).

A key insight is that the mapping G from the semantic space $\mathcal{A}$ to the weight space $\mathcal{V}$, implemented by the semantic encoder and the factual vision decoder $\mathcal{D}_{\mathcal{V}}^{F} \circ \mathcal{E}_{\mathcal{A}}$, is constrained by our framework to preserve this geometric structure. We approximate this mapping G as an affine transformation. This is a reasonable approximation because MLP-based mappings are piecewise affine, for example, when using ReLU activations. For semantically similar classes, where counterfactual intervention is most critical, it is likely they fall within the same linear region, making the transformation effectively affine for those inputs (Appendix A.3).

Because an affine transformation preserves collinearity, a counterfactual representation lying on the factual boundary in semantic space will be mapped to the corresponding decision boundary in the weight space. Our separation loss $\mathcal{L}_{C-F}^{\mathcal{V}}$ then explicitly pushes the factual weights away from this boundary defined by the counterfactual weights. This creates a clear and robust classification margin (Appendix A.3). The strong empirical performance shown in our ablation studies further provides a practical validation of this affine approximation.

4 Experiments

In this section, we first describe the experimental setup. Next, comparative results with baselines are presented. Subsequently, we conduct the ablation study and provide the robustness, applicability, and generalizability analysis.

4.1 Experimental Setup

Datasets. The method is evaluated on three benchmarks for ZSL: CUB (Wah et al. 2011), AWA2 (Xian et al. 2018a), and SUN (Patterson et al. 2014). To ensure consistency with previous studies, this work adheres to the standard dataset splits and evaluation protocols proposed by the study (Xian et al. 2018a). CUB is a fine-grained benchmark for bird species recognition, comprising 200 distinct species (150 seen and 50 unseen classes). AWA2, consisting of 40 seen and 10 unseen classes, features 50 animal classes. SUN encompasses 717 indoor and outdoor scene classes, split into 645 seen and 72 unseen classes. Since our focus is on image-free ZSL, the models are trained exclusively using the semantic class attributes, without access to the images.

Baselines. Our method is compared with representative baselines from three main ZSL paradigms: semantic embedding, classifier weight mapping, and classifier weight generation. Specifically: 1) ConSE (Norouzi et al. 2014) forms the representation of test images by a convex combination of semantic embeddings of the seen classes. 2) VGSE (Xu et al. 2022) learns new semantic embeddings by clustering visual parts and generalizes them via a class-relation module. 3) COSTA (Mensink, Gavves, and Snoek 2014) forms unseen classifiers by blending seen classifiers, with mixing weights derived from co-occurrence data. 4) SubReg (Akyürek et al. 2021) constrains new class classifiers to the subspace of base classifiers using a dedicated regularization loss. 5) wDAE (Gidaris and Komodakis 2019) trains a GNN-based denoising autoencoder to reconstruct clean classifier weights from noisy inputs. 6) ICIS (Christensen et al. 2023) synthesizes classifier weights from semantics without images, learning a bidirectional mapping via a dual autoencoder.

Metrics. The method is evaluated under two different settings: standard ZSL and generalized ZSL (GZSL). For the former, the average per-class top-1 accuracy (**T**) is reported, with the search space restricted to the set of unseen classes in the testing phase. For the latter, the search space is adjusted to the union of both seen and unseen classes, and denote their accuracies as **u** and **s**, respectively. For comprehensive measurement, the harmonic mean **H** is also used for GZSL[4],

[4]In GZSL, the recognition of unseen classes (**u**) requires overcoming the model's bias towards seen classes (**s**). This correction process often sacrifices part of the performance on seen classes. Therefore, the harmonic mean (**H**) is considered as a trade-off.

Application Scenario	Method	CUB				AWA2				SUN			
		T	H	u	s	T	H	u	s	T	H	u	s
Expert-constructed Class Attributes	ConSE	41.9	0.9	0.5	88.0	44.0	5.7	3.0	96.1	44.4	0.1	0.1	47.9
	COSTA	31.9	0.0	0.0	87.6	40.9	0.0	0.0	96.1	19.9	0.0	0.0	50.1
	SubReg	37.6	0.0	0.0	87.6	37.5	0.0	0.0	96.1	48.3	0.0	0.0	50.1
	wDAE	38.2	0.0	0.0	87.3	37.0	0.3	0.1	96.0	49.9	0.0	0.0	49.3
	VGSE	45.1	44.8	39.2	52.3	55.4	47.3	31.8	92.4	42.7	3.1	42.5	1.6
	ICIS	60.6	56.5	45.8	73.7	64.6	51.6	35.6	93.3	51.8	32.7	45.2	25.6
LLM-generated View Attributes	ConSE	46.8	0.8	0.4	87.9	51.8	5.0	2.6	96.3	40.1	0.4	0.2	49.6
	COSTA	40.9	0.0	0.0	87.9	56.5	0.0	0.0	96.3	31.0	0.0	0.0	52.3
	SubReg	61.0	3.2	1.6	87.9	65.0	0.1	0.1	96.2	55.3	1.9	1.0	52.3
	wDAE	57.3	4.7	2.4	87.6	65.9	0.3	0.1	96.1	55.8	5.5	2.9	50.7
	VGSE	48.1	48.0	39.5	61.7	59.5	55.6	41.7	83.3	44.0	17.2	41.1	11.0
	ICIS	58.0	54.7	43.5	73.6	64.7	53.0	37.4	92.4	55.8	33.4	26.1	47.5
	Ours	**67.3**	**61.3**	58.4	64.4	**76.3**	**67.9**	55.7	86.8	**60.9**	**41.3**	41.9	40.6
	ΔImp	10%	12%	-	-	16%	22%	-	-	9%	20%	-	-

Table 1: Comparison of the proposed method with other baselines in terms of top-1 accuracy (%) and harmonic mean. The row "ΔImp", which can be calculated by $(p_2 - p_1)/p_1$, represents the improvement of our proposed method (p_2) compared with the SOTA baseline (p_1). Higher accuracy or harmonic mean across three widely used datasets is highlighted in **bold**.

which is calculated as $H = (2 \times s \times u)/(s + u)$.

Settings. The base classifier weights for the seen classes are sourced from the final classification layer of a pre-trained ResNet101 (He et al. 2016) model. For the generation of dual factual-counterfactual descriptions, we utilize GPT-4o (Achiam et al. 2023) as the main LLM and CLIP (Radford et al. 2021) for embeddings, where all components are replaceable. Both encoders are single-layer linear mappings whose outputs are passed to the decoders through a ReLU activation function, which are two-layer MLPs with a hidden dimension of 4096. During training, we use the Adam optimizer (Kinga, Adam et al. 2015) with a learning rate of 10^{-5} and set the batch size to 16, 10, and 24 for CUB, AWA2, and SUN, respectively. More detailed experimental setup and descriptions are presented in Appendix B.1.

4.2 Comparison with State-of-the-Art Methods

To make the experimental results more comparable, two different application scenarios are constructed in this experiment: 1) expert-constructed class attributes only; 2) LLM-generated view attribute construction. In the second scenario, the baselines apply mean pooling to integrate the LLM-based multidimensional attributes to replace the factual refinement in our method. The experiment is conducted on three datasets, and the results are reported in Table 1.

Regardless of the scenario, the experimental results confirm that our method consistently outperforms all baselines across the three benchmark datasets. Meanwhile, the assistance of counterfactuals not only enhances the recognition of unseen classes for ZSL, but also achieves a better balance between the accuracy of seen and unseen classes for GZSL. This suggests that the introduction of semantic fusion and counterfactual assistance can mitigate confounding factors to some extent that affect class discrimination, thereby enabling the model to focus on the core features of each class.

4.3 Ablation Study

To independently validate the contribution of each key component, we perform a series of ablation studies, including counterfactual, factual, and weight alignment ablations. In each study, we start from our full model and replace or remove only one component at a time. Specifically: 1) Replace the counterfactual refinement branch of SCF with mean pooling for fusion. 2) Replace inter-class counterfactuals with arbitrary interventions. 3) Remove all modules related to counterfactuals. 4) Replace the factual refinement branch of SCF with mean pooling for fusion. 5) Replace the CWA module with a basic encoder-decoder model, which assigns semantic attributes to classifier weights.

As demonstrated in Table 2, the replacement or removal of any individual component degrades the performance of the model, highlighting the contribution of each component. Based on the factual representation, the assistance of counterfactuals can refine the factual part to better capture the core attributes. Furthermore, compared with arbitrary interventions, inter-class counterfactuals can select meaningful and effective attributes for intervention, fundamentally enabling the model to learn true class-discriminative criteria as much as possible.

In addition to validating the core modules, we also provide a detailed breakdown of the loss functions within the CWA module in Appendix B.2 . That study begins from a basic encoder-decoder model and incrementally adds loss components to verify their contributions. The results confirm that while factual and counterfactual reconstruction and alignment losses are necessary, they alone cannot fully constrain class discrimination in the latent space. The key contributor is our visual separation loss ($\mathcal{L}^{\mathcal{V}}_{C-F}$), which significantly enhances the model's discriminative ability by explic-

Method Ablation	CUB				AWA2				SUN			
	T	H	u	s	T	H	u	s	T	H	u	s
Ours	**67.3**	**61.3**	58.4	64.4	**76.3**	**67.9**	55.7	86.8	**60.9**	**41.3**	41.9	40.6
w/o Counterfactual refinement branch	64.6	59.5	55.5	64.2	70.6	64.8	52.2	85.4	57.2	39.8	41.9	38.0
w/o Inter-class counterfactuals	61.5	57.6	51.0	66.3	66.2	61.7	47.9	86.7	56.9	37.0	47.2	30.5
w/o Counterfactual assistance	60.6	57.4	50.2	67.1	65.5	60.3	46.3	86.9	56.7	36.6	47.7	29.7
w/o Fact refinement branch	57.2	54.0	51.1	57.1	59.3	54.7	40.3	86.5	56.4	34.0	49.5	26.0
w/o Weight alignment	57.4	54.0	45.4	66.7	55.7	49.9	34.8	88.1	54.9	31.6	24.9	46.3

Table 2: Ablation study on the core modules of our proposed method in terms of top-1 accuracy T (%) and harmonic mean H.

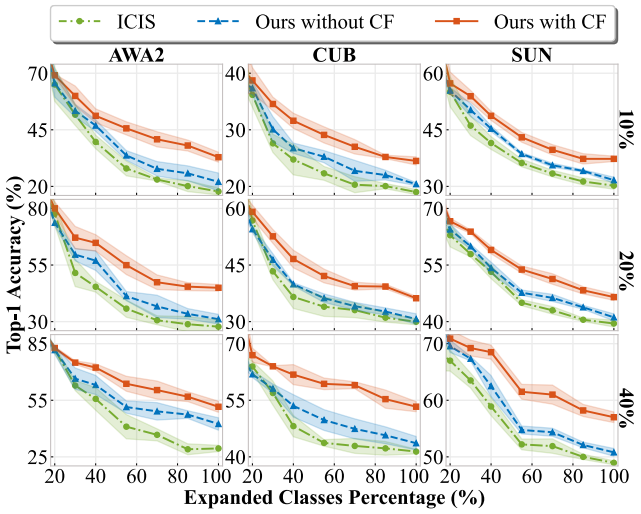

Figure 3: Comparative accuracy of the SOTA baseline (ICIS) and our method (with and without counterfactual assistance) across proportions of seen classes and expanded classes.

itly forcing the factual and counterfactual representations apart in the visual weight space.

4.4 More Empirical Analysis

Robustness Analysis. To explore the robustness of our method and its ability in data-scarce scenarios, we conduct experiments on three datasets by setting the proportion of seen classes to 10%, 20%, and 40%, respectively. On this basis, we measure the class expansion ability of our method in terms of top-1 accuracy (**T**), and the performance is presented in Figure 3. We observe that our method demonstrates considerable robustness in data-scarce scenarios. Meanwhile, as the expansion of classes increases, our method can expand the classes to more than twice that of the baseline, and the performance disparity between our method and the other baselines becomes increasingly apparent. Notably, when the expansion reaches 100%, our method requires only 20% of the seen classes to approximately achieve the performance of the baselines with 40% of the seen classes, which further emphasizes that counterfactual assistance enables the model to expand more classes by focusing on discriminative features. A more comprehensive experimental analysis is provided in Appendix B.6.

Embed Model	CLIP		SBERT		LLaMA		Qwen	
LLM	T	H	T	H	T	H	T	H
GPT-4o	76.3	67.9	69.1	60.9	69.1	59.8	69.2	63.0
GPT-4o-mini	74.6	64.0	66.1	58.6	67.4	58.9	65.6	56.6
Gemini-2.5	72.9	62.3	65.0	58.1	66.7	56.0	65.8	54.1
LLaMA-3.1	71.6	62.6	66.9	58.8	68.3	59.1	68.8	60.5
Qwen-2.5	65.8	58.5	65.0	54.6	65.7	54.4	68.6	57.1

Table 3: Comparative analysis of our proposed method with various LLMs and embedding models on the AWA2 dataset.

Applicability Analysis. To evaluate the practical applicability of our method, we conduct experiments on three datasets using different LLMs and embedding models. The evaluated LLMs include GPT-4o, GPT-4o-mini, Gemini-2.5-flash (Team et al. 2025), LLaMA-3.1 (Touvron et al. 2023) and Qwen-2.5 (Bai et al. 2023). For embedding models, in addition to CLIP, we also select SBERT (Reimers and Gurevych 2019), LLaMA, and Qwen. The results on AWA2 are shown in Table 3. We adopt a multi-session strategy to parallelize queries, efficiently generating view and counterfactual descriptions in batches of 10 classes. On average, for each LLM mentioned, the full inference process to collect all required semantic descriptions for the AWA2, CUB, and SUN datasets takes approximately 15, 30, and 50 minutes, respectively, and is a one-time cost. It is clear that no matter which of the settings is chosen, our method outperforms the SOTA baseline, providing broad prospects and flexibility for the application of our model. Additional datasets and more detailed experiments are presented in Appendix B.8.

Cross-Dataset Generalization. To demonstrate the generalizability of our method, a classifier pre-trained on ImageNet (Deng et al. 2009) is expanded to recognize classes from several other datasets without requiring any images. The required semantic attributes of the baselines are derived from two alternative knowledge sources, including ConceptNet (Speer and Lowry-Duda 2017) and Wiki2Vec (Yamada et al. 2018). Table 4 shows the results when generalizing from ImageNet to the unseen classes of the other three datasets. Our method consistently outperforms the baselines and demonstrates improved generalization to unseen classes,

Method	SubReg		wDAE		VGSE		ICIS		Ours	
Dataset	**WV**	**CN**	**WV**	**CN**	**WV**	**CN**	**WV**	**CN**	**WV**	**CN**
CUB	3.7	4.6	3.6	6.0	7.3	11.1	3.5	6.6	**19.5**	**21.2**
AWA2	18.4	59.9	62.2	59.9	52.1	64.7	65.5	64.8	**84.8**	**78.2**
SUN	1.6	11.7	7.4	12.1	7.2	11.3	9.1	13.6	**21.3**	**22.0**

Table 4: Top-1 accuracy of baselines and our method when transferred from ImageNet to three benchmark datasets. The columns "WV" and "CN" represent Wiki2Vec and ConceptNet, which are knowledge sources provided for baselines.

regardless of the granularity or domain of the seen classes.

In addition to the main results presented above, we conduct more experiments from additional perspectives, including qualitative confusion analysis, robustness and applicability assessments, and the adaptability to class incremental learning. These analyses are detailed in Appendix B.

5 Conclusion

In this paper, we introduced a novel counterfactual-driven framework for zero-shot classifier expansion, that mitigates reliance on superficial correlations. Our automated pipeline uses a LLM to generate factual and inter-class counterfactual descriptions. These descriptions are then leveraged by a contrastive fusion module and a dual-decoder alignment architecture, where a visual separation loss explicitly sculpts the classification boundaries in the weight space. Our experimental results demonstrated the superiority of our method, as it consistently and significantly outperformed SOTA baselines across the CUB, AWA2, and SUN benchmarks. Furthermore, extensive ablation studies validated the effectiveness of each component, particularly the substantial improvements brought by the counterfactual-guided separation loss. By integrating causality, this work offers a more robust solution for classifier expansion and a reusable paradigm for other zero-shot learning tasks.

Acknowledgments

This research was supported in part by the National Key R&D Program of China (No. 2021ZD0111700), in part by the National Nature Science Foundation of China (No. 62406302, 62137002, 62576327), in part by the Natural Science Foundation of Anhui province (No. 2408085QF195), in part by the Fundamental Research Funds for the Central Universities under Grant WK2150110035.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akyürek, A. F.; Akyürek, E.; Wijaya, D. T.; and Andreas, J. 2021. Subspace regularizers for few-shot class incremental learning. *arXiv preprint arXiv:2110.07059*.

Atzmon, Y.; Kreuk, F.; Shalit, U.; and Chechik, G. 2020. A causal view of compositional zero-shot recognition. *Advances in Neural Information Processing Systems*, 33: 1462–1473.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Chen, S.; Fu, D.; Chen, S.; Ye, S.; Hou, W.; and You, X. 2024. Causal visual-semantic correlation for zero-shot learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 4246–4255.

Chen, S.; Hou, W.; Hong, Z.; Ding, X.; Song, Y.; You, X.; Liu, T.; and Zhang, K. 2023. Evolving semantic prototype improves generative zero-shot learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, 4611–4622. PMLR.

Christensen, A.; Mancini, M.; Koepke, A.; Winther, O.; and Akata, Z. 2023. Image-free classifier injection for zero-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19072–19081.

De Lange, M.; Aljundi, R.; Masana, M.; Parisot, S.; Jia, X.; Leonardis, A.; Slabaugh, G.; and Tuytelaars, T. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7): 3366–3385.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Gidaris, S.; and Komodakis, N. 2019. Generating classification weights with GNN denoising autoencoders for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21–30.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hou, W.; Fu, D.; Li, K.; Chen, S.; Fan, H.; and Yang, Y. 2025. ZeroMamba: Exploring Visual State Space Model for Zero-Shot Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3527–3535.

Kinga, D.; Adam, J. B.; et al. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations*.

Kolling, C.; More, M.; Gavenski, N.; Pooch, E.; Parraga, O.; and Barros, R. C. 2022. Efficient counterfactual debiasing for visual question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3001–3010.

Lai, H.; Yao, Q.; Jiang, Z.; Wang, R.; He, Z.; Tao, X.; and Zhou, S. K. 2024. Carzero: Cross-attention alignment for radiology zero-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11137–11146.

Liu, L.; Sun, S.; Zhi, S.; Shi, F.; Liu, Z.; Heikkilä, J.; and Liu, Y. 2025. A causal adjustment module for debiasing scene

graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 12562–12580.

Liu, Y.; Li, J.; and Gao, X. 2020. A Simple Discriminative Dual Semantic Auto-Encoder for Zero-Shot Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Mensink, T.; Gavves, E.; and Snoek, C. G. 2014. COSTA: Co-occurrence statistics for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2441–2448.

Misra, I.; Gupta, A.; and Hebert, M. 2017. From red wine to red tomato: Composition with context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1792–1801.

Naeem, M. F.; Khan, M. G. Z. A.; Xian, Y.; Afzal, M. Z.; Stricker, D.; Van Gool, L.; and Tombari, F. 2023. I2MVFormer: Large language model generated multi-view document supervision for zero-shot image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15169–15179.

Norouzi, M.; Mikolov, T.; Bengio, S.; Singer, Y.; Shlens, J.; Frome, A.; Corrado, G. S.; and Dean, J. 2014. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*.

Patterson, G.; Xu, C.; Su, H.; and Hays, J. 2014. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108(1–2): 59–81.

Pearl, J. 2009. Causal inference in statistics: An overview. *Statistics Surveys*, 3: 96–146.

Pearl, J.; and Mackenzie, D. 2018. *The book of why: The new science of cause and effect*. Basic Books.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv preprint arXiv:1908.10084*.

Song, L.; Yang, C.; Li, X.; and Shang, X. 2024. A robust dual-debiasing VQA model based on counterfactual causal effect. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 4242–4252.

Speer, R.; and Lowry-Duda, J. 2017. ConceptNet at SemEval-2017 task 2: Extending word embeddings with multilingual relational knowledge. *arXiv preprint arXiv:1704.03560*.

Sun, X.; Gu, J.; and Sun, H. 2021. Research progress of zero-shot learning. *Applied Intelligence*, 51(6): 3600–3614.

Team, L.; Modi, A.; Veerubhotla, A. S.; Rysbek, A.; Huber, A.; Anand, A.; Bhoopchand, A.; Wiltshire, B.; Gillick, D.; Kasenberg, D.; et al. 2025. Evaluating Gemini in an arena for learning. *arXiv preprint arXiv:2505.24477*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset. *Computation & Neural Systems Technical Report*.

Xian, Y.; Lampert, C. H.; Schiele, B.; and Akata, Z. 2018a. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9): 2251–2265.

Xian, Y.; Lorenz, T.; Schiele, B.; and Akata, Z. 2018b. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5542–5551.

Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; and Akata, Z. 2022. VGSE: Visually-grounded semantic embeddings for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9316–9325.

Yamada, I.; Asai, A.; Sakuma, J.; Shindo, H.; Takeda, H.; Takefuji, Y.; and Matsumoto, Y. 2018. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. *arXiv preprint arXiv:1812.06280*.

Yang, M.; Xu, C.; Wu, A.; and Deng, C. 2022. A decomposable causal view of compositional zero-shot learning. *IEEE Transactions on Multimedia*, 25: 5892–5902.

Yue, Z.; Wang, T.; Sun, Q.; Hua, X.-S.; and Zhang, H. 2021. Counterfactual zero-shot and open-set visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15404–15414.

Zhou, D.-W.; Wang, Q.-W.; Qi, Z.-H.; Ye, H.-J.; Zhan, D.-C.; and Liu, Z. 2024. Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 9851–9873.

Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; and He, Q. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1): 43–76.