Information Processing and Management 63 (2026) 104414

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Superpixel-based Visual Feature Enhancement for Compositional Zero-Shot Learning

Wenlong Du [a], Xianglin Bao [a], Xiaofeng Xu [a,b,*], Xingyu Lu [c], Ruiheng Zhang [d]

[a] School of Computer and Information, Anhui Polytechnic University, Wuhu 241000, China
[b] Industrial Innovation Technology Research Co., LTD, Anhui Polytechnic University, Wuhu 241000, China
[c] School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[d] School of Mechatronical Engineering, Beijing Institute of Technology, Beijing 100081, China

ARTICLE INFO

Keywords:
Compositional zero-shot learning
Attribute-object combinations
Superpixel segmentation
Fourier spectral layer
Attention fusion

ABSTRACT

Compositional Zero-Shot Learning (CZSL) is a challenging machine learning task that recognizes new compositional concepts by leveraging learned concepts such as attribute-object combinations. Previous research depended on visual attributes derived from networks pre-trained in object categorization. These approaches are limited in capturing the subtleties of attribute distinctions and fail to account for the critical contextual interactions between attributes and visual objects. To address this problem, in this work, we draw inspiration from superpixels and introduce the Superpixel-based Visual Feature Enhancement (SVFE) model for the compositional zero-shot learning task. In the proposed approach, an innovative superpixel integration strategy is designed to meticulously disentangle and represent the visual concepts of states and objects with finer granularity. Then, we introduce a novel Fourier spectral layer that harnesses the frequency domain to capture global image features and dynamically adjusts component contributions to enhance the local detail representation. Furthermore, we propose a long-range fusion module to optimize the synergy between the local and global features, thereby fortifying the model's acuity in discerning intricate compositional relationships. Through rigorous experiments on standard CZSL benchmark datasets, the proposed SVFE model demonstrates significant improvement over other state-of-the-art methods in both open-world and closed-world CZSL scenarios.

1. Introduction

In the ever-changing world, the ability to recognize and understand novel combinations of concepts is not just a hallmark of human intelligence but also a crucial capability for artificial intelligence systems. Consider the cognitive ease with which we identify a "green tiger", a creature that, while fantastical, is instantly recognizable due to its constituent parts "green" and "tiger" (Xu, Tsang, & Liu, 2021). This innate ability to decompose and recombine known attributes to form an understanding of unseen entities is the essence of compositional zero-shot learning (CZSL). CZSL is a novel machine learning paradigm that aims to equip machines with the ability to generalize from known compositions to infer the properties of unseen combinations, much like how humans leverage existing knowledge to comprehend new phenomena (Mancini, Naeem, Xian, & Akata, 2021). This capability is not only a testament to the flexibility and robustness of learning algorithms but also holds significant practical implications for real-world applications.

* Corresponding author at: School of Computer and Information, Anhui Polytechnic University, Wuhu 241000, China.
E-mail addresses: 2230911101@stu.ahpu.edu.cn (W. Du), baoxianglin@ahpu.edu.cn (X. Bao), xuxiaofeng@ahpu.edu.cn (X. Xu), ee_luxingyu@njust.edu.cn (X. Lu), ruiheng.zhang@bit.edu.cn (R. Zhang).

https://doi.org/10.1016/j.ipm.2025.104414
Received 19 October 2024; Received in revised form 23 August 2025; Accepted 19 September 2025
Available online 3 October 2025

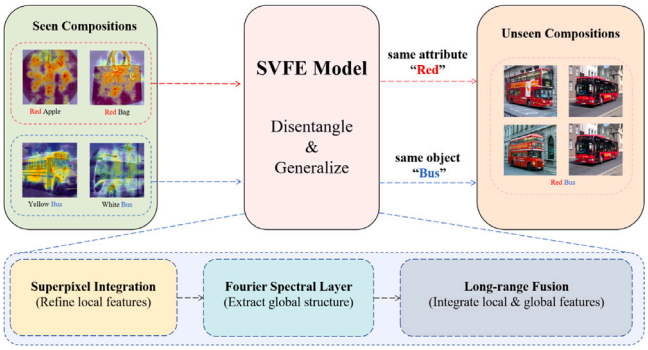

Fig. 1. Conceptual diagram of compositional zero-shot learning with SVFE model. SVFE first learns from seen attribute-object compositions (such as "red apple" and "yellow bus") and then leverages this knowledge to identify and construct novel unseen compositions (such as "red bus").

Nowadays, compositional zero-shot learning has attracted increasing attention and achieved much progress. However, existing CZSL methods struggle with the nuanced challenge of extracting fine-grained image features that capture the subtleties of visual content due to the unique characteristics of different domains. The visual appearance and distribution of categories can vary greatly across domains, making it difficult for models to generalize from seen domains to unseen ones. In the context of CZSL, fine-grained features take on a unique significance. Unlike conventional definitions that focus on local characteristics such as a bus emblem, fine-grained features in CZSL refer to the specific attributes or objects within an image that are essential for distinguishing categories across different domains. When effectively identified and utilized, these features can help bridge the domain gap by capturing the subtle yet important details that differentiate categories.

Initial approaches leveraging pre-trained networks for attribute extraction demonstrated efficacy in forming attribute-object compositions (Nagarajan & Grauman, 2018; Purushwalkam, Nickel, Gupta, & Ranzato, 2019; Xu, Bao, Lu, Zhang, Chen, & Lu, 2023; Xu, Tsang, Cao, Zhang, & Liu, 2019), yet their dependence on fixed representations inherently constrained adaptability to novel compositions and nuanced attribute variations. Subsequent metric learning methods addressed global interdependencies through shared embedding spaces (Atzmon, Kreuk, Shalit, & Chechik, 2020), but fundamentally struggled with fine-grained disentanglement as spatial insensitivity entangled localized cues like color distribution with object features. This limitation precipitated compositional overfitting, particularly evident when holistic representations failed to distinguish region-specific patterns such as differentially manifested "striped" attributes.

Building upon these foundations, generative and graph-based techniques explicitly pursued disentanglement through symmetry principles and prototype propagation (Li, Xu, Mao, & Lu, 2020; Ruis, Burghouts, & Bucur, 2021). Notwithstanding these advances, three persistent challenges emerged: transformation rules exhibited structural rigidity that impeded adaptation to non-uniform attribute distributions; graph convolutions induced contextual oversmoothing, which homogenized textural boundaries; and prototype misalignment occurred when attribute locations diverged from object anchors (Zhang et al., 2024, 2022). These collective shortcomings in granular isolation, adaptive configuration, and local context preservation culminated in recognition failures for compositions demanding precise attribute-object interplay—most notably the "green tiger" paradigm where subtle color-texture interactions evade distinction.

Recent methodological innovations have sought to address these gaps, though significant limitations endure. Prompt-based frameworks, such as ASP (Munir, Qureshi, Khan, & Ali, 2024), enhance attribute localization through attention mechanisms but remain hindered by frozen backbone architectures that perpetuate spatial misalignments. Vision-language approaches, including OWC (Jayasekara, Pham, Saini, & Shrivastava, 2025), unify compositional reasoning while still depending on heuristic filtering for implausible pairs. Simultaneously, feature enhancement strategies such as ProCC (Huo et al., 2024) achieve progressive cross-primitive alignment without attaining pixel-wise adaptability, and categorical formalisms like CatCom (Chytas, Kim, & Singh, 2025) overlook indispensable low-level visual cues during compositional reasoning.

To address the above limitations, in this work, we introduce a novel Superpixel-based Visual Feature Enhancement (SVFE) approach that integrates superpixel algorithms and architectural innovations. Superpixels, as aggregates of pixels based on low-level attributes, offer an efficient means of representing image data while mitigating redundant and noisy pixel-level information (Achanta et al., 2012; Achanta & Süsstrunk, 2017). By incorporating superpixels, we can adeptly capture global information within images and more precisely represent local features, thereby enhancing the model's comprehension of image intricacies. To capture expansive global information, we design the Fourier spectral layer to transform the image features into the frequency domain with an optimized computational expenditure and dynamically adjust the component contributions. Furthermore, we propose the long-range fusion module, a sophisticated and lightweight decomposition fusion module, to adeptly disentangle the linguistic characteristics into the distinct attribute and object features. Then, these attribute and object features are seamlessly integrated with the finely resolved image features, enhancing the model's ability to discern and learn from the intricacies of visual data. The long-range fusion module equipped with a parallel processing mechanism could adeptly capture and integrate long-range dependencies, which ensures the effective merging of local features and thereby enhances the model's overall representational power. As shown in Fig.

1, by integrating superpixel-extracted fine-grained features into our model, we capture critical cross-domain discriminative details, bridging the domain gap to enhance generalization and recognition accuracy for unseen target categories.

Considering the challenge of recognizing a novel composition like a “green tiger,” where conventional methods fail due to their inability to associate unseen attribute-object combinations and capture subtle features such as color gradients in green fur against stripes, our SVFE model demonstrates critical enhancements. It first segments images into superpixels, aggregating pixel features to precisely capture local color and texture details alongside their spatial relationships. This enables accurate identification of fine-grained attributes like green hue distribution. Subsequently, the Fourier spectral layer transforms features into the frequency domain, amplifying imperceptible color variations and texture patterns overlooked in the spatial domain. Distinct green shades and complex stripe details thus become more prominent. Finally, the long-range fusion module integrates these localized features with global context through parallel processing. By merging superpixel-level details with holistic dependencies across the feature map, the model effectively relates the green attribute to the tiger object via learned representations. This integrated approach overcomes prior generalization limitations in compositional zero-shot learning, advancing fine-grained feature extraction for novel compositions.

The main contributions of this paper are summarized as follows:

(1) We innovatively integrate superpixel algorithms into the CZSL framework, effectively disentangling visual concepts of states and objects, providing a finer granularity for feature extraction.
(2) A novel Fourier spectral layer is introduced to enhance the global image feature capture by transforming to the frequency domain and dynamically adjusting the component contributions.
(3) We propose a long-range fusion module that optimizes the integration of local and global features, bolstering the model’s ability to recognize complex compositional relationships.
(4) Rigorous experiments demonstrate the superior performance of the proposed method in both open-world and closed-world CZSL scenarios.

The rest of this paper is organized as follows. Section 2 reviews related work and Section 3 provides the preliminary formulation. The methodology of the proposed SVFE model is introduced in Section 4. Section 5 presents the experimental results and analysis. Section 6 gives the discussion and Section 7 concludes the paper.

2. Related work

2.1. Compositional zero-shot learning

Compositional zero-shot learning is a specialized subset of zero-shot learning (ZSL) that aims to recognize unseen attribute-object compositions by leveraging seen compositions during training. The concept of CZSL was first explored by Misra et al. who proposed projecting composed primitives and visual features into a joint embedding space to address the task (Misra, Gupta, & Hebert, 2017). Over time, various approaches have been developed to enhance the performance and generalization capabilities of CZSL models.

Early works in visual attribute learning have laid the groundwork for CZSL. Ferrari and Zisserman introduced the idea of learning visual attributes using a probabilistic generative model (Ferrari & Zisserman, 2007). Lampert et al. advanced this by utilizing visual attributes to detect unseen objects through attribute-based multi-label classification (Lampert, Nickisch, & Harmeling, 2009). These foundational studies paved the way for understanding how visual properties can be learned and transferred across different objects. With the rapid development of neural networks, Nagarajan et al. treated attributes as matrix operators applied to object vectors, providing a novel perspective on attribute-object interactions (Nagarajan & Grauman, 2018). Purushwalkam et al. introduced a task-driven modular architecture that reweights a set of sub-tasks to learn unseen compositions, emphasizing the flexibility and adaptability required for CZSL (Purushwalkam et al., 2019). Wei et al. utilized generative adversarial networks to generate attribute-object compositions that match visual features, showcasing the potential of generative models in this domain (Wei et al., 2020).

Recent advancements have focused on improving the disentanglement of visual features to better represent attributes and objects. Atzmon et al. approached CZSL from a causal perspective, aiming to learn disentangled representations that can generalize well to unseen compositions (Atzmon et al., 2020; Liu & Ozay, 2023). Ruis et al. proposed learning prototypical representations of objects and attributes, highlighting the importance of prototype-based learning in CZSL (Ruis et al., 2021). Li et al. leveraged a Siamese contrastive space to disentangle visual features, subsequently entangling them with a generative model to enhance the learning process (Li et al., 2020). Marcus et al. extracted visual similarity from spatial features to disentangle attributes and objects, demonstrating the effectiveness of spatial feature extraction in improving CZSL models (Rohrbach, 2016). Furthermore, the shift from word compositionality to visual disentanglement has also been significant (Jiang, Ye, Wang, Shen, Zhang, & Zhang, 2024; Shuang et al., 2025). Zhang et al. treated CZSL as a domain generalization task, learning attribute- and object-invariant domains to improve generalization to unseen compositions (Hu, Zhao, Peng, & Gu, 2022; Panda & Mukherjee, 2024). This approach underscores the importance of domain invariance in achieving robust CZSL performance. In addition, the open-world setting of CZSL, where all possible compositions are considered during testing, has gained attention. Studies by Mancini et al. and Nayak et al. have explored this setting, proposing models that can handle the vast compositional space without prior knowledge of test-time compositions (Dong, Fu, Hwang, Sigal, & Xue, 2022; Mancini et al., 2021; Nayak, Yu, & Bach, 2023).

Notably, CLIP-based prompt engineering (Radford et al., 2021) enables scalable compositional reasoning through learnable soft prompts (Nayak et al., 2023), demonstrating the efficacy of vision-language synergy. Emerging techniques refine this paradigm.

ASP's attention mechanisms (Munir et al., 2024) improve attribute localization but inherit spatial constraints from frozen backbones, while OWC's unified framework (Jayasekara et al., 2025) introduces heuristic filtering for open-world feasibility. Concurrently, ProCC (Huo et al., 2024) and CatCom (Chytas et al., 2025) explore cross-primitive compatibility and categorical formalisms, yet lack pixel-level adaptability and low-level feature integration, respectively. These collective efforts highlight CZSL's evolving landscape, where cross-domain integration drives performance gains.

In summary, the field of CZSL has evolved from basic visual attribute learning to sophisticated models that disentangle visual features and leverage generative models. The transition from closed-world to open-world settings has further challenged researchers to develop more robust and generalizable models. As CZSL continues to progress, the integration of new methodologies and the exploration of novel settings will likely drive future advancements in this domain.

2.2. Superpixel algorithms

Superpixel algorithms have been extensively studied and can be broadly categorized into several approaches, including graph-based methods, clustering-based methods, and others such as watershed transform, geometric flows, and mean-shift.

Graph-based superpixel methods represent images as graphs where pixels form nodes connected by similarity edges. The normalized cuts algorithm minimizes cut values during partitioning (Felzenszwalb & Huttenlocher, 2004; Shi & Malik, 2000). However, their non-differentiable discrete optimizations hinder deep learning integration.

Clustering-based approaches primarily adapt k-means variants. SLIC (Achanta et al., 2012) generates uniform superpixels using spatial-color features (Li & Chen, 2015). Recent deep learning advancements enable end-to-end solutions like Superpixel Sampling Networks (SSN) (Jampani, Sun, Liu, Yang, & Kautz, 2018), which learn task-specific superpixels through differentiable loss functions.

Superpixels benefit vision tasks by providing structured image representations. They reduce search spaces in semantic segmentation while improving boundary accuracy. As detection proposals, they enhance object localization. Crucially, they suppress noise while better capturing local structures during feature extraction.

In the context of compositional zero-shot learning, the model is challenged to identify and combine attribute-object pairs that were not explicitly present in the training data. The advantage of superpixel technology lies in its ability to assist the model in more accurately locating and identifying visual features related to attributes and objects. For instance, when it comes to identifying the combination of "red car," superpixel segmentation can highlight the red regions in the image and combine them with the shape features of the car, thereby helping the model comprehend the association between the red attribute and the car object. In addition, superpixel segmentation reduces noise interference, enabling the model to focus on key features and consequently improving its recognition ability and generalization performance for unseen combinations.

3. Preliminary

3.1. Compositional zero-shot learning

Let us denote by $\mathcal{S}$ the set of possible states and by $\mathcal{O}$ the set of possible objects in the domain of interest. The compositional space $\mathcal{C}$ is defined as the Cartesian product of states and objects, i.e., $\mathcal{C} = \mathcal{S} \times \mathcal{O}$, representing all potential state-object compositions. Given a training set $\mathcal{T} = \{(x_i, c_i)\}_{i=1}^{N}$, where $x_i \in \mathcal{X}$ represents an image and $c_i \in C_s$ is a composition with $C_s \subset \mathcal{C}$, the goal of CZSL is to learn a mapping $f : \mathcal{X} \rightarrow C_t$, where C_t is the target compositional space, which may include compositions not present in C_s (i.e., $\exists c \in C_t$ s.t. $c \notin C_s$).

Open-World and Closed-World Scenarios. In the closed-world setting (Mancini et al., 2021), the target compositional space C_t is a strict subset of $\mathcal{C}$, i.e., $C_s \subset C_t \subset \mathcal{C}$, where C_s is the subset of compositions seen during training. This scenario is characterized by $|C_t| \ll |\mathcal{C}|$, indicating that the number of unseen compositions is typically less than the number of seen compositions, posing a relatively constrained generalization challenge.

In contrast, in the open-world setting (Mancini et al., 2021), as an extension of the CZSL problem, the output space of the model encompasses the entire set of possible compositions, $C_t \equiv \mathcal{C}$. This presents a significantly more complex challenge due to two main factors: (i) $|C_t| \gg |C_s|$, implying a vast increase in the number of potential unseen compositions, and (ii) the presence of a multitude of distractor compositions that are not present in the actual test set, complicating the discrimination of valid unseen compositions.

The objective in both settings is to learn a mapping $f : \mathcal{X} \rightarrow C_t$ from the input image space $\mathcal{X}$ to the target compositional space C_t. However, the open-world setting requires the model to not only generalize from seen to unseen compositions but also to discern and filter out implausible compositions from a much larger and less constrained output space.

3.2. Superpixel segmentation

Let $\mathcal{I}$ be an input image with a set of pixels $\mathcal{P} = \{p_1, p_2, \ldots, p_N\}$, where N is the total number of pixels. The objective of superpixel segmentation within the context of neural networks is to learn a mapping $\mathcal{F}$, parameterized by θ, that assigns each pixel to a superpixel in a manner that optimizes certain criteria reflective of the segmentation quality. Specifically the neural network learns a mapping $\mathcal{F}_\theta : \mathcal{P} \rightarrow \mathcal{S}$, where $\mathcal{S} = \{s_1, s_2, \ldots, s_M\}$ is the set of superpixels and M is the desired number of superpixels. The mapping is represented by an association matrix $Q \in \mathbb{R}^{N \times M}$, where each entry q_{nm} denotes the strength of association between pixel p_n and superpixel s_m. During the training Process, the network is trained to minimize a loss function $\mathcal{L}(\theta)$, which quantifies the discrepancy between the predicted association matrix $\hat{Q} = \mathcal{F}_\theta(\mathcal{P})$ and the ground-truth segmentation Q^*. The loss function encapsulates the criteria for homogeneity, compactness, and boundary preservation. Throughout the whole process, the network must ensure that for each pixel p_n, the sum of associations across all superpixels equals one, i.e., $\sum_{m=1}^{M} q_{nm} = 1$, reflecting a probabilistic assignment of pixels to superpixels (see Fig. 2).

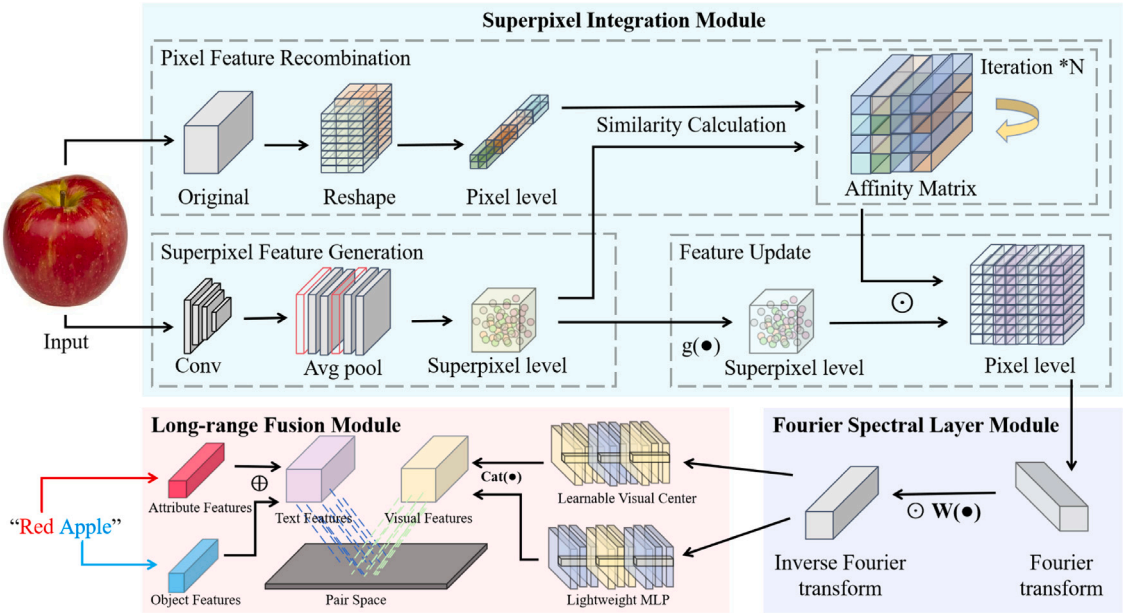

Fig. 2. The framework of the proposed SVFE model. SVFE integrates superpixel segmentation for refining image primitives, a Fourier spectral layer for extracting global features, and a long-range module for merging local and global features.

Table 1
Notations and descriptions.

Notation	Description	Notation	Description
I	Input image	S_i	The ith superpixel
N	Number of superpixels	f_p	Feature vector of pixel p
$\hat{f}_i$	Feature representation of superpixel S_i	$\tilde{f}_p$	Recombined pixel-level feature
$\mathbf{X}$	Input feature tensor (Fusion module)	$\mathbf{x}$	Input feature tensor (Fourier layer)
A	Affinity matrix between superpixels	T	Total number of iterations
$\mathbf{Q}^*$	Ground-truth association matrix	$\hat{\mathbf{Q}}$	Predicted association matrix
M, N	Dimensions of input tensor	$\mathbf{W}$	Learnable complex weight matrix
$\mathcal{F}$	2D Fourier transform	$\mathcal{F}^{-1}$	Inverse Fourier transform
b	Batch size	c	Number of channels
h, w	Spatial dimensions	X_{lvc}	Output of Local Visual Center
R	Reshaping function	γ	Channel-wise importance weights
$\mathbf{v}$	Normalized visual feature matrix	$\mathbf{T}$	Text feature matrix
α	Scaling factor ($\exp(s)$)	$\mathcal{L}$	Loss function

4. Methodology

4.1. Superpixel integration

In this section, we present a novel superpixel attention mechanism customized for the compositional zero-shot learning task. The proposed superpixel integration approach endeavors to exploit the synthesis of novel concepts from familiar states and objects, acquiring knowledge of visible compositional concepts (state and object) during training, and discerning unseen compositional concepts during inference, mirroring human cognitive processes.

The proposed superpixel integration strategy fundamentally differs from the Superpixel Sampling Network (SSN) (Jampani et al., 2018) by extending beyond mere superpixel generation to emphasize feature extraction and utilization. While SSN primarily clusters deep network features to create superpixels, based on the SSN, our approach employs average pooling to aggregate pixel-level features into superpixel-level representations, preserving local image information. Crucially, we introduce a novel mechanism for pixel feature recombination and iterative updating, where an affinity matrix characterizes superpixel relationships and reconstructs pixel-level features. This iterative refinement of representations at both levels enables a more nuanced understanding of image structures, enhancing precision and granularity in feature extraction for compositional zero-shot learning. The superpixel module, built upon SSN, is integrated into SVFE's end-to-end training pipeline. All parameters are optimized jointly with the Fourier spectral layer and long-range fusion module using a unified loss function.

The primary steps of the superpixel integration approach include the superpixel feature generation, pixel feature recombination, and iterative update. For clarity, all symbols used in this section are summarized in Table 1.

4.1.1. Superpixel feature generation

During the superpixel feature generation stage, the goal is to convert pixel-level features into superpixel-level features to comprehensively encapsulate local information within the image. This phase encompasses two critical steps, i.e., the superpixel segmentation step and the feature aggregation step.

The module firstly executes superpixel segmentation on the input image I, partitioning it into a set of superpixels $\{S_i\}_{i=1}^{N}$, where N represents the number of superpixels. Specifically, the module employs a three-layer convolutional encoder: The first layer consists of a 3×3 convolution with stride 1 and padding 1, processing the input's 3 channels to produce 64 output channels, followed by ReLU activation; subsequently, the second layer applies another 3×3 convolution (stride 1, padding 1) that transforms the 64 input channels into 128 output channels, also activated by ReLU; finally, the third layer executes a 1×1 convolution to generate M output channels corresponding to the target number of superpixels. For each superpixel S_i, it aggregates its internal pixel features to derive the superpixel-level feature representation $\hat{f}_i$. Specifically, the module computes the superpixel features via average pooling, which entails averaging the feature vectors of all pixels within the superpixel. This process can be mathematically expressed as:

$$\hat{f}_i = \frac{1}{|S_i|} \sum_{p \in S_i} f_p, \tag{1}$$

where f_p denotes the feature vector of pixel p, and $|S_i|$ signifies the number of pixels in superpixel S_i.

Through superpixel feature generation, the module transforms the pixel-level features of the original image into superpixel-level feature representations, adeptly capturing local information within the image. This transformation facilitates the provision of more effective inputs for subsequent tasks.

4.1.2. Pixel feature recombination

In the pixel feature recombination stage, the module amalgamates the superpixel-level feature representations into pixel-level feature representations for further processing within the model. This phase involves two main steps: computing the affinity matrix and recombining features.

The module computes the affinity matrix between superpixels to delineate their relationships. We define an affinity matrix A, where each element A_{ij} represents the degree of association between superpixels S_i and S_j. This association can be computed based on features such as spatial position, color, and texture of the superpixels. Typically, it employs a Gaussian kernel function to compute the similarity between two superpixels, expressed as:

$$A_{ij} = \exp\left(-\frac{d(S_i, S_j)^2}{2\sigma^2}\right), \tag{2}$$

where $d(S_i, S_j)$ signifies the distance between superpixels S_i and S_j, which can be spatial distance, color distance, or texture distance, and σ represents the bandwidth parameter of the Gaussian kernel.

Subsequently, it utilizes the affinity matrix A to reassemble the superpixel-level features to derive pixel-level feature representations. Specifically, for each pixel p, we weight-sum the feature vectors of all superpixels using the affinity matrix A to derive the pixel-level feature representation $\tilde{f}_p$, as follows:

$$\tilde{f}_p = \sum_{j=1}^{N} A_{ij} \hat{f}_j, \tag{3}$$

where N denotes the number of superpixels, $\hat{f}_j$ signifies the feature vector of superpixel S_j, and A_{ij} represents the element of the affinity matrix A.

To ensure the dimensional consistency between pixel-level and superpixel-level features during the similarity calculation, the module employs unfolding and folding operations. The unfolding operation involves sliding a 3×3 window over the pixel-level feature map with a stride of 1 and padding of 1. This process converts the spatial dimensions of the pixel-level features into a channel dimension, effectively transforming the pixel-level features into a format that matches the dimensions of the superpixel-level features. Specifically, for an input feature map of size [B, C, H, W], where B is the batch size, C is the number of channels, and H and W are the spatial dimensions, the unfolding operation reshapes it into [B, C*9, H*W], where 9 corresponds to the 3×3 window size. This allows the pixel-level features to be compatible with the superpixel-level features for the calculation of their pairwise similarities. After computing the affinity matrix and performing feature recombination, the folding operation is applied as the transpose of the unfolding process. It restores the features to their original spatial dimensions by converting the channel dimension back into spatial dimensions, resulting in a feature map of size [B, C, H, W]. This meticulous process ensures that the pixel-level features are appropriately aligned and compatible with the superpixel-level features throughout the model's processing pipeline.

Through pixel feature recombination, we amalgamate the superpixel-level feature representations into pixel-level feature representations, providing finer-grained feature information for subsequent tasks and contributing to the enhancement of the model's performance.

4.1.3. Iterative update

In the iterative update stage, the module iteratively refines the pixel-level feature representations through multiple iterations of computation. This process encompasses two critical steps: iterative computation and feature update.

Iterative Computation. The module conducts multiple iterations of computation, wherein each iteration refines the pixel-level feature representations. Assuming T iterations of computation, in the tth iteration, it leverages the current pixel-level feature representations $\tilde{f}_p^{(t-1)}$ for computation and derive the updated pixel-level feature representations $\tilde{f}_p^{(t)}$. This process can be represented as:

$$\tilde{f}_p^{(t)} = F(\tilde{f}_p^{(t-1)}), \tag{4}$$

where $F(\cdot)$ represents the iterative update function, which accepts the current pixel-level feature representations as input and yields the updated pixel-level feature representations as output.

Feature Update. The module performs the feature update step, wherein it utilizes the newly computed pixel-level feature representations to update the original superpixel-level feature representations. This process can be represented as follows:

$$\hat{f}_i^{(t)} = G(\tilde{f}_p^{(t)}), \tag{5}$$

where $G(\cdot)$ represents the feature update function, which accepts the updated pixel-level feature representations as input and yields the updated superpixel-level feature representations as output. The function $g(\cdot)$ is constructed as a composite function that includes a convolutional operation, batch normalization, and a non-linear activation function. Specifically, the pixel-level features are first processed by a convolutional layer with a kernel size of 3×3, which aggregates local information from the pixel-level features. Then batch normalization is applied to normalize the feature maps, ensuring that the mean and variance of the features are stabilized across different batches. Finally, a ReLU activation function is applied element-wise to introduce non-linearity into the model, enhancing its ability to learn complex feature representations.

The "Iteration *N" in the Affinity Matrix section refers to the same iterative process described here. The number of iterations is controlled by the parameter n_{iter}, which determines how many times the model refines the feature representations. In our implementation, n_{iter} is set to 3 to balance refinement effectiveness and computational efficiency.

Via iterative updates, the model systematically refines both pixel-level and superpixel-level feature representations, thereby augmenting the model's performance and accuracy. This iterative refinement maximizes the utilization of information at both superpixel and pixel levels, facilitating the acquisition of richer and more precise feature representations. Ultimately, we integrate the updated superpixel features with the transpose of the affinity matrix to reconstruct the pixel features, organizing the outcome into the original input feature map format. Subsequently, we yield the processed pixel features as the final output result.

Critically, all superpixel module parameters underwent systematic optimization through extensive experimental trials. Ablation studies across both closed-world and open-world settings quantified the impact of parameter variations. Finalized configurations (e.g., iteration count = 3) achieved an optimal accuracy-efficiency balance on the benchmark datasets.

4.1.4. Loss function

In the proposed SVFE model, training the superpixel module requires a well-designed loss function to ensure high-quality superpixel segmentation and effective image feature representation. This loss function comprises two main components: the superpixel segmentation loss and the feature reconstruction loss.

The superpixel segmentation loss aims to optimize the quality of superpixel segmentation by accurately assigning each pixel to the corresponding superpixel. We utilize the cross-entropy loss to measure the discrepancy between the predicted association matrix and the ground-truth association matrix.

Let the ground-truth association matrix be denoted as $\mathbf{Q}^*$, and the predicted association matrix as $\hat{\mathbf{Q}}$. The superpixel segmentation loss is defined as follows:

$$\mathcal{L}_{\text{seg}} = -\sum_{n=1}^{N}\sum_{m=1}^{M} Q_{nm}^* \log(\hat{Q}_{nm}) \tag{6}$$

Here, N represents the total number of pixels in the image, M is the number of superpixels, Q_{nm}^* indicates the probability that pixel n belongs to superpixel m in the ground truth, and $\hat{Q}_{nm}$ is the probability predicted by the model.

The feature reconstruction loss ensures that the reconstructed pixel-level features are as close as possible to the original pixel-level features. We employ the mean squared error loss to quantify the difference between the original and reconstructed pixel-level features.

Let the original pixel-level features be denoted as $\mathbf{g}_p$, and the reconstructed pixel-level features as $\tilde{\mathbf{g}}_p$. The feature reconstruction loss is defined as:

$$\mathcal{L}_{\text{rec}} = \frac{1}{N}\sum_{p=1}^{N} \|\mathbf{g}_p - \tilde{\mathbf{g}}_p\|^2 \tag{7}$$

Here, N is the total number of pixels, and $\|\cdot\|$ denotes the Euclidean norm.

4.2. Fourier spectral layer

To enhance the representational capabilities of vision transformers for CZSL, we introduce a novel Fourier spectral layer designed to leverage the power of the Fast Fourier Transform (FFT) for manipulating the frequency components of image features. This method

is meticulously crafted to refine the model's ability to discern novel compositions. Here is a detailed description of the approach:

The process begins with applying a two-dimensional FFT to the input feature tensor x. The FFT is mathematically represented by the following operation:

$$\mathcal{F}(x)_{(u,v)} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} x_{(m,n)} \cdot e^{-2\pi i \left(\frac{um}{M} + \frac{vn}{N}\right)}, \tag{8}$$

where $\mathcal{F}$ denotes the two-dimensional Fourier transform operator. This operator transforms the spatial domain data into the frequency domain, enabling the separation of the input signal into its constituent frequencies. M and N are the dimensions of the input tensor, and (u, v) represent the frequency indices. $x_{(m,n)}$ denotes the pixel value or feature response at position (m, n) in the spatial domain of the input feature tensor. Here, m and n are the indices for rows and columns, respectively, defining the spatial location of each pixel in the image. This equation shows that each point in the frequency domain is a result of a weighted sum of all points in the spatial domain, with the weights being complex exponentials that oscillate at different frequencies.

After the transformation, a learnable complex weight matrix $\mathbf{W}$ is introduced to modulate the frequency components. This matrix plays a pivotal role in adjusting the frequency spectrum, enabling the model to focus on or downplay specific frequencies to better represent the key features. The modulation process is executed through an element-wise multiplication between the frequency components obtained from the transformation and the complex weight matrix $\mathbf{W}$. The operation is denoted by the element-wise product symbol $\odot$, which implies that each frequency component is individually scaled by the corresponding weight in $\mathbf{W}$. The initialization of $\mathbf{W}$ is drawn from a normal distribution, and it is carefully scaled to ensure that the modulation is nuanced, allowing for subtle yet effective adjustments to the frequency components. As training progresses, the complex weights within $\mathbf{W}$ are fine-tuned through backpropagation. This training mechanism allows the model to identify and adapt to the most important frequency components for capturing salient features within the spectrum. By emphasizing relevant frequencies and suppressing less important ones, the model enhances its capacity to represent and distinguish compositional structures in images.

Finally, to revert the transformed features back into the spatial domain, the inverse FFT is applied:

$$\mathcal{F}^{-1}(x)_{(m,n)} = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} x_{(u,v)} \cdot e^{2\pi i \left(\frac{um}{M} + \frac{vn}{N}\right)}, \tag{9}$$

where $\mathcal{F}^{-1}$ represents the inverse Fourier transform operator. This operator effectively reverses the Fourier transform process, converting the frequency domain representation back into the spatial domain. The exponential term serves as the complex conjugate of that used in the forward transform, ensuring that the original spatial domain signal is accurately reconstructed.

The designed Fourier layer transforms features into the frequency domain, enabling access and manipulation of global frequency components. This is highly beneficial for capturing long-range dependencies and contextual information that may be spread across the image. Compared to large spatial domain convolutions, which need many parameters and computations for a similar receptive field, the Fourier layer does this more efficiently, with fewer parameters and lower computational cost. Moreover, the Fourier layer allows dynamic adjustment of component contributions through a learnable complex weight matrix. This flexibility lets the model emphasize or suppress specific frequencies as needed. For example, it can amplify high-frequency components for fine details or prioritize low-frequency components for overall image structure. Large convolutions in the spatial domain typically cannot match this adaptability, as they have fixed kernel weights and less flexibility in adjusting component contributions. The Fourier layer also offers a unique perspective for feature representation by operating in the frequency domain. It provides an additional dimension of information not directly available in the spatial domain, leading to more comprehensive and nuanced feature representations. In contrast, large spatial domain convolutions are limited to capturing features based on spatial patterns and may not fully utilize frequency domain information.

4.3. Long-range fusion

This module is designed to effectively integrate global and local feature representations, addressing the challenges of feature fusion in CZSL. As shown in Fig. 3, the module comprises the local visual center (LVC) and the light multi-layer perceptron (LightMLP), culminating in a feature fusion step that combines these local and global insights.

4.3.1. Local visual center

The local visual center module orchestrates a hierarchical pipeline to extract fine-grained local features critical for attribute-object composition recognition. The process begins with channel compression, where a 1×1 convolutional layer reduces the input feature tensor $\mathbf{X} \in \mathbb{R}^{b \times c_i \times h \times w}$ (with b as batch size, c_i as input channels, and $h \times w$ as spatial dimensions) to a compact representation, stabilized by batch normalization and activated via ReLU. This step preserves spatial context while eliminating redundant channel information.

Spatial feature enrichment follows through a 3×3 convolutional layer, which captures local structural patterns (e.g., edges, textures) within the compressed features. The resulting spatial descriptors are normalized and activated to retain discriminative details, then projected back to the original channel dimension via another 3×3 convolution for channel restoration, balancing spatial precision with feature richness.

The core of LVC lies in adaptive feature modulation. Global spatial information of the restored features $\mathbf{X}_e$ is aggregated via global average pooling and encoded into a latent embedding, which is then transformed by a fully connected layer to generate

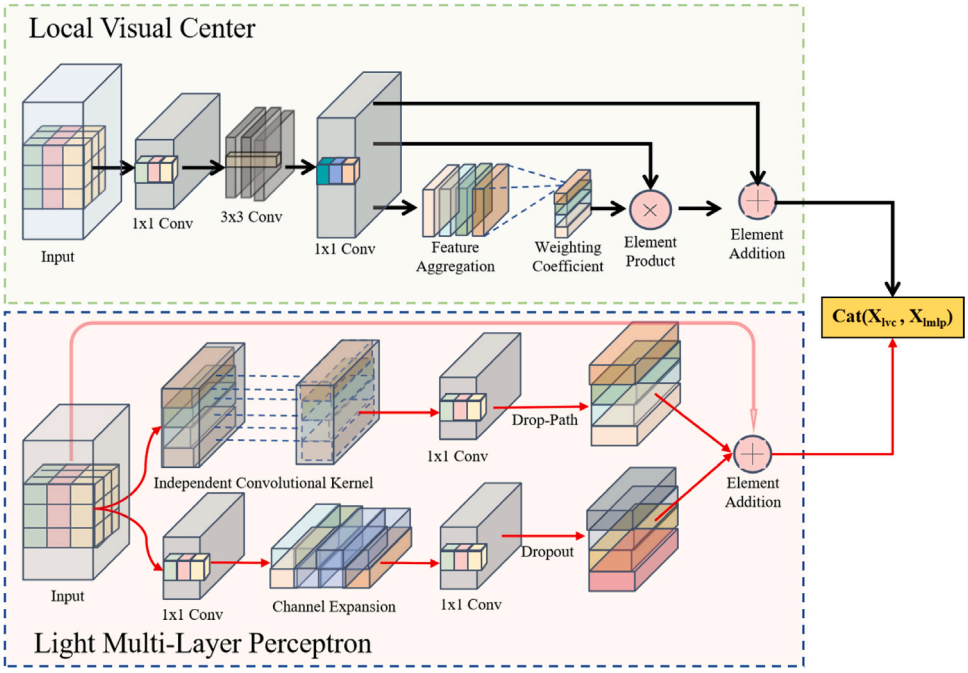

Fig. 3. Architecture of the proposed local visual center module and light multi-layer perceptron module.

channel-wise importance weights $\boldsymbol{\gamma} \in \mathbb{R}^{c_i}$. These weights are reshaped by a function $\mathcal{R}$ to match the tensor dimensions of $\mathbf{X}_e$, enabling gated residual fusion:

$$\mathbf{X}_{\text{lvc}} = \mathbf{X}_e + \mathbf{X}_e \odot \mathcal{R}(\boldsymbol{\gamma}) \tag{10}$$

Here, $\odot$ denotes element product, and $\mathcal{R}$ reshapes $\boldsymbol{\gamma}$ to $(b, c_i, 1, 1)$. This operation selectively amplifies attribute-sensitive channels while preserving spatial structure, ensuring robust discrimination of complex compositional patterns in zero-shot scenarios. The output $\mathbf{X}_{\text{lvc}}$ integrates local detail with adaptive channel weighting, forming an optimized representation for compositional reasoning.

4.3.2. Light multi-layer perceptron

The light multi-layer perceptron module is designed to capture long-range dependencies in feature maps while maintaining computational efficiency, integrating local feature extraction with global context modeling. The module processes input features $\mathbf{X} \in \mathbb{R}^{b\times c\times h\times w}$, where b denotes the batch size, c is the number of input channels, and $h \times w$ represents the spatial dimensions.

Branch one employs depthwise separable convolutions to extract fine-grained local features. First, a depthwise convolution (dconv) processes each channel independently using a $k \times k$ kernel, capturing per-channel spatial patterns while preserving the channel dimension. This is followed by a pointwise convolution (pconv) with a 1×1 kernel, which integrates channel information and projects the features to the desired output dimension. The resulting output of Branch 1 is denoted as $\mathbf{Y}_p$, and a drop-path layer is applied to stochastically skip this branch during training for regularization.

Branch two leverages MLP operations to model global dependencies. A 1×1 convolution first expands the channel dimension to a hidden size c_{hidden}, enabling feature abstraction, followed by another 1×1 convolution to reduce the channels back to c. This two-stage transformation, denoted as $\mathbf{Z}_2$, is complemented by a dropout layer to mitigate overfitting.

The final output of the LightMLP module fuses the original input features with the processed outputs from both branches via a residual connection:

$$\mathbf{X}_{\text{lmlp}} = \mathbf{X} + \text{DropPath}(\mathbf{Y}_p) + \text{Dropout}(\mathbf{Z}_2) \tag{11}$$

Here, DropPath$(\cdot)$ and Dropout$(\cdot)$ denote the drop-path and dropout regularization operations, respectively. This formulation ensures that the model balances local spatial details (captured by Branch 1) with global contextual relationships (modeled by Branch 2), enabling effective recognition of long-range attribute-object dependencies in compositional zero-shot learning scenarios. The lightweight architecture of LightMLP, combining depthwise separable convolutions and linear MLP layers, ensures efficient computation while maintaining representational power.

Table 2
Statistics of the public CZSL benchmark datasets. The datasets are summarized in terms of the number of state and object concepts, as well as the number of images across training, validation, and test splits. s and o denote the number of state and object concepts, and i represents the number of images. C_s and C_u are the pair concepts of seen and unseen classes.

			Train		Validation			Test		
Dataset	s	o	c_s	i	c_s	c_u	i	c_s	c_u	i
MIT-States	115	245	1262	30 338	300	300	10 420	400	400	12 995
UT-Zappos	16	12	83	22 998	15	15	3214	18	18	2914

4.3.3. Text feature extraction

In the SVFE model, text feature extraction is achieved through a customized text encoder, which is based on the pre-trained CLIP text encoder architecture and optimized for compositional zero-shot learning tasks. First, we construct a soft prompt template "a photo of [ATTR] [OBJ]", where [ATTR] and [OBJ] are placeholders for attributes and objects. For each attribute and object word, we obtain its word vector representation through the token embedding layer of CLIP, and calculate the mean of all token vectors from the start token to the EOS token as the feature embedding for the attribute/object. Meanwhile, the context part "a photo of" is replaced with learnable context vectors, soft-prompt, which are dynamically optimized during training to capture task-specific semantic patterns.

During the model's forward pass, given an attribute-object pair, we generate the corresponding token tensor. This involves placing the learnable context vectors in the corresponding positions of the prompt and filling the placeholder positions with the precomputed attribute and object embeddings. This forms a complete representation of the compositional concept. This token tensor is then input into the text encoder, sequentially passing through the token embedding layer, positional embeddings, Transformer encoder layers, and layer normalization. Finally, the vector at the EOS token position is extracted as the text feature representation and mapped to a joint embedding space aligned with visual features through the text projection layer.

To enhance the model's ability to disentangle compositional relationships, we perform a decomposition operation on the text features. According to the indices of the attribute-object pairs, the text features are separated into independent attribute and object feature matrices. Through average pooling, we aggregate the feature representations of the same attributes/objects. This structured text representation allows the model to explicitly model the compositional relationships between attributes and objects, providing fine-grained semantic guidance for subsequent visual-semantic alignment.

4.3.4. Feature fusion

The fusion of local and global features is achieved through concatenation followed by a refining convolutional layer, enhancing the model's ability to extract comprehensive patterns from the fused representation.

$$x_{fused} = \mathrm{Conv}\left(\mathrm{Cat}\left(x_{lvc}, x_{lmlp}\right)\right), \tag{12}$$

where $\mathrm{Cat}(\cdot)$ performs the concatenation along the channel dimension, merging the local features x_{lvc} with the global features x_{lmlp}. The function $\mathrm{Conv}(\cdot)$ represents a convolutional layer that integrates these features, allowing the model to leverage both local detail and global context for improved pattern recognition in CZSL tasks.

The LVC and LightMLP both aim to enhance feature representation but differ in their mechanisms and feature emphasis. The LVC extracts fine-grained local features using convolutional and pooling operations. It captures detailed visual information within localized regions by employing multiple convolutional layers followed by pooling operations, which progressively reduce spatial dimensions while preserving and enhancing local feature representations. Meanwhile, the LightMLP focuses on capturing long-range dependencies across the entire feature map. It uses fully connected layers to integrate contextual information from all parts of the feature map, providing a holistic understanding of the image context. When compared to multi-scale methods that process and combine features at various scales, our Long-range Fusion approach offers a more effective integration of local and global features. Multi-scale methods often rely on simple concatenation for fusion, which may not fully utilize the complementary nature of local and global information. Our method combines the LVC and LightMLP modules, leveraging their complementary strengths. The LVC provides detailed local features, while the LightMLP offers global context. By concatenating and refining these features through a convolutional layer, our approach achieves a more synergistic fusion of local and global information. This enhances the model's ability to recognize novel compositions in Compositional Zero-Shot Learning.

In the SVFE model, both text and visual features are directly mapped to the pair space through similarity calculations, enabling the establishment of geometric correspondences between images and compositional concepts within a unified embedding framework. The core formula driving this mapping is:

$$\text{logits} = \alpha \cdot \left(\frac{\mathbf{v}}{\|\mathbf{v}\|_2}\right)\left(\frac{\mathbf{T}}{\|\mathbf{T}\|_2}\right)^{\top} \tag{13}$$

Here, $\mathbf{v} \in \mathbb{R}^{B\times d}$ denotes the normalized visual feature matrix for a batch of B visual inputs, with ℓ_2 normalization ensuring each vector has a unit length. The text feature matrix $\mathbf{T} \in \mathbb{R}^{P\times d}$ corresponds to P attribute-object pairs, also normalized. The scaling factor $\alpha = \exp(s)$, derived from CLIP's logit-scale parameter s, dynamically adjusts the range of similarity scores. Each element logits_{ij} in the output matrix $\text{logits} \in \mathbb{R}^{B\times P}$ quantifies the match between the ith image and the jth attribute-object pair.

This mechanism computes cosine similarities via vector dot products. The visual features **v** are a weighted combination of original CLIP features (weighted by ω) and superpixel-enhanced features (weighted by $1-\omega$). Similarly, the text features **T** integrate soft-prompt-optimized context vectors with disentangled attribute/object embeddings. By constructing this joint embedding space, the model effectively captures cross-modal semantic relationships through high-dimensional dot product operations, providing a robust foundation for discriminative similarity measurement in open-world compositional zero-shot recognition.

5. Experiments

5.1. Experimental setup

5.1.1. Datasets

To evaluate the effectiveness of the proposed SVFE model, we conduct extensive experiments on two CZSL benchmark datasets: MIT-States (Isola, Lim, & Adelson, 2015) and UT-Zappos (Yu & Grauman, 2014, 2017). These datasets encompass a wide range of object-state combinations, providing a robust evaluation of models' abilities to generalize to both seen and unseen compositions. The MIT-States dataset, introduced by Isola et al. comprises 59,000 images with 245 objects and 115 states. In the closed-world settings, the search space contains 1262 seen compositions and 300 unseen for validation, and 400 unseen for testing. While in the open-world settings, it results in 28,175 unique object-state pairings. This dataset's rich variety and complexity serve as a critical benchmark for evaluating a model's generalization capabilities. Conversely, the UT-Zappos dataset, crafted by Yu and Grauman, focuses on attribute combinations within footwear, offering 50,025 images across four main categories with 12 objects and 16 states. For the closed-world experiments, it is constrained to the 83 seen and 15/18 (validation/test) unseen compositions. While in the open-world settings, it contains 192 compositions. Although it contains fewer object categories, the diversity of attributes provides an excellent basis for evaluating model performance on fine-grained attribute variations. To better understand the characteristics of these datasets, we provide a comparison in Table 2. By conducting experiments on these datasets, we can thoroughly evaluate the performance of models in various compositional tasks and verify their ability to generalize to diverse combinations.

5.1.2. Metrics

Following the setting of prior work (Mancini et al., 2021), we discuss the metrics employed to evaluate the performance of our CZSL model. The metrics used include Seen (S), Unseen (U), Harmonic Mean (H) of the S and U metrics, and Area Under the Curve (AUC). Each of these metrics provides a unique perspective on model performance, particularly in the context of generalizing to both seen and unseen object-state combinations. The Seen (S) metric evaluates the model's performance on object-state combinations that were present in the training set. The Unseen (U) metric measures the model's ability to recognize object-state combinations that it has never encountered during training. The Harmonic Mean (H) of the Seen and Unseen metrics provides a balanced evaluation of the model's performance across both seen and unseen combinations. The Area Under the Curve (AUC) provides a comprehensive view of the model's performance across different thresholds by evaluating the trade-off between precision and recall for recognizing object-state combinations. By employing these metrics, we ensure a thorough and balanced evaluation of our models.

5.1.3. Implementation details

The model has been meticulously implemented with PyTorch (Paszke et al., 2019), a dynamic and versatile deep learning framework. Our implementation parameters are based on the state-of-the-art Vision Transformer (ViT) architecture, specifically the "$ViT-L/14$" variant, which serves as the backbone for feature extraction, contributing 425 million parameters. The training regimen incorporates a learning rate of 0.0001, alongside weight decay set to 0.00001 to mitigate overfitting, with regularization through attribute dropout at 0.3. The training process is orchestrated with a batch size of 128, leveraging gradient accumulation over 2 steps for stable updates. For computational efficiency, each epoch requires approximately 5 min for UT-Zappos and 20 min for MIT-States on a single NVIDIA RTX 3090 GPU, with convergence achieved in 20 epochs. During evaluation, the system processes 100 images per second (0.01s per image) at a batch size of 128. The model's performance is evaluated using various metrics, with particular focus on "$best_unseen$" for novel composition generalization, augmented by attribute-object fusion and soft prompt weights ("att_obj_w", "sp_w"). Evaluation occurs in both closed and open-world settings (the latter with unconstrained test-time search space), implementing a feasibility threshold of 0.4 to filter implausible compositions.

5.1.4. Implementation details

The model has been meticulously implemented with PyTorch (Paszke et al., 2019), a dynamic and versatile deep learning framework. Our implementation parameters are based on the state-of-the-art Vision Transformer (ViT) architecture, specifically the "$ViT-L/14$" variant, which serves as the backbone for feature extraction, contributing 425 million parameters. The training regimen incorporates a learning rate of 0.0001, alongside weight decay set to 0.00001 to mitigate overfitting, with regularization through attribute dropout at 0.3. The training process is orchestrated with a batch size of 128, leveraging gradient accumulation over 2 steps for stable updates. For computational efficiency, each epoch requires approximately 5 min for UT-Zappos and 20 min for MIT-States on a single NVIDIA RTX 3090 GPU, with convergence achieved in 20 epochs. The proposed modules introduce negligible computational overhead, increasing training time by $< 0.5\%$ per epoch versus the ViT-L/14 baseline. During evaluation, the system processes 156 images per second at a batch size of 128. Inference speed remains at 100 images/second, maintained through optimized tensor operations and replacement of spatial convolutions with frequency-domain processing in the Fourier layer. The model's performance is evaluated using various metrics, with particular focus on "$best_unseen$" for novel composition generalization, augmented by attribute-object fusion and soft prompt weights ("att_obj_w", "sp_w"). Evaluation occurs in both closed and open-world settings (the latter with unconstrained test-time search space), implementing a feasibility threshold of 0.4 to filter implausible compositions.

Table 3
Closed-world CZSL results on MIT-States and UT-Zappos datasets. S and U are the accuracies on seen and unseen compositions, respectively. H is the harmonic mean of U and S, and AUC is the area under the curve. Boldface indicates the best. "-" indicates that no reported results are available.

Method	MIT-States				UT-Zappos			
	AUC	*H*	*S*	*U*	*AUC*	*H*	*S*	*U*
AoP (Nagarajan & Grauman, 2018)	1.6	9.9	14.3	17.4	25.9	40.8	59.8	54.2
LE+ (Naeem et al., 2021)	2.0	10.7	15.0	20.1	25.7	41.0	53.0	61.9
TMN (Purushwalkam et al., 2019)	2.9	13.0	20.2	20.1	29.3	45.0	58.7	60.0
SymNet (Li et al., 2020)	3.0	16.1	24.2	25.2	23.4	40.4	49.8	57.4
CompCos (Mancini et al., 2021)	4.5	16.4	25.3	24.6	28.1	43.1	59.8	62.5
CGE (Naeem et al., 2021)	5.1	17.2	28.7	25.3	26.4	41.2	56.8	63.6
Co-CGE (Mancini et al., 2022)	1.1	6.4	31.1	5.8	23.1	40.3	62.0	44.3
SCEN (Li, Yang, Wei, Deng, & Yang, 2022)	5.3	18.4	29.9	25.2	32.0	47.8	63.5	63.1
CSP (Nayak et al., 2023)	19.4	36.3	46.6	49.9	33.0	46.6	64.2	66.2
PromptCompVL (Xu et al., 2022)	18.3	35.3	48.5	47.2	32.2	46.1	64.4	64.0
DFSP (Lu et al., 2023)	20.8	37.7	47.1	**52.8**	33.5	47.1	63.3	**69.2**
FOMA (Dai et al., 2024)	–	–	–	–	33.1	47.3	60.3	68.0
DBC (Zhang et al., 2025)	–	–	–	–	30.9	45.8	60.1	63.7
CSP+TPT (Zhou & Ma, 2024)	20.1	37.1	47.1	50.7	33.5	47.3	64.4	67.0
SVFE(ours)	**21.4**	**38.1**	**49.2**	52.1	**34.6**	**48.1**	**64.4**	65.7

5.2. Comparison with state-of-the-arts

The proposed SVFE model demonstrates empirically validated superiority over existing approaches across both closed-world and open-world CZSL benchmarks, as detailed in Tables 2 and 3, directly overcoming critical limitations in prior methodologies. Early prototype methods (AoP Nagarajan & Grauman, 2018, LE+ Naeem, Xian, Tombari, & Akata, 2021, TMN Purushwalkam et al., 2019) constrain fine-grained attribute disentanglement by projecting primitives into fixed embedding spaces with holistic features that conflate local cues. While disentanglement and optimization methods (SymNet Li et al., 2020, CompCos Mancini et al., 2021, CGE Naeem et al., 2021, Co-CGE Mancini, Naeem, Xian, & Akata, 2022, DBC Zhang, Liang, Du, Chen, & Ma, 2025) explicitly separate attributes and objects using graphs or generative architectures, their rigid transformation rules struggle with non-uniform spatial configurations, and graph convolutions cause oversmoothing that blurs textures. Prompt-based and vision-language methods (CSP Nayak et al., 2023, PromptCompVL Xu, Kordjamshidi, & Chai, 2022, HPL Wang, Yang, Wei, & Deng, 2023, DFSP Lu, Guo, Liu, & Guo, 2023, LeMA Kim, Lee, & Choi, 2024, CSP+TPT Zhou & Ma, 2024, ASP Munir et al., 2024, OWC Jayasekara et al., 2025), though leveraging CLIP's alignment, remain spatially insensitive as frozen encoders cannot resolve misalignment between localized attributes and object anchors through prompt tuning alone. Image feature enhancement methods (GIPCOL Xu, Chai, & Kordjamshidi, 2024, FOMA Dai, Huang, Zhang, Gong, & Wang, 2024, ProCC Huo et al., 2024, CatCom Chytas et al., 2025) augment visual features via convolutions but fail to jointly capture pixel-level granularity and global frequency context without dynamic modulation.

In contrast, SVFE systematically bridges these gaps. Its superpixel integration enables finer attribute/object separation than disentanglement methods through iterative pixel-superpixel recombination. The Fourier layer dynamically amplifies discriminative frequencies, outperforming static spatial convolutions. The long-range fusion module mitigates oversmoothing by integrating localized superpixel details with global dependencies, ensuring precise attribute-object interplay.

The SVFE model has achieved the state-of-the-art performance in the closed-world setting, as evidenced by the experimental results detailed in Table 3. In the open-world setting, SVFE continues to demonstrate its superiority, as illustrated in Table 4. The model's performance in this scenario is particularly remarkable, given the vast and unbounded composition space it must navigate. SVFE's ability to maintain high accuracy and AUC in the face of such complexity is a testament to its advanced feature extraction and representation capabilities. Moreover, we can see that while the SVFE model shows a slight dip in performance for the Unseen metric compared to models like DFSP (Lu et al., 2023), this is a deliberate trade-off. The SVFE model is designed to excel at capturing comprehensive image features for seen compositions, which are more likely to be encountered in practical scenarios. This strategic focus enables our model to achieve optimal results in the AUC, HM, and Seen metrics, ensuring high reliability and accuracy in the majority of real-world applications. By prioritizing the representation of seen compositions, the model establishes a robust feature space that supports both strong performance on familiar compositions and effective generalization to novel ones, thus striking a balance that optimizes overall performance across the board.

5.3. Ablation study

The ablation study, meticulously conducted, aims to assess the distinct contributions of the three core components of our SVFE model: the superpixel integration strategy (SIS), the Fourier spectral layer (FSL), and the long-range fusion (LF) module. Through a series of controlled experiments, we elucidate the significance of each component and its collective synergistic impact on the SVFE model's overall performance. The results are detailed in Tables 5 and 6 for the MIT-States dataset under closed and open-world settings, respectively, and in Tables 7 and 8 for the UT-Zappos dataset under analogous conditions.

Table 4
Open-world CZSL results on MIT-States and UT-Zappos datasets. Boldface indicates the best. "-" indicates that no reported results are available.

Method	MIT-States				UT-Zappos			
	AUC	H	S	U	AUC	H	S	U
AoP (Nagarajan & Grauman, 2018)	0.7	4.7	16.6	5.7	13.7	29.4	50.9	34.2
LE+ (Naeem et al., 2021)	0.3	2.7	14.2	2.5	16.3	30.5	60.4	36.5
TMN (Purushwalkam et al., 2019)	0.1	1.2	12.6	0.9	8.4	21.7	55.9	18.1
SymNet (Li et al., 2020)	0.8	5.8	21.4	7.0	18.5	34.5	53.3	44.6
CompCos (Mancini et al., 2021)	1.6	8.9	25.4	10.0	21.3	36.9	59.3	46.8
CGE (Naeem et al., 2021)	1.0	6.0	32.4	5.1	23.1	39.0	61.7	47.7
Co-CGE$_{closed}$ (Mancini et al., 2022)	1.1	6.4	31.1	5.8	23.1	40.3	62.0	44.3
Co-CGE$_{open}$ (Mancini et al., 2022)	2.3	10.7	30.3	11.2	23.3	40.8	61.2	45.8
KG-SP (Karthik, Mancini, & Akata, 2022)	1.3	7.4	28.4	7.5	26.5	42.3	61.8	52.1
CSP (Nayak et al., 2023)	5.7	17.4	46.3	15.7	22.7	38.9	64.1	44.1
PromptCompVL (Xu et al., 2022)	6.1	17.7	48.5	16.0	21.6	37.1	64.6	44.0
HPL (Wang et al., 2023)	6.9	19.8	46.4	18.9	24.6	40.2	63.4	48.1
DFSP (Lu et al., 2023)	6.7	19.2	47.1	18.1	27.6	42.7	63.5	57.2
GIPCOL (Xu et al., 2024)	6.3	17.9	48.5	16.0	23.5	40.1	**65.0**	45.0
CSP+TPT (Zhou & Ma, 2024)	6.2	17.6	47.1	16.5	24.5	39.3	64.4	45.0
LeMA (Kim et al., 2024)	2.4	11.1	30.3	11.5	24.3	42.0	47.8	**61.3**
DBC (Zhang et al., 2025)	–	–	–	–	23.8	39.8	60.1	48.0
ProCC (Huo et al., 2024)	1.6	7.8	27.6	10.6	23.6	39.9	62.2	48.8
ASP (Munir et al., 2024)	1.4	7.7	27.1	8.4	25.9	43.1	61.0	48.6
CatCom (Chytas et al., 2025)	2.2	10.6	31.1	12.1	22.1	37.4	62.6	48.0
OWC (Jayasekara et al., 2025)	3.1	12.4	36.3	12.5	–	–	–	–
SVFE(ours)	**7.1**	**19.8**	**49.0**	**18.2**	**28.3**	**44.1**	64.5	55.0

Table 5
Results of ablation studies in the closed-world CZSL scenario on MIT-States dataset.

Method	AUC	HM	S	U
Backbone	19.7	36.5	48.1	49.6
$+SIS$	20.6 (↑ 0.9)	37.3 (↑ 0.8)	46.8 (↓ 1.3)	52.1 (↑ 2.5)
$+FSL$	20.4 (↑ 0.7)	37.2 (↑ 0.7)	46.6 (↓ 1.5)	51.9 (↑ 2.3)
$+LF$	21.0 (↑ 1.3)	37.6 (↑ 1.1)	48.0 (↓ 0.1)	52.1 (↑ 2.5)
$+SIS+FSL+LF$	21.4 (↑ 1.7)	38.1 (↑ 1.6)	49.2 (↑ 1.1)	52.1 (↑ 2.5)

Table 6
Results of ablation studies in the open-world CZSL scenario on MIT-States dataset.

Method	AUC	HM	S	U
Backbone	6.4	18.6	47.9	16.9
$+SIS$	6.8 (↑ 0.4)	19.5 (↑ 0.9)	46.7 (↓ 1.2)	18.6 (↑ 1.7)
$+FSL$	6.7 (↑ 0.3)	19.4 (↑ 0.8)	46.6 (↓ 1.3)	18.3 (↑ 1.4)
$+LF$	6.9 (↑ 0.5)	19.6 (↑ 1.0)	48.1 (↑ 0.2)	18.3 (↑ 1.4)
$+SIS+FSL+LF$	7.1 (↑ 0.7)	19.8 (↑ 1.2)	49.0 (↑ 1.1)	18.2 (↑ 1.3)

Table 7
Results of ablation studies in the closed-world CZSL scenario on UT-Zappos dataset.

Method	AUC	HM	S	U
Backbone	29.8	43.3	61.4	64.0
$+SIS$	31.0 (↑ 1.2)	44.9 (↑ 1.6)	61.9 (↑ 0.5)	65.4 (↑ 1.4)
$+FSL$	29.9 (↑ 0.1)	43.4 (↑ 0.1)	62.3 (↑ 0.9)	64.5 (↑ 0.5)
$+LF$	33.5 (↑ 3.7)	47.7 (↑ 4.4)	61.5 (↓ 0.1)	66.7 (↑ 2.7)
$+SIS+FSL+LF$	34.6 (↑ 4.8)	48.1 (↑ 4.8)	64.4 (↑ 3.0)	65.7 (↑ 1.7)

The superpixel integration strategy is pivotal in refining the granularity of feature extraction and capturing finer details within image data. Integrating superpixel technology allows the model to more effectively represent and disentangle visual concepts, thereby enhancing its discrimination and generalization capabilities. The SIS significantly bolsters the model's performance across various metrics, particularly evident in the substantial improvements in AUC and HM scores for both datasets under closed and open settings.

The Fourier spectral layer is introduced to capture global image features through the transformation of these features into the frequency domain. This innovative layer enables the model to dynamically modulate component contributions, thereby enhancing the representation of local details. Ablation studies consistently show improved performance metrics when the FSL is incorporated.

Table 8
Results of ablation studies in the open-world CZSL scenario on UT-Zappos dataset.

Method	AUC	HM	S	U
Backbone	23.4	39.8	61.4	48.3
$+SIS$	25.4 (↑ 2.0)	41.9 (↑ 2.1)	61.9 (↑ 0.5)	52.7 (↑ 4.4)
$+FSL$	24.9 (↑ 1.5)	40.8 (↑ 1.0)	62.2 (↑ 0.8)	52.8 (↑ 4.5)
$+LF$	26.1 (↑ 2.7)	42.5 (↑ 2.7)	61.5 (↓ 0.1)	52.4 (↑ 4.1)
$+SIS+FSL+LF$	28.3 (↑ 4.9)	44.1 (↑ 4.3)	64.5 (↑ 3.1)	55.0 (↑ 6.7)

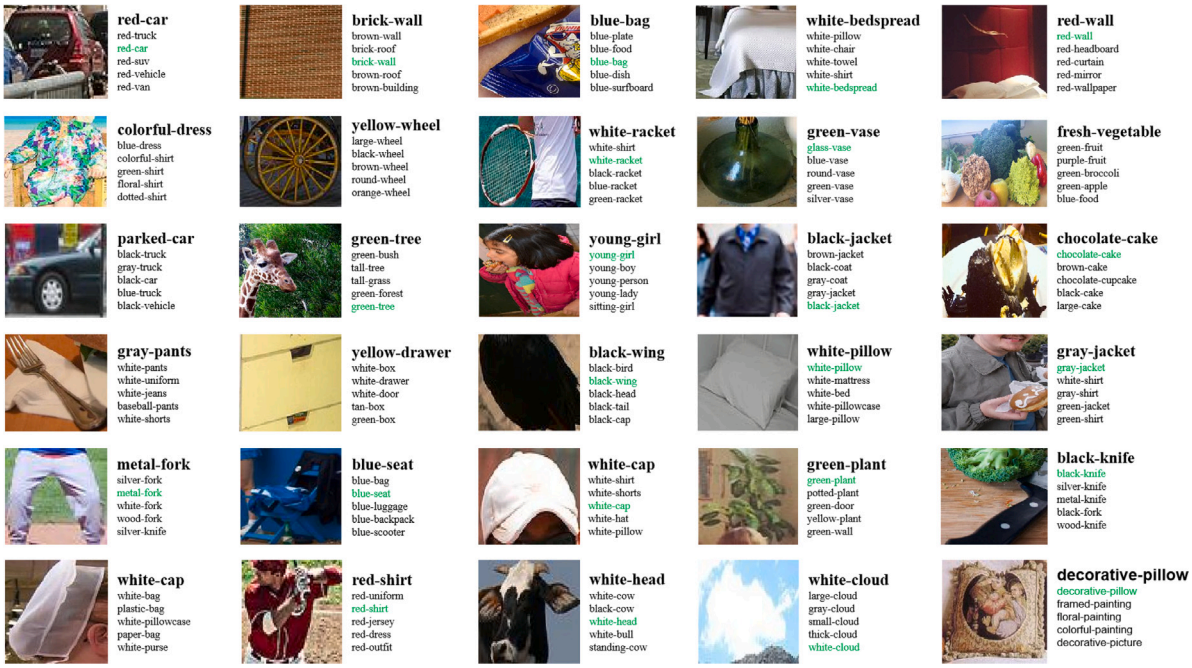

Fig. 4. Top-5 Image-to-Text Retrieval. For each image, the model retrieves the five closest matching textual descriptions, reflecting the visual content accurately.

The long-range fusion module is engineered to optimize the synergy between local and global features, significantly enhancing the model's ability to discern intricate compositional relationships. The substantial impact of the long-range fusion module on the model's performance is evident. Notably, the activation of the long-range fusion module correlates with enhanced S and U metrics, underscoring the model's improved capacity to recognize both seen and unseen compositions.

Upon integrating all three modules into the SVFE model, the collective effect yields a robust framework that excels in the challenging domain of compositional zero-shot learning. As evidenced in Tables 5, 6, 7, and 8, the integrated implementation of these modules achieves superior performance across all metrics, thereby solidifying the significance of each component within the overall framework. Regarding the suboptimal performance on certain unseen data, a plausible explanation is that our model excels at recognizing fine-grained image features, particularly for the categories it has seen. This heightened sensitivity to nuanced features enables the model to achieve higher accuracy for seen categories, given their thorough learning during the training phase. Conversely, the lack of training samples for unseen categories constrains the model's ability to fully capture their subtleties, leading to slightly inferior performance. The ablation study affirms the significant and complementary contributions of the three core components to our SVFE model's performance. Each component contributes unique advantages, and their integration culminates in a robust framework that excels in the challenging domain of Compositional Zero-Shot Learning. Future work will focus on optimizing computational efficiency and expanding the model's applicability, building on the robust foundation established by SVFE.

5.4. Qualitative analysis

The qualitative analysis of the SVFE model offers an insightful examination of its performance across various retrieval tasks, highlighting its strengths in compositional zero-shot learning scenarios.

Image-to-Text Retrieval. The image-to-text retrieval task highlights the SVFE model's interpretative capabilities by associating visual content with appropriate textual descriptions. As illustrated in Fig. 4, for the image "red car", the model successfully retrieved terms that captured the ‹red› attribute and related vehicle types, including "red-truck" "red-suv" and "red-van" alongside the exact

Fig. 5. Top-5 Text-to-Image Retrieval. For each textual query, the model retrieves the five most relevant images, demonstrating its ability to translate textual descriptions into visual representations.

Fig. 6. Visual Concept Retrieval. The model demonstrates the ability to identify and retrieve images based on shared visual concepts such as attributes or objects.

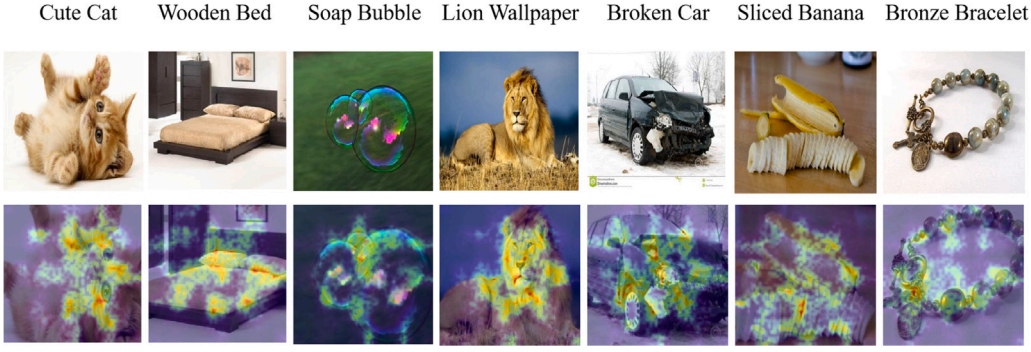

Fig. 7. Visualization of image feature extraction on MIT-States. The top row displays the original image, with its associated attributes and objects labeled above. The second row depicts the activation heatmaps generated by our SVFE model.

match “red-car”. This demonstrates the model’s proficiency in recognizing visual attributes and associating them with a diverse range of relevant objects. In the case of “fresh vegetable”, although the exact term was not retrieved, the model’s response with “green-fruit” and “green-broccoli” was contextually relevant, indicating its ability to generalize from visual cues to a broader semantic spectrum. The slight discrepancy in the second example may be attributed to the model’s broader interpretation of ‹freshness›, associated with the vibrant colors in the image, resulting in a slightly different but contextually fitting set of descriptions.

Text-to-Image Retrieval. In the text-to-image retrieval task, the SVFE model demonstrates exceptional accuracy in mapping textual descriptions to visual content. As shown in Fig. 5, when queried with the composite concept “Satin Sandals”, the model’s

top-4 retrieval results were strikingly consistent, all depicting “Satin Sandals” as requested. This high precision indicates the model’s proficiency in interpreting and prioritizing the integration of material and object attributes from text. The model’s ability to focus on the ‹satin› attribute and accurately pair it with ‹sandals› demonstrates its nuanced understanding of the query. However, the fifth result, “Leather Sandals”, while not an exact match, is semantically close, indicating the model’s capacity for generalization. This slight deviation may be due to the model’s attempt to find the closest match in the absence of an exact ‹satin› example, highlighting its robustness in handling variations.

Visual Concept Retrieval. The visual concept retrieval task demonstrates the SVFE model’s effectiveness in identifying shared attributes or objects across images. As depicted in Fig. 6, when retrieving for the concept of “green cup”, the model accurately listed images containing the ‹green› attribute across various items, such as vests, shirts, and walls. This indicates the model’s proficiency in recognizing and categorizing images by color attributes. However, when retrieving the ‹cup› object, the model’s precision was less consistent, yielding items such as “glass-container” and “white-cup”. This imprecision may stem from the variability in the shape and material of cups, which may not be sufficiently distinctive in the visual data for accurate disambiguation. The model’s challenge in differentiating similar objects underscores the need for a more nuanced approach to object recognition, possibly involving enhanced training on object-specific features.

Visual analysis for image feature extraction. Based on the visualization results in Fig. 7, the SVFE model demonstrates exceptional capability in extracting fine-grained compositional features through its integrated modules: For the “Cute Cat” example, the superpixel integration strategy precisely isolates the cat’s facial texture and whisker details (local attribute “cute”) while the Fourier spectral layer amplifies the global frequency patterns of fur softness across the body, with the long-range fusion module linking these localized textures to the holistic “cat” object. In “Wooden Bed”, superpixels capture high-frequency wood grain details at joint regions, while Fourier components enhance the structural continuity of the bed frame, enabling the fusion module to associate “wooden” material properties with the “bed” object. The “Broken Car” case highlights SVFE’s spatial sensitivity—superpixels activate strongly around windshield cracks (localizing broken), while the frequency domain dynamically amplifies fracture patterns against the car’s global contour, allowing the fusion module to correlate damage attributes with specific object regions. Similarly, for “Sliced Banana”, superpixels extract incision textures and flesh details, while Fourier modulation emphasizes the banana’s curved form, ensuring the model recognizes “sliced” as an attribute modifying the entire object. The “Bronze Bracelet” heatmap shows concentrated activation on metallic reflections (superpixel-level material capture) with frequency enhancement of specular highlights. Crucially, all heatmaps align with semantic labels, proving SVFE’s unified approach—superpixels isolate attribute-specific regions (e.g., fracture points, material spots), Fourier layers reinforce object-wide patterns (e.g., animal contours, structural shapes), and the fusion module establishes precise attribute-object interdependencies, overcoming the spatial insensitivity of prior methods.

6. Discussion

The proposed SVFE model presents a significant advancement in addressing the persistent challenges of compositional zero-shot learning. Existing CZSL approaches often rely on holistic feature representations that struggle to capture the nuanced interactions between attributes and objects (Nagarajan & Grauman, 2018). Our framework offers a compelling alternative through its integrated architecture that combines superpixel processing, frequency domain analysis, and long-range feature fusion. The consistent performance improvements across both closed-world and open-world CZSL scenarios strongly validate its effectiveness.

Theoretical Implications. Firstly, our work establishes a unified framework that connects low-level image processing with high-level semantic reasoning. This approach diverges from methods that treat attribute and object learning as separate optimization problems (Li et al., 2020; Ruis et al., 2021). Instead, SVFE employs an end-to-end architecture where each component informs and refines the others. The superpixel integration strategy draws inspiration from biological vision systems (Achanta et al., 2012), enabling the model to group pixels into perceptually coherent regions before semantic analysis. Then, the Fourier spectral layer introduces a novel perspective on feature representation. By operating in the frequency domain, the model gains efficient access to global image statistics that spatial convolutions often miss. This capability is particularly valuable for capturing long-range dependencies between attributes and objects. The learnable complex weight matrix serves as an adaptive filter, dynamically enhancing discriminative frequencies while suppressing irrelevant information. Lastly, the long-range fusion module completes this integrated approach through its dual-path architecture. By simultaneously modeling local attribute-sensitive patterns and global compositional relationships, it avoids the oversmoothing effects that plague graph-based methods (Zhang et al., 2022). The visualization results in Fig. 7 demonstrate this capability, showing precise activation alignments with semantic concepts.

Practical Implications. From an application perspective, SVFE’s performance improvements translate directly to enhanced reliability in real-world scenarios. Content-based retrieval systems could better handle queries involving novel attribute-object combinations, while autonomous systems might more accurately interpret instructions containing unseen compositions (Xu et al., 2021). The superpixel-based approach additionally confers inherent robustness to pixel-level variations like lighting changes and minor occlusions. Although the superpixel generation process introduces some computational overhead, the resulting gains in accuracy and generalization justify this cost for precision-sensitive applications. This trade-off represents a significant advantage over prompt-based methods, which remain constrained by frozen backbone architectures (Munir et al., 2024; Nayak et al., 2023). The model’s ability to adapt its feature extraction specifically for the CZSL task enables it to resolve spatial misalignments that prompt tuning alone cannot address.

Comparison with Existing Methods. Our work moves beyond several established paradigms in CZSL research. Unlike disentanglement methods that often suffer from oversmoothing issues (Zhang et al., 2022), SVFE’s iterative pixel-superpixel recombination

provides superior spatial sensitivity for modeling non-uniform attribute distributions. Where prompt-based frameworks rely on frozen feature extractors (Munir et al., 2024), our approach enables full end-to-end adaptation of the visual encoder. Compared to other feature enhancement strategies (Huo et al., 2024), SVFE's fusion of spatial and frequency domains offers a unique advantage. The Fourier layer captures global context more efficiently than stacks of spatial convolutions, while the long-range fusion module provides more sophisticated integration than simple concatenation operations. This comprehensive approach explains the consistent performance improvements demonstrated in our ablation studies.

Limitations and Future Directions. Despite its strengths, SVFE shows room for improvement in handling extremely rare or challenging unseen compositions. This limitation reflects a deliberate design choice to prioritize robust feature learning over specialized optimization for the unseen set. Future work might incorporate meta-learning techniques (Wei et al., 2020) to better address these edge cases without compromising overall performance. The computational aspects of superpixel generation also present opportunities for optimization. Developing more efficient neural superpixel formulations could reduce overhead while maintaining performance benefits. Additionally, exploring co-optimization strategies between visual and textual pathways would help create a more integrated cross-modal framework, moving beyond the current reliance on pre-trained text encoders.

7. Conclusion

In our paper, we introduce a pioneering approach to handle the compositional zero-shot learning task. Through the integration of superpixels and innovative architectural designs, the Superpixel-based Visual Feature Enhancement model significantly enhances the feature extraction and representation, delivering remarkable results on CZSL benchmarks. The model's ability to generalize to unseen compositions underscores its robustness and potential for real-world applications. The proposed approach overcomes previous limitations in fine-grained feature capture, offering a more nuanced comprehension of visual content. Rigorous experimentation on CZSL benchmark datasets confirms the model's superior performance, establishing a new benchmark in the CZSL domain. Future work will concentrate on optimizing computational efficiency and broadening the model's applicability, leveraging the robust foundation established by SVFE.

CRediT authorship contribution statement

Wenlong Du: Writing – original draft, Software, Methodology, Investigation. **Xianglin Bao:** Writing – review & editing, Formal analysis, Data curation. **Xiaofeng Xu:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Xingyu Lu:** Writing – review & editing, Validation, Resources. **Ruiheng Zhang:** Writing – review & editing, Visualization, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62406004, 62001229, 62201058), in part by the Key Project of Natural Science Category in Anhui Province's Higher Education Scientific Research, China (No. 2024AH050122), and in part by the Anhui Future Technology Research Institute Enterprise Cooperation Project, China (No. 2023qyhz14).

Data availability

Data will be made available on request.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(11), 2274–2282.

Achanta, R., & Süsstrunk, S. (2017). Superpixels and polygons using simple non-iterative clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4895–4904).

Atzmon, Y., Kreuk, F., Shalit, U., & Chechik, G. (2020). A causal view of compositional zero-shot recognition. In *Proceedings of the international conference on neural information processing systems: vol. 33*, (pp. 1462–1473).

Chytas, S. P., Kim, H. J., & Singh, V. (2025). Understanding multi-compositional learning in vision and language models via category theory. In A. s. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Proceedings of the European conference on computer vision* (pp. 324–341). Cham.

Dai, F., Huang, S., Zhang, M., Gong, B., & Wang, D. (2024). Focus-consistent multi-level aggregation for compositional zero-shot learning. *Clinical Orthopaedics and Related Research*, arXiv:2408.17083.

Dong, H., Fu, Y., Hwang, S. J., Sigal, L., & Xue, X. (2022). Learning the compositional domains for generalized zero-shot learning. *Computer Vision and Image Understanding*, *221*, Article 103454.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International Journal of Computer Vision*, *59*, 167–181.

Ferrari, V., & Zisserman, A. (2007). Learning visual attributes. In *Proceedings of the international conference on neural information processing systems: vol. 20*.

Hu, Z., Zhao, H., Peng, J., & Gu, X. (2022). Region interaction and attribute embedding for zero-shot learning. *Information Sciences*, *609*, 984–995.
Huo, F., Xu, W., Guo, S., Guo, J., Wang, H., Liu, Z., et al. (2024). Procc: Progressive cross-primitive compatibility for open-world compositional zero-shot learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(11), 12689–12697.
Isola, P., Lim, J. J., & Adelson, E. H. (2015). Discovering states and transformations in image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1383–1391).
Jampani, V., Sun, D., Liu, M.-Y., Yang, M.-H., & Kautz, J. (2018). Superpixel sampling networks. In *Proceedings of the European conference on computer vision* (pp. 352–368).
Jayasekara, H., Pham, K., Saini, N., & Shrivastava, A. (2025). Unified framework for open-world compositional zero-shot learning. In *2025 IEEE/CVF winter conference on applications of computer vision* (pp. 2706–2714).
Jiang, C., Ye, Q., Wang, S., Shen, Y., Zhang, Z., & Zhang, H. (2024). Mutual balancing in state-object components for compositional zero-shot learning. *Pattern Recognition*, *152*, Article 110451.
Karthik, S., Mancini, M., & Akata, Z. (2022). KG-SP: Knowledge guided simple primitives for open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9336–9345).
Kim, S., Lee, S., & Choi, Y. S. (2024). Focusing on valid search space in open-world compositional zero-shot learning by leveraging misleading answers. *IEEE Access*, *12*, 165822–165830.
Lampert, C. H., Nickisch, H., & Harmeling, S. (2009). Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 951–958).
Li, Z., & Chen, J. (2015). Superpixel segmentation using linear spectral clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1356–1363).
Li, Y.-L., Xu, Y., Mao, X., & Lu, C. (2020). Symmetry and group in attribute-object compositions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11316–11325).
Li, X., Yang, X., Wei, K., Deng, C., & Yang, M. (2022). Siamese contrastive embedding network for compositional zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9326–9335).
Liu, S., & Ozay, M. (2023). Task guided representation learning using compositional models for zero-shot domain adaptation. *Neural Networks*, *165*, 370–380.
Lu, X., Guo, S., Liu, Z., & Guo, J. (2023). Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 23560–23569).
Mancini, M., Naeem, M. F., Xian, Y., & Akata, Z. (2021). Open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5222–5230).
Mancini, M., Naeem, M. F., Xian, Y., & Akata, Z. (2022). Learning graph embeddings for open world compositional zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1.
Misra, I., Gupta, A., & Hebert, M. (2017). From red wine to red tomato: Composition with context. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1792–1801).
Munir, A., Qureshi, F. Z., Khan, M. H., & Ali, M. (2024). Attention based simple primitives for open-world compositional zero-shot learning. In *2024 international conference on digital image computing: techniques and applications* (pp. 714–721).
Naeem, M. F., Xian, Y., Tombari, F., & Akata, Z. (2021). Learning graph embeddings for compositional zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 953–962).
Nagarajan, T., & Grauman, K. (2018). Attributes as operators: factorizing unseen attribute-object compositions. In *Proceedings of the European conference on computer vision* (pp. 169–185).
Nayak, N. V., Yu, P., & Bach, S. H. (2023). Learning to compose soft prompts for compositional zero-shot learning. In *International conference on learning representations*.
Panda, A., & Mukherjee, D. P. (2024). Compositional zero-shot learning using multi-branch graph convolution and cross-layer knowledge sharing. *Pattern Recognition*, *145*, Article 109916.
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the international conference on neural information processing systems*: *vol. 32*.
Purushwalkam, S., Nickel, M., Gupta, A., & Ranzato, M. (2019). Task-driven modular networks for zero-shot compositional learning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3592–3601).
Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PMLR.
Rohrbach, M. (2016). Attributes as semantic units between natural language and visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. arXiv:1604.03249.
Ruis, F., Burghouts, G., & Bucur, D. (2021). Independent prototype propagation for zero-shot compositionality. In *Proceedings of the international conference on neural information processing systems*: *vol. 34*, (pp. 10641–10653).
Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888–905.
Shuang, F., Li, J., Huang, Q., Zhao, W., Xu, D., Han, C., et al. (2025). Visual primitives as words: Alignment and interaction for compositional zero-shot learning. *Pattern Recognition*, *157*, Article 110814.
Wang, H., Yang, M., Wei, K., & Deng, C. (2023). Hierarchical prompt learning for compositional zero-shot recognition. In *Proceedings of the thirty-second international joint conference on artificial intelligence* (pp. 1470–1478).
Wei, K., Deng, C., Yang, X., et al. (2020). Lifelong zero-shot learning. In *Proceedings of the thirty-second international joint conference on artificial intelligence* (pp. 551–557).
Xu, X., Bao, X., Lu, X., Zhang, R., Chen, X., & Lu, G. (2023). An end-to-end deep generative approach with meta-learning optimization for zero-shot object classification. *Information Processing & Management*, *60*(2), Article 103233.
Xu, G., Chai, J., & Kordjamshidi, P. (2024). GIPCOL: Graph-injected soft prompting for compositional zero-shot learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 5774–5783).
Xu, G., Kordjamshidi, P., & Chai, J. (2022). Prompting large pre-trained vision-language models for compositional concept learning. arXiv preprint arXiv: 2211.05077.
Xu, X., Tsang, I. W., Cao, X., Zhang, R., & Liu, C. (2019). Learning image-specific attributes by hyperbolic neighborhood graph propagation. In *Proceedings of the thirty-second international joint conference on artificial intelligence* (pp. 3989–3995).
Xu, X., Tsang, I. W., & Liu, C. (2021). Complementary attributes: A new clue to zero-shot learning. *IEEE Transactions on Cybernetics*, *51*(3), 1519–1530.
Yu, A., & Grauman, K. (2014). Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 192–199).
Yu, A., & Grauman, K. (2017). Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *Proceedings of the IEEE international conference on computer vision* (pp. 5570–5579).
Zhang, R., Li, L., Zhang, Q., Zhang, J., Xu, L., Zhang, B., et al. (2024). Differential feature awareness network within antagonistic learning for infrared-visible object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(8), 6735–6748.

Zhang, T., Liang, K., Du, R., Chen, W., & Ma, Z. (2025). Disentangling before composing: Learning invariant disentangled features for compositional zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *47*(2), 1132–1147.

Zhang, R., Xu, L., Yu, Z., Shi, Y., Mu, C., & Xu, M. (2022). Deep-irtarget: An automatic target detector in infrared imagery using dual-domain feature extraction and allocation. *IEEE Transactions on Multimedia*, *24*, 1735–1749.

Zhou, S., & Ma, J. (2024). Test-time prompt tuning for compositional zero-shot learning. In *2024 3rd International conference on electronics and information technology* (pp. 745–749).