IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 September 2024, accepted 1 October 2024, date of publication 8 October 2024, date of current version 12 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3476082

RESEARCH ARTICLE

ClusterE-ZSL: A Novel Cluster-Based Embedding for Enhanced Zero-Shot Learning in Contrastive Pre-Training Cross-Modal Retrieval

UMAIR TARIQ[1], ZONGHAI HU[1], KHAWAJA TAUSEEF TASNEEM[2], MD BELAL BIN HEYAT[3], MUHAMMAD SHAHID IQBAL[4], AND KAMRAN AZIZ[5]

[1]School of Electronic Engineering, Beijing University of Posts and Telecommunication, Beijing 100876, China
[2]Information Technology Department, College of Computing and Informatics, Saudi Electronic University, Riyadh 11673, Saudi Arabia
[3]CenBRAIN Neurotech Center of Excellence, School of Engineering, Westlake University, Hangzhou, Zhejiang 310024, China
[4]School of Computer Science and Technology, Anhui University, Hefei 230000, China
[5]Laboratory of Aerospace Information Security and Trusted Computing Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China

Corresponding author: Zonghai Hu (zhhu@bupt.edu.cn)

This work was supported in part by National Key R&D Program of China (2022YFB3605601 and 2017YFB0403602).

ABSTRACT Zero-shot learning (ZSL) in a multi-model environment presents significant challenges and opportunities for improving cross-modal retrieval and object detection in unseen data. This study introduced a novel embedding approach of vector space clustering to address image-to-text and text-to-image retrieval problems effectively. We proposed an iterative training strategy; unlike the CLIP model, which directly compares visual and textual modalities, our model concatenates by clustering trained image and text features in common vector space. We use cross-modal contrastive and multi-stage contrast loss to improve the unsupervised learning of our model. This integration makes it possible to achieve proper clustering on embedding, which enhances the image-text matching problem in zero-shot learning tasks. We rigorously evaluate our model performance on standard benchmark datasets, including Flickr30K, Flickr8K, and MSCOCO 5K, achieving notable improvements with accuracies of 91.3%, 88.8%, and 90.3%, respectively. The results demonstrate the better performance of our model over existing methods but also show its effectiveness in enhancing cross-modal retrieval in zero-shot learning.

INDEX TERMS Contrastive learning, embedded, cluster, self-supervised learning, embedded computing, cross-modal retrieval, multi-model machine learning.

I. INTRODUCTION

The rapid increase in multimedia data, the issue of cross-modal retrieval has become an important research topic. This cross-modal domain combines aspects of computer vision and Natural Language Processing (NLP) to solve the information search problem across modalities [1]. The process involves a query in one modality and a database containing various modalities, aiming to pinpoint the most relevant matches from the database. The inherent challenge lies in bridging the substantial semantic gaps between modalities like text and images, each encoding information differently within their respective embedding spaces [2]. Cross-modal retrieval essentially seeks to synchronize these disparate embedding spaces into a unified, comparable format. Typically, this is achieved by mapping visual and textual data into a shared embedding space, simplifying the retrieval process to the nearest neighbor search within the Euclidean space. The development of sophisticated techniques to facilitate efficient retrieval has practical value. The multimodal domain relied on feature extraction [3] and traditional machine-learning strategies [4], [5]. This model is designed mainly to improve

The associate editor coordinating the review of this manuscript and approving it for publication was Diego Oliva.

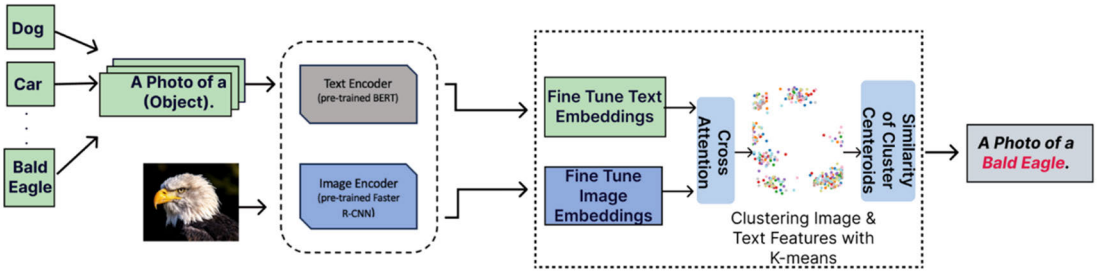

FIGURE 1. Architecture of our proposed ClusterE-ZSL model.

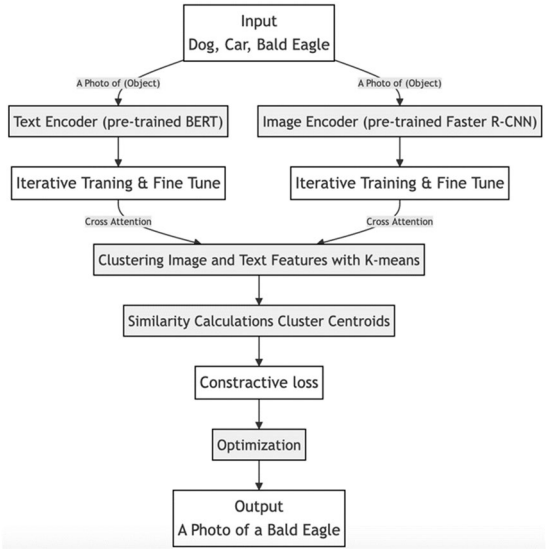

FIGURE 2. Simplified two-dimensional projection of image-text feature clustering used for illustrative purposes. Actual clustering is performed in a higher dimensional space, where each point represents an image or text feature vector derived from deep learning models.

the outcomes of zero-shot learning tasks by incorporating text and image encoders, fine-tuning, and clustering strategies. Figure 1. shows a multimodal deep learning architecture that integrates image and textual descriptions of images to improve robust zero-shot learning techniques. It uses two paths: A Faster R-CNN pre-trained model is used to encode the images, and a BERT pre-trained model encodes the corresponding texts for semantic embeddings for clustering.

These are then refined to attune well with the textual and visual data. A cross-attention layer focuses on the crucial information in both embeddings to enhance the interaction between these two modalities.

The resulting combined embeddings, K-means clustering, allow the model to identify semantically close embeddings, improving the categorization and retrieval of similar images or texts.

Recent studies on CLIP-based zero-shot have focused on refining the alignment between visual objects and textual descriptors [6]. The CLIP-decoder emphasizes the local alignment of these elements, while others extend to mapping complex phrases and significant image regions into a multimodal embedding space [7]. Figure 2. The block diagram represents a zero-shot learning architecture for cross-modal retrieval.

The process consists of refinement and optimization cycles carrying out cross-attention and then the final clustering of the image and text features by K-means. Contrastive loss is facilitated by calculating the similarities of two clusters between their centroids, which are, in turn, used to fine-tune the model for efficient cross-modal retrieval capabilities.

Our model differs from previous methods by leveraging an iterative training and contrastive learning approach in self-supervised learning with unlabeled data to foster understanding. This method enhances the proximity of similar pairs and distances from dissimilar features within the embedding space. Then, the combined embeddings are clustered with the help of the K-means algorithm ($k = 100$). These cluster embeddings assist with detecting patterns and associations involved in data, which also supports more efficient zero-shot learning. In terms of architectural design for cross-modal learning. The cross-modal embeddings process modalities independently, often employing a cross-attention mechanism to merge multimodal information at intermediate layers [8], [9]. Our approach proposes several novelties that differentiate our methodology from other approaches in the context of zero-shot learning and cross-modal retrieval. Not only does our model improve the zero-shot object detection based on vector space clustering, but it also presents a fresh approach to modeling the mapping from image and text data to a shared vector space. This integration is quite different from traditional methods, and the models that proposed it enable more refined and accurate matching of features, which has not been one of the focuses of most models. It is evident that FRCNN, as used for the image encoder, and DistilBERT, as used for the text encoder, provide fresh ways to approach the processing and analysis of multimodal data by making strategic points for movement within the vector space. This enhances the current models' effectiveness by finding the interconnection between text and picture information, expanding the abilities of the real-time search.

This work also shows comparable results to the existing state-of-the-art methods in baselines such as Flickr30K and MSCOCO while improving accuracy and retrieval measures.

In the area of zero-shot learning (ZSL) for cross-modal retrieval, a significant semantic gap between the modalities; various datasets need to scale better; it is highly dependent on accurate labeling, and there needs to be more model robustness and flexibility. The computational utilization of existing models could be more feasible in low-resource environments. Regarding these issues, our approach helps solve them by using vector space clustering embeddings and a cluster-based ZSL framework that deals with the problem by improving the integration of the modal features into the space with ZSL. Analyzing the results presented in this work, it can be noted that this method allows for a more effective overcoming of the semantic gap and improves the scalability and possibility of unseen data categories in the model. In addition, the process of iterative training and the depreciation of the model's dimension result in decreased computational requirements compared to our preliminary work without significantly compromising the measured performance, which makes the presented solution more applicable to real-world scenarios, which require constant adaption to new data, as well as overall computational efficiency. Compared with other approaches of the state of the art, the proposed model has better performance in terms of number accuracy, efficiency, and scalability. Our method has been tested on standard datasets, including Flickr30K, Flickr8K, and MSCOCO; nevertheless, significant enhancements were observed compared to CLIP-based models and other transformer-based architectures.

The proposed framework ensures that the problem can be solved in a more viable, expansive, and efficient approach than conventional methods that have consistently been implemented. The efficiency of the cluster-based structure, particularly its capacity to pre-train image and text representations for retrieval, this method surpasses the existing models, which require real-time computation of embeddings during retrieval queries. Several significant contributions of the 'ClusterE-ZSL' framework are stated below:

- Each stream dedicated to a specific modality (text or image) was trained using a contrastive learning approach in an asymmetric cross-modal contrastive learning framework.
- The ClusterE-ZSL improves the model by using embeddings at the cluster level, contributing to the generalization of unseen data. The clustering strategy assists in creating more comprehensive clusters that encompass diverse features and attributes, which is crucial when the model novices more unidentified case circumstances during the appraisal stage.
- The ClusterE-ZSL approach involves applying specialized methods to align image and text embeddings in the cluster properly. This alignment is essential, especially during cross-modal retrieval, where the mapping between the modalities directly affects performance.

The findings from the research work will be of great importance to the further development of zero-shot learning as well as cross-modal image-text retrieval. It clearly emphasized that the industry of digital media, online advertising, and automated content moderation, to name some, could immensely benefit from the direction and means that provide a better way of matching the images with the text without the need for labeling. The described approach can be used with other elaborate data or can be expanded to other kinds of tasks, for instance, multimodal sentiment analysis or automated tag assignment, which can envision new applications of AI in a variety of industries.

The organization of this study follows Section II: Related Work, which presents the analysis of related work with an explanation of how our work continues and differs from prior studies. Section III: Methods explains how the technical infrastructure of the new approach is constructed and the features different from Zero-Shot Learning and Vector Space Clustering. Section IV: Experiments demonstrate the outcomes of experiments, and the deployment of our methods in datasets proves our proposed approaches successful. Section V: To further qualify the effects of individual and synergistic elements in the Ablation Study, we break down the essential sections of our system. Finally, Section VI: Conclusion presents the overview and analysis of the results obtained and highlights the prospects for further research into the topic, stating the significance of our contribution to zero-shot learning and cross-modal retrieval.

II. RELATED WORK

The growth of innovative connected devices and social networks, there has been a growing multiplication of multimedia content on the Web [10]. This massive amount of data is in objects of different types, such as text, image, video, and audio, that are different in format but are semantically connected. Due to the increase in text and image data, the need for efficient search systems has continued to grow [11]. These systems include simple and essential to advanced types: single-model, cross-modal, and more. There are various relations between cross-modal retrieval and information retrieval (IR) since the main problem of cross-modal retrieval originated from the field of IR when the primary task was to search relevant documents or images for textual queries. Early approaches employed more frequentist 'black-box' approaches, which entailed hand-crafting most of the features and using basic matching techniques. For example, when a search of images was done without deep learning, the method [12], [13] of using histograms in image retrieval or using text-based meta tags in image search was used [14].

The semantic gap, where the text contains rich detail, and visuals are more abstract. Although it was still under development, methods like support vector machines (SVM) [15], [16] and decision trees [17], [18] were incorporated for better results in the process of retrieval [19]. These methods used hand-designed features for text and images that seek to create representations bridging gaps between these two domains. A significant development during this phase was the use of Canonical Correlation Analysis (CCA), which principally sought to find correlated subspaces for two different modalities to enhance the retrieval process. Cross-modal

TABLE 1. Comparative analysis of existing studies.

Reference	Year	Primary Contribution	Pros	Cons
[11]	2021	Data Efficient Language-supervised Zero-shot Recognition with Optimal Transport Distillation	Introduces optimal transport distillation for data efficiency in zero-shot recognition.	Primarily language-driven, may not be applicable to purely visual tasks.
[10]	2022	ECCV-Model	Comprehensive overview of cutting-edge research in computer vision.	As an edited volume, lacks the depth of a single-topic study.
[26]	2022	DiffCSE: Difference-based Contrastive Learning for Sentence Embeddings	Innovative contrastive learning technique improves sentence embedding robustness.	Specific to sentence embeddings; limited application to multimodal tasks.
[29]	2022	Modality-Aware Representation Learning for Zero-shot Sketch-based Image Retrieval	Focuses on modality-specific challenges in sketch-based retrieval.	Limited to sketch-based retrieval, not generalizable to other modalities.
[19]	2023	Combined scaling for zero-shot transfer learning	Explores scalability in zero-shot learning with promising results.	Focused on scaling, less on novel algorithmic development.
[23]	2023	Align before Fuse Vision and Language Representation Learning with Momentum Distillation	Advances integration of vision and language through novel pre-fusion alignment.	Momentum distillation complexity may hinder practical application.
[30]	2023	Clustering-based Image-Text Graph Matching for Domain Generalization	Enhances domain generalization in image-text graph matching.	Results still preliminary, published as a preprint.
[24]	2023	Deep Image Clustering Based on Label Similarity and Maximizing Mutual Information across Views	Innovative approach to image clustering using label similarity.	Focuses exclusively on images, not applicable to text.
[28]	2023	Large Language Models Enable Few-Shot Clustering	Utilizes cutting-edge large language models for clustering.	Computational intensity may limit accessibility for some users.
[20]	2023	ZeroCap: Zero-Shot Image-to-Text Generation for Visual-Semantic Arithmetic	Pioneers zero-shot image-to-text generation using visual-semantic arithmetic.	Still in conceptual stage with limited empirical validation.
[21]	2023	Latent Embeddings for Zero-shot Classification	Develops novel latent embedding techniques for zero-shot classification.	May require extensive tuning for optimal performance.
Our Study	2024	Clustered Embeddings for Improved Zero-Shot Learning in Contrastive Image-Text Matching	Introduces a novel cluster-based approach for zero-shot learning using contrastive learning for image-text matching. Achieves high accuracy across multiple benchmarks.	May require further optimization for scalability in larger datasets

retrieval entered the age of deep learning and was greatly enhanced. CNNs and RNNs enabled features to be extracted from images and text automatically, resulting in improved representation [20]. The DeViSE (Deep Visual-Semantic Embedding model) was highly influential in utilizing deep learning to map images into the text space for retrieval with substantial improvement in precision [21].

In parallel to these developments, the field of metric learning commenced its progression. Finding the distance function can precisely help measure the similarity between different modalities. Several methods, like triplet loss and contrastive loss, helped align the embeddings to ensure similar items are pulled closer while dissimilar ones are pushed far, irrespective of the modality [22]. Contrastive learning was a robust self-supervised learning paradigm useful in situations with limited labeled data. Comparing positive to antagonistic pairs allowed models to learn general and accurate representations of CV features even without explicit guidance [23]. This approach has been widely used in recent study works to improve cross-modal retrieval.

Exploring model architectures has led to the development of two distinct approaches: single-stream and two-stream architectures in the object recognition task. Single-stream models analyze inputs from different forms in parallel, usually integrating them at the initial stage of the system [24]. This integration can yield better intra-modality representations but at the potential loss of inter-modality flexibility and, sometimes, accuracy in modality-specific details. On the one hand, two-stream architectures process each modality independently at some point [20]. Two networks can focus detailed processing on their specialty, the auditory or the language modeling before the embeddings are combined for further integration and analysis. Some examples of cross-modal architectures include cross-attention mechanisms that enable the model to selectively direct attention to features belonging to the other group, thus improving interaction between the two [25].

Previous studies attempted to enhance these capabilities even further by including additional intricate mechanisms, such as the transformer models, which provide a sophisticated approach to managing sequences and attention across modes [26], [27]. In Table 1. comparative analysis of existing studies within cross-modal scenarios proved that this technique intensifies the correlation between the text and images, especially when confronted with intricate queries and heterogeneous data [28].

Thus, further development of unsupervised and semi-supervised approaches and consideration of techniques that can be learned continuously are considered worthwhile avenues for future research. These methods try to update bias based on new data at any time to improve their applicability to real-life problems in which data distributions may change.

III. METHODS

We implement symmetric multi-model contrastive learning to optimize our model's performance and reasoning capabilities. This learning approach intensifies the interaction within and between the modalities, enhancing the representation learning of both encoders. This method boosts the model's zero-shot learning, ensuring robustness and generalization across different modalities. We incorporate specialized losses to refine our training process: the cross-modal contrastive loss and the image-text matching loss. These losses are pivotal in guiding the model to resolve ambiguities and enhance alignment between the encoded modalities.

A. MODEL FRAMEWORK

Our proposed model architecture is structurally similar to CLIP. It consists of Image and Text encoders, as presented in Figure 1. We have opted for the pre-trained BERT model for the text encoder, renowned for its efficiency and effectiveness in distilling complex textual data into meaningful representations. For the images, we use the Faster R-CNN (FRCNN) as the vision encoder, which is highly regarded for its precision in detecting and encoding detailed aspects of images. These choices ensure that our model benefits from state-of-the-art technologies in both text and image processing [31], [32]. The model is fine-tuned to perform high-level associative and analytic tasks by focusing on cross-modal contrast and image-text matching. We use heard examples in image text matching. During training, we select the two most similar negative examples in each batch as 'hard examples.' This approach is based on the premise that if the model can successfully learn to differentiate these challenging cases, it will inherently improve its ability to distinguish fewer complex examples. By clustering the training data around more complicated examples, we create an environment where the model is continually challenged, accelerating the learning process and improving the efficacy of the model representations. Our model leverages a balanced combination of advanced encoding techniques and strategic learning methodologies to achieve superior zero-shot learning and cross-modal retrieval performance. Integrating model architectures, alongside sophisticated contrastive learning and loss functions, sets a strong foundation for the model to excel in understanding and connecting complex multi-model information.

B. DATA AUGMENTATION

We apply some augmentation to enhance the model's robustness for image and text data before feeding image-text pairs into our model as text data [33], [34]. The image input size is 224 × 224, enhancing the model's robustness to extract an object by 32 × 32 like [35]. The image I have multiple objects, and each of them has a bonding box $B_i = (x_i, y_i, w_i, h_i)$ where each box extracts an object from the input image. The extraction of an object is performed by equation (1).

$$I_{obj}^{(i)} = I[x_i : x_i + w_i, y_i : y_i + h_i] \tag{1}$$

where $I_{obj}^{(i)}$ represents the sub-image containing i-th object.

Through data augmentation, we form positive image and text single-model contrast pairs. We introduce a random crop for each image at a scale of 0.2 to 1, a random contrast ratio, a random Gaussian blur, a random greyscale, and a random horizontal flip. We use stop words for text data to filter meaningless words that appear frequently. We also adopt synonym replacement, random insertion, random swap, word repetition, and random deletion.

C. CLUSTERING

Since our dataset lacks predefined labels, we employ unsupervised clustering techniques to aid in identifying challenging examples for training. Similar to [36]. Once our model can effectively differentiate the most complex examples within a set, it becomes more adept at distinguishing the more straightforward cases. This principle underlies our adoption of clustering throughout the training process, which is applied to cross-modal. For the visual component of our data, we utilize the FRCNN model, which outputs features in a 512-dimensional space that provides a rich, detailed representation of each image. We then apply the K-means clustering algorithm to these features, organizing them into k distinct centers. This clustering groups similar images together and facilitates more focused and challenging contrastive learning tasks within those groups. We use BERT to extract text features for text data, generating a 300-dimensional vector for each text sample. Following the extraction, we apply k-means clustering to these text features, with the same number of clusters used for the images set at k centers. We switch from random to sequential data loading during training sessions to enhance the training challenge and effectiveness. Figure 3 presents the clustering of image-text pairs. This change ensures that all contrastive learning tasks occur within the same cluster [37]. By increasing the task difficulty in this manner, the model must develop more robust and nuanced data representations. The outcomes of this clustering are illustrated in Figure 4, presenting examples of how the data has been grouped according to similarity in both image and text modalities. In the case of the K-means clustering, the number of 100 clusters chosen was based on accuracy decomposition and computational cost. The proved elbow method supported the decision, which establishes the sum of squares equal to the distance from the point to its corresponding set center against the number of these sets.

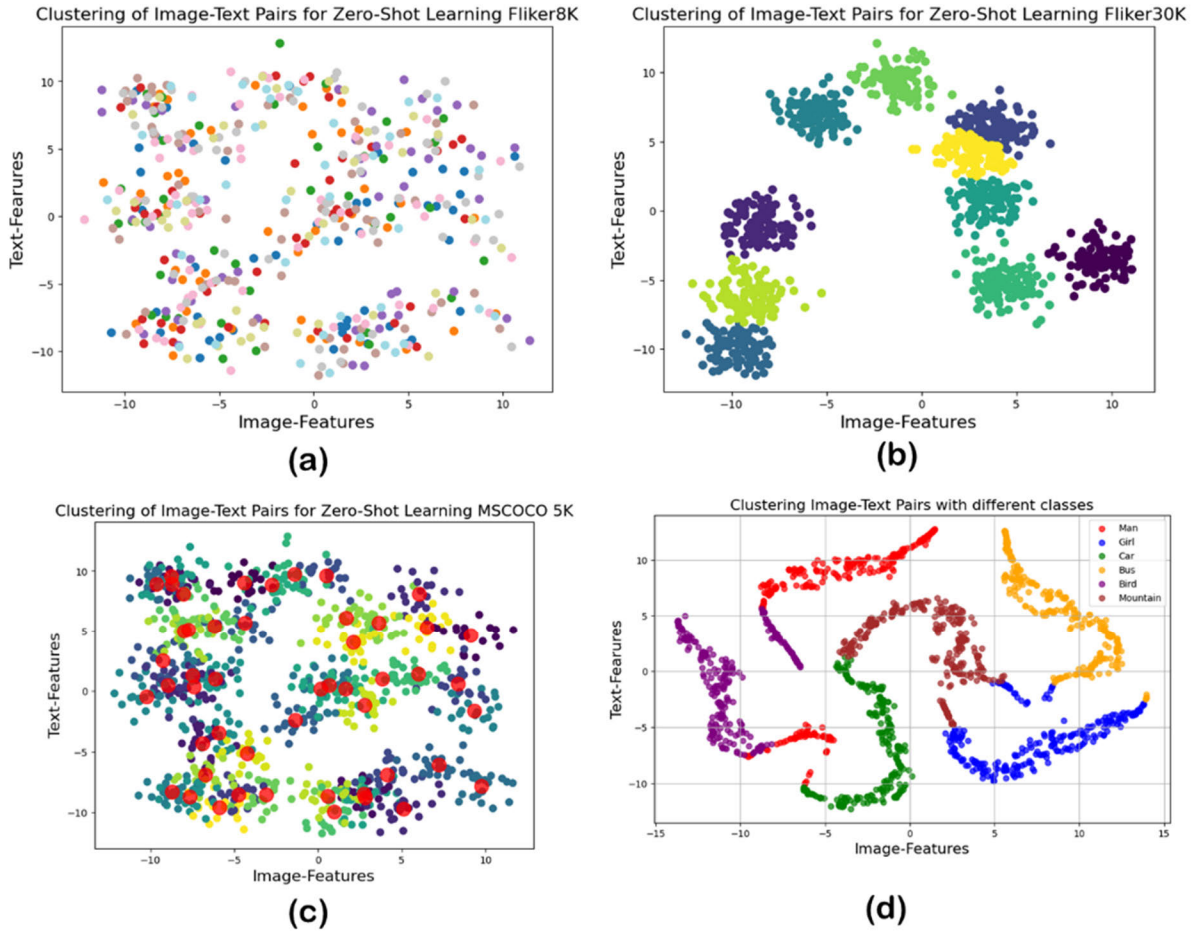

FIGURE 3. Clustering of Image-Text Pairs for Zero-Shot Learning. (a) Displays the clustering pattern on the Flickr30K dataset, showing how features group image-text pairs. (b) Illustrates similar clustering on the Flickr8K dataset. (c) Shows clustering results on the MSCOCO 5K dataset with varied clusters. (d) Shows clustering results on the MSCOCO 5K dataset with varied clusters. (d) Presents detailed class-specific clustering for categories like Man, Girl, Car, Bird, Bus, and Mountain, clearly segregating content types in a combined dataset.

D. IMAGE MODEL CONTRAST

Self-Attention is the core mechanism in equation (2), which involves mapping between a query and a set of key-value pairs, to a weighted sum of the values. Q, K, and V denote matrices packing together sets of queries, keys, and values [38]. The dot product of Q and K is scaled inversely by $\sqrt{d}$k, where dk is the dimension of the query and key vectors.

$$Attention\,(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

The encoder extracts richer information from different representation subspaces at other positions using the attention mechanism.

$$Z_0 = \left[x_{CLS}, x_i^1 E, x_i^2 E, \ldots x_i^N E\right] + E_{gpos,} E \in R^{(l^2.c)xD}, E_{pos} \in R^{(N+1)xD} \tag{3}$$

Apart from paired multi-model data, we can improve the performance of the single-model encoders by training unpaired l single-model data contrastively in equation (3). For the image-model contrast, we apply the image augmentation mentioned above to input images in a minibatch of N examples, resulting in $2N$ augmented images. We treat the other $2(N-1)$ augmented examples within a minibatch as negative examples. We use a PCA with one hidden layer to obtain the feature z_i.

$$Z_\iota = AM\,(LN\,(Z_{\iota-1})) + Z_{\iota-1}, \iota = 1 \ldots L \tag{4}$$

$$Z_\iota = PCA\,(LN\,(Z_{\iota-1})) + Z_{\iota-1}, \iota = 1 \ldots L \tag{5}$$

In equation (4) and (5) AM is the attention mechanism, PCA denotes the principal component analysis block, and LN denotes layer normalization. We employ a temperature-scaled cross-entropy loss similar to the infoNCE loss used in SimCLR, defined for the positive image pair (with label 1) and antagonistic pairs (with label 0) in equation (6).

$$Z = LN(z_L^0) \tag{6}$$

FIGURE 4. Examples of K-means Clustering on Image Datasets. (a) Shows initial similar clusters, demonstrating basic theme groupings. (b) shows another set of cluster examples, revealing broader image diversity.

The cumulative loss for image single-model contrast within a batch is expressed as equation (7). This framework ensures efficient learning from image and text data, optimizing the model's ability to process and align multi-model inputs.

$$L_{iuc} = \frac{1}{2n} \sum_{k=1}^{N} [1\,(2k-1, 2k) + (2k, 2k-1)] \quad (7)$$

E. TEXT MODEL CONTRAST

Text-model contrast is the same as image-model contrast. It is instrumental in amplifying the semantic gap between negative text pairings, thereby enhancing the capability of the text encoder to capture more refined semantic representations [39]. The text encoder operates similarly to the image encoder, with the primary distinction being the placement of the [EOT] (End of Token) end of the token sequence. We employ techniques such as synonym replacement, random insertion, swapping of words, repetition of words, and random deletion to construct positive text pairs. Conversely, negative pairs are constituted by other sentences within the same batch. The loss function for text-model contrast, denoted as , L_{iuc} the function used in image single-model contrast, aligns closely with the framework established in SimCSE. The hyperparameter τ is set at 0.05 for the text-model contrast to maintain consistency in the training process. To empirically validate the effectiveness of this approach, we select challenging examples from both image and text datasets. Subsequent visualization of patch-to-patch and token-to-token cosine similarity between the features extracted by our ClusterE-ZSL model. This enhancement further confirms the efficacy of our model contrastive learning during training and validation and its impact on zero-shot learning performance.

F. CROSS-MODAL CONTRAST

Cross-modal contrast enhances the separation between negative image-text pairs within our model. We use the embedding from the BERT to represent the textual content's overarching features. Similarly, for images, we employ the embedding from the Faster R-CNN [40]. Within a cluster, we calculate the similarity between the query and the cluster center as a measure of pairs of image-text similarity [41]. Cosine similarity between these pairs of cluster centers is calculated and denoted as C. The cross-modal contrast involves calculating a loss that effectively measures and optimizes the distance between non-matching cluster center pairs, thereby improving the model's ability to distinguish between them. This is implemented through the following loss function for cross-modal contrast:

$$L_{itc} = -\frac{1}{2n} \sum_{i=2}^{n} logp^{i2t}\,(I_i, T_i) - \frac{1}{2n}(\sum_{t=2}^{n} logp^{t2i}\,(I_t, T_t)) \quad (8)$$

The cross-entropy loss for image-text contrast is represented in equation (8) where $\sum_{i=2}^{n} logp^{i2t}\,(I_i, T_i) - \frac{1}{2n}(\sum_{t=2}^{n} logp^{t2i}\,(I_t, T_t)$. The probability distributions measure the likelihood of each image matching its corresponding text, as derived from the cosine similarity measures. The τ, temperature parameter in the softmax calculation within these probability distributions [42], is made learnable in our model, allowing it to adjust dynamically to the data characteristics, and n represents the batch size [26]. This approach improves the model's ability to correlate relevant image-text pairs. It solidifies the distinctions between mismatched pairs, enhancing overall accuracy and effectiveness in cross-modal

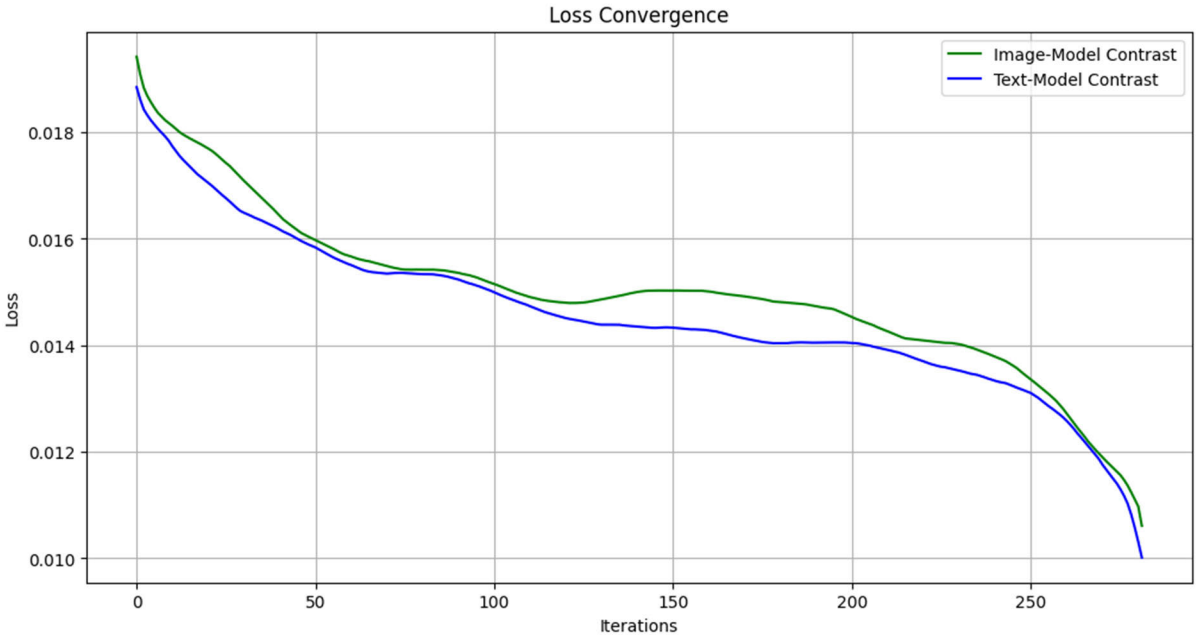

FIGURE 5. Text and image contrast loss during multiple iterations.

retrieval with zero-shot learning.

$$c^{i2t}(I_i, T_k) = \frac{\exp(c_{i,k/\tau})}{\sum_{j=1}^{n} \exp(c_{i,j/\tau})} \tag{9}$$

The softmax operation in equations (9) is a critical component of our cross-modal contrast methodology, which is used to compute the probability distributions for image-to-text and text-to-image retrieval tasks. Specifically, for a given batch of size n, these operations normalize the cosine similarities between images $\| \ . \ \|$ and text T_k, as well as between text T_j and image I_j, and apply a temperature scaling factor τ to fine-tune the sharpness of the distribution.

$$c^{t2i}\left(T_j, I_k\right) = \frac{\exp(c_{i,k/\tau})}{\sum_{i=1}^{n} \exp(c_{i,j/\tau})} \tag{10}$$

where k represents the index hard examples matching i or j, τ is learnable in cross-modal contrast, and n is the batch size in equation (10). We perform a rigorous analysis to evaluate the model's capacity for discrimination by selecting the most challenging or ''hard'' examples from both image and text domains within the batch. The effectiveness of this strategy is visually assessed through the generation of heat maps, which depict the cosine similarities on a token-to-token and patch-to-patch basis for our enhanced model. The hyperparameter $\tau\tau$ is particularly crucial as it is adaptively learned during training. This enables a dynamic adjustment to the scale of similarities, thereby improving the differentiation between positive and negative pairs. Figure 5. show image and text model contrastive loss.

G. ZERO-SHOT LEARNING

We present a cluster-based zero-shot learning method to enhance the efficiency of the model in matching input queries with trained classes. In zero-shot learning, only training data has labels. Our method works based on two assumptions: first, that data belonging to the target class will be substantially different from training data and, therefore, spatially far from it. K-clusters (centroids) and K-thresholds were extracted from labeled data in training. For every respective cluster, we establish the distance between the cluster's center and its most distant point. During the prediction phase, we establish the distance between the center of the clusters and its most distant point. We categorize instances according to these formed clusters to make the predictions in the third phase. If the distance of a new data point to the closest centroid is beyond a threshold set for that particular cluster, the latest data point is labeled as belonging to the target class for that cluster.

$$T_k = max_{x \epsilon S_k} \|X - C_k\| \tag{11}$$

To calculate the threshold for each cluster, we calculate the T_k as the distance between the centroids C_k where S_k is the set of all data points belonging to cluster k, and $\| \ . \ \|$ the distance in equation (11). We set the classification rule for new data point X, to determine the nearest cluster centroid $c_{nearest}$ and calculate the distance D from x to $c_{nearest}$ in equations (12) and (13).

$$c_{nearest} = argmin_{c_k} \| X - C_k \| \tag{12}$$

$$D = \| X - c_{nearest} \| \tag{13}$$

H. CROSS-MODAL RETRIEVAL

Cross-modal retrieval is a complex task that is essential for identifying whether the image and text pair belong to the same or different category when evaluated globally [44]. This process starts with assessing the matching between features of an image and those arising from text data. To optimize this capacity, we combine it with other tasks that utilize multi-task learning strategies, such as cross-modal contrast. This integration enhances cross-modal contrast learning since it involves the incorporation of features from both image and text. The accumulative losses are summed up to optimize the model's overall performance. This structured training approach ensures that the model learns to identify matching pairs accurately and enhances its ability to distinguish between non-matching pairs, thereby improving its zero-shot learning across all modalities involved. We particularly use a contrastive loss function which is crucial in learning discriminative features for the zero-shot learning model. The contrastive loss is intended to reduce the distance between the feature vectors of positive samples and maximize the distance for negative samples by using equation (14).

$$L_{constrastive}\,(i,j) = y.d(i,j)^2 + (1-y)\,.\max(0, m - d(i,j)^2) \tag{14}$$

where $d(i, j)$ are the Euclidian distance between an embedded feature of image and text pairs, y is the binary label (1 for positive and 0 for negative), and m is the margin.

I. TRAINING STRATEGY

In the training process, we use a two-stage training strategy. We optimize the image-model and text-model contrast task and then optimize the cross-modal self-supervised, which reduces overfitting. We design an iterative training procedure to implement a training strategy. In the first step, image and text models are trained separately by the contrast learning approach to enhance their representation learning. To improve cross-modal retrieval performance in zero-shot learning, we use contrastive embeddings as input for training. Both mapped image and text embeddings are subjected to contrastive learning at this stage. In an iterative training process, the model is trained with cluster data. During each epoch, our model gains updated weights and learns training patterns. As in most experiments, the subsequent experiments employ the default and primary text-splitting approach for the training data. Our iterative two-stage training strategy enhanced the efficiency of our model in zero-shot learning and image-text retrieval tasks.

IV. EXPERIMENTS

A. DATASET DETAILS

We use well-established public datasets, Fliker30k, Fliker8k, and MSCOCO5k, in the image-text retrieval domain. These datasets comprise a diverse collection of images: 31,783 in Fliker30k, 8,000 in Fliker8k, and 5,000 for MSCOCO5k. Every image in these datasets has associated five captions generated by human annotators with text descriptions. the partition of our data set is used. It includes 5,000 images for validation, 1,000 for 168 testing, and the rest for training. Also, previous studies used 20,000 training images, 4,000 validation images, and 5,00 test images, and we chose an equivalent split strategy for all datasets. The presented results average over five-fold of 1,000 test images or evaluations on complete sets. Dataset description details are mentioned in Table 2.

TABLE 2. Detail description of the datasets such as Fliker30K, Fliker8K, and MSCOCO.

Description	Specifications
Dimensions	Two (Images and Text)
Format for Text	.txt
Format for Images	JPEG
Input Size	224x224
Bonding Box Coordinates	[0.1,0.2,0.3,0.4]

B. IMPLEMENTATION DETAILS

Our experimental setup utilizes the Fliker30k caption dataset, which includes a comprehensive collection of 31,783 images. To enhance the robustness of our evaluation, we shuffle and then re-divide the original training and validation datasets. As a result, we allocate a set of 5,000 images strictly for testing purposes, with the remaining images designated for the training phase. For the initial training of our model, we employ a substantial batch size of 512 and conduct the training over 10 epochs. The specific two-stage training strategy implemented has been outlined previously. For the first seven epochs, we utilize a conservative learning rate of 1-3, which is set to 1e-6. For the rest of the training, the learning rate for image-text contrast is 3e-6, and the learning rate for image-text contrast is 2e-5. The model configuration is set to process sentences up to a maximum token length of 77, ensuring the inclusion of most caption lengths without truncation. We set 10 epochs for Fliker30K 10 for Fliker8K and 10 for MOSCOCO. This means the model stabilized after these epochs. This parameter is critical in balancing precision and generalization when computing the contrastive loss.

C. IMPLEMENTATION PLATFORM

We use the Adam optimizer and train our model on a T4 32 GB GPU, supported by a 2vCPU @ 2.2 GHz, 128 GB RAM, and a Linux 338 operating system. In our study, training one epoch on the Flickr30K dataset took an average of 0. 34 hours, where the model attains the best convergence for Flickr30K within four epochs. For the Flickr8K and MSCOCO datasets, convergence at the seventh epoch takes about 0. 8 hours to run each epoch.

D. HYPERPARAMETER SETTINGS

The following hyperparameters are used in this study to obtain state-of-the-art results. Batch size: 512, head learning

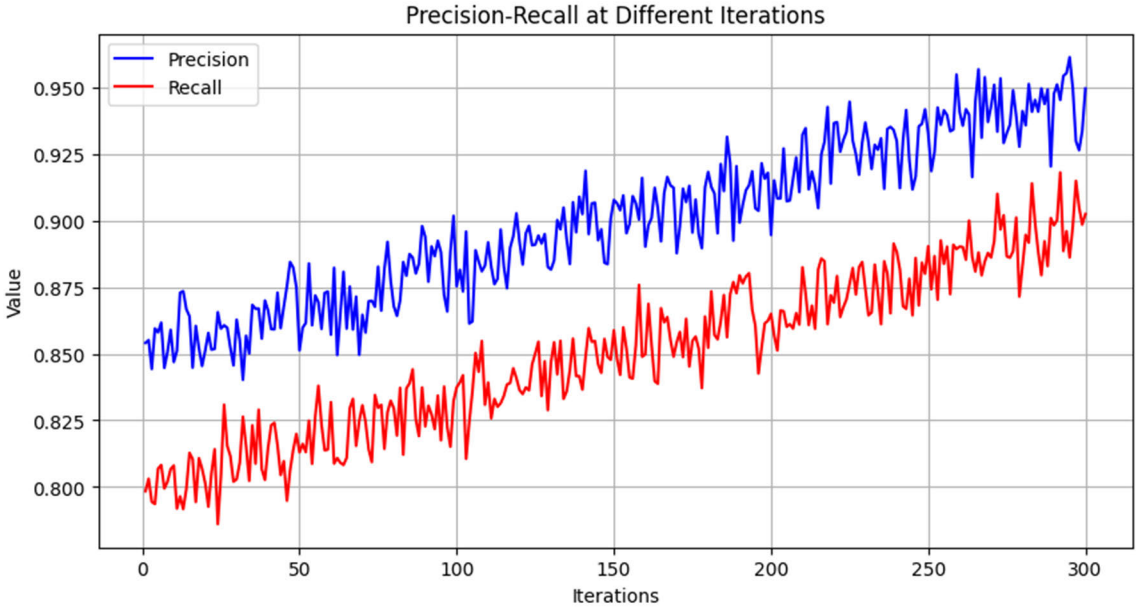

FIGURE 6. Precision-recall curves on different iterations.

TABLE 3. Training-testing loss and accuracy of our zero-shot learning model for image-text retrieval.

	Training		Testing	
Dataset	Loss	Accuracy%	Loss	Accuracy%
Fliker30K	0.307	91	2.35	91.3
Fliker8K	0.765	89	2.14	88.5
MSCOCO5K	0.371	90	2.26	90.3

rate: 1e-3, image encoder learning rate: 1e-4, text encoder learning rate: 1e-5, weight decay: 1e-3, patience: 1, factor: 0.8, Kernal size, 7×7 with stride 2, 3×3 with stride two and 1×1 with stride 1. Epochs: 10, but the model stabilized at the fourth epoch. Device: CUDA model name: Faster-RCNN, image embedding dimension: 2048, text encoder model: BERT, text embedding dimension: 768, text tokenizer: BERT, maximum token length: 77, pre-trained models: yes, trainable models: yes, temperature: 0.05, image size: 224, number of projection layers: 1, projection dimension: 256 and dropout rate: 0.1.

V. RESULTS AND DISCUSSION

In our study, we benchmarked our model against various baseline models, including Faster-RCNN-based architectures as pre-training models. To evaluate the effectiveness of our model in cross-modal retrieval tasks, we utilized the Recall at K (r@k) metric. This metric measures the ability of the system to retrieve the correct results within the top-K results for a given query, either retrieving texts for an image query or images for a text query.

Our findings show that our model significantly outperforms the traditional CNN-RNN-based models [45]. Even though existing pre-trained models such as CLIP and Unicoder-VL have the advantage of more extensive training datasets and enhanced object detection capabilities, our model demonstrates superior performance in both image-to-text and text-to-image retrieval tasks. Our model achieves this using the pre-trained weights for both image and text encoders. Table 3. Show the impact of different modes of inputs.

A. PERFORMANCE OF THE SYSTEM ON FLIKER30K

Several models were evaluated based on image-to-text and text-to-image retrieval results on the Flickr30K test set, and the promising result of the proposed ClusterE-ZSL is showcased, as indicated in Table 4. ClusterE-ZSL performs reasonably well regarding recall in image-to-text retrieval, with the required recall set at 89.8% at r@10. These results demonstrate its effectiveness in identifying the relative image descriptions, outperforming many models like VSE++ [46], DVSA [47], ImageBERT [48], and SGRAF+VSL [49]. In the text-to-image retrieval aspect, the effectiveness of ClusterE-ZSL is further proven with a score of 91.3% at r@10, which improves performances in this dataset. These results signify that our model identifies the right images for

TABLE 4. Image-to-text and text-to-image retrieval results on Flickr30K test set.

	Image-Text Retrieval			Text-Image Retrieval		
Model	r@1%	r@5%	r@10%	r@1%	r@5%	r@10%
CAMP [29]	51.5	77.1	85.3	68.1	89.7	95.2
DVSA [47]	25.2	37.7	50.5	22.2	48.2	61.4
VSE++[46]	39.6	70.1	79.5	52.9	80.5	87.2
MFTN [48]	54.3	79.6	87.5	70.7	90.2	94.0
CLIP-Zeroshot [6][7]	68.7	87.6	95.2	88.0	83.3	88.4
SGRAF+VSL [49]	79.5	81.3	71.9	60.2	84.3	89.4
VSRN [11]	62.0	73.4	81.0	76.0	90.0	90.4
SGRAF [59]	58.5	89.0	88.8	77.0	81.0	87.1
ImageBERT [48]	65.3	71.2	77.0	46.7	58.0	61.3
CAC [50]	80.1	76.1	87.7	87.0	76.0	84.9
Our ClusterE-ZSL	**62.1**	**78.2**	**89.8**	**89.7**	**90.2**	**91.3**

textual queries and can outperform other models like CLIPzeroshot and VSRN I text-image retrieval tasks. The ability to perform well on both image and text queries also reflects the cross-modal similarity of ClusterE-ZSL, showing that the proposed model can handle complex zero-shot learning scenarios.

B. PERFORMANCE OF THE SYSTEM ON FLIKER8K

In Table 5, the performance of the models for the Flickr8K test set for the image-to-text and text-to-image image retrieval tasks are presented, thereby highlighting the usage of the ClusterE-ZSL model, which performs well in comparison to the other models used. ClusterE-ZSL performs well in image-to-text retrieval, with 83.6% at r@10, and text-image retrieval, with 88.3% at r@10. These results show that our constructive learning strategy improves zero-shot learning compared to existing models like VSE++ and even the most complex models like ImageBERT. This demonstrates the model's ability to select pertinent images using retrieved text queries efficiently and effectively while retaining improved performance against models like VSRN and surpassing models such as MTFN [50], CAC [51], and SCG [52]. The high average score on both retrieval tasks confirms ClusterE-ZSL's feature learning. It helps establish it as a prominent participant for future solutions in the context of Zero-Shot Learning, especially when incorporating textual and visual data that need alignment and translation. Figure 6. presents two recall and precision confidence curves during our model's different learning iterations and performances. Recall generally decreases as confidence increases, and the Precision-Confidence curve demonstrates that precision typically improves with increased confidence. Each line represents a different iteration, illustrating variations in how the model's confidence impacts its recall and precision.

C. PERFORMANCE OF THE SYSTEM ON MSCOCO

In Table 6. Model performance on the MSCOCO5K dataset has been presented. It is observable that the ClusterE-ZSL model performs better than other existing models. Precisely, in the image-to-text retrieval tasks, ClusterE-ZSL attains recall rates of 85.8% at r@10 at r@10, text-to-image accuracy is 84.3 at r@10%, and it outperforms the existing models such as ImageBERT [53], VSE++ [54], and DVSA [55]. This demonstrates that ClusterE-ZSL recognizes the meaning of the data well when pairing images with textual descriptions. These results prove ClusterE-ZSL's reliability in efficiently image-text matching tasks. The F1-confidence curve shows the extent to which the model's F1 score- a measure of its accuracy obtained from a harmonic mean of precision and recall increased confidence level. We analyze the optimal confidence threshold to achieve maximum accuracy; an acute increase in the F1 score is evident for the data confidence levels compared to the non-contrastive learning method [56].

D. ABLATION STUDY

To better understand our model's various components, we performed several ablation studies based on the following aspects: single-model contrast, image-text matching, iterative training strategy, and clustering. These studies either eliminated or incorporated these components separately to analyze their effects on the model's performance. We compared the modified models in the following test set: MSCOCO 5K [57], Flickr30K, and Fliker8K [58].

1) EFFECTIVENESS OF SINGLE-MODEL CONTRAST

The ablation values show that effectively using the influence of different contrast modes on our model significantly increases its performance. The simple CLIP model [60] includes no other training mechanism except for cross-modal contrastive learning; our model also integrates a single-model contrast. On the Fliker30K test set, our model improves the retrieval metrics of interest by 1% than the CLIP [59] in all categories except for r@10, which is 0.2% of the raw CLIP performance. This apparent enhancement proves experimentally the benefits of incorporating the interaction strategy during the training courses. It reinforces its importance in

TABLE 5. Image-to-text and text-to-image retrieval results on Flickr30K test set.

	Image-Text Retrieval			Text-Image Retrieval		
Model	r@1%	r@5%	r@10%	r@1%	r@5%	r@10%
CAMP [29]	32.3	41.1	61.5	45.2	55.1	61.2
DVSA [47]	34.1	41.3	62.1	46.3	56.4	64.1
VSE++[46]	44.3	53.1	63.2	51.1	65.3	67.2
MTFN [48]	45.1	54.4	65.5	52.6	65.1	68.0
CLIP-Zeroshot [6][7]	53.5	67.3	77.1	65.1	78.1	87.3
SGRAF+VSL [49]	62.6	69.2	74.2	69.3	73.2	78.3
VSRN [11]	64.1	71.5	78.6	69.7	75.0	81.2
ImageBERT [48]	55.6	64.4	68.2	46.8	56.2	67.2
SCG [51]	45.2	53.2	64.1	51.2	64.1	71.8
CAC [50]	62.4	68.3	75.5	67.6	71.2	74.4
Our ClusterE-ZSL	**66.2**	**74.6**	**83.6**	**68.4**	**77.2**	**88.3**

FIGURE 7. Qualitative retrieval results for image and text query top-5 results are shown. Green denotes the top-matched image and text.

enhancing the model's ability to analyze the embeddings of training information and its compiled relationship within the overall scope of zero-shot learning and cross-modal tasks. Figure 7. presents the image-to-text and text-to-image retrieval results for image and text query top-5 results are shown.

In Figure 7. Green Text denotes the top-matched image and text. Due to the iterative training of clusters for contrastive zero-shot learning, our model can retrieve the best match.

2) EFFECTIVENESS OF EMBEDDING CLUSTER

The experimental evaluations on the MSCOCO and Flickr30K data sets provide valuable findings regarding clustering the training data in single-model and cross-modal training as compared to [61]. In cross-modal training tasks, employing only cluster data provides robust results compared to mixed data strategies. This approach approves that cluster learning is a more structured approach that can optimize performance in zero-shot learning.

3) EFFECTIVENESS OF ITERATIVE TRAINING STRATEGY

Another aspect of the model that was investigated is the effect of various training strategies on performance. Using multiple iterations of the training set is the same as expanding the dataset, but it is more cost-effective and improves the model's performance. The change in this test configuration alters tasks

TABLE 6. Image-to-text and text-to-image retrieval results on MSCOCO 5K test set.

	Image-Text Retrieval			Text-Image Retrieval		
Model	r@1%	r@5%	r@10%	r@1%	r@5%	r@10%
CAMP [29]	35.0	44.2	64.2	66.1	71.2	78.3
DVSA [47]	21.4	35.6	51.2	28.1	47.1	62.1
VSE++ [46]	46.1	62.4	76.2	54.1	79.3	77.2
MTFN [48]	47.2	61.4	79.5	67.1	73.1	76.0
SGRAF+VSL [49]	62.5	71.3	77.9	59.1	68.4	79.3
VSRN [11]	63.0	76.4	79.0	78.0	81.0	83.1
ImageBERT [48]	52.4	61.3	67.1	41.7	52.5	62.1
SCG [51]	59.1	68.7	76.1	63.4	76.2	81.3
CAC [50]	61.4	78.1	82.3	75.3	77.2	79.4
Our ClusterE-ZSL	**68.1**	**77.2**	**85.8**	**72.7**	**83.2**	**84.3**

involving zero-shot learning in image-to-text or text-to-image retrieval, as the first type generally becomes more challenging due to having more options [62].

E. LIMITATION OF THE STUDY

The limitation of this work is the performance of the proposed model highly depends on the quality and robustness of the annotated data used in the training process, which may not always be feasible with small datasets, as Fliker30k shows better performance than 8k and 5k datasets. Also, the computational complexity required for the proposed model, especially when working with large data volumes or many model interactions, poses a significant drawback in scenarios with strict computational resource availability.

VI. CONCLUSION

We propose a novel multi-model pre-training approach for zero-shot learning and cross-modal retrieval tasks. We introduced multi-stage learning of image and text-model training to improve the cross-modal performance. Our work further improves the original CLIP model with an iterative training strategy. By utilizing clustering embedding, the experiments and comparison of our model on the Flickr30K, Flickr8K, and MSCOCO datasets have shown better precision and outperform the existing models, which are dependent on a large-scale dataset. The evaluation of the ablation studies validates the efficiency of our cluster-based zero-shot contrastive learning and iterative training methods. In future work, we will use adaptive clustering techniques to enhance the model performance and robustness with multi-lingual capabilities.

ACKNOWLEDGMENT

The authors thank Beijing University of Posts and Telecommunication for providing resources to conduct and implement this research project.

AUTHOR'S CONTRIBUTION

Umair Tariq and Zonghai Hu: conceptualization, methodology, formal analysis, software, validation, and writing—original draft; Khawaja Tauseef Tasneem: data curation, methodology, investigation, visualization, and writing review and editing; Md Belal Bin Heyat: conceptualization, validation, project administration, resources, methodology, investigation, and writing—original draft; Muhammad Shahid Iqbal and Kamran Aziz: conceptualization, formal analysis, funding acquisition, resources, supervision, project administration, and writing review and editing. All authors read and agreed to the publication.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest to report regarding the present study.

DATA AVAILABILITY STATEMENT

The data presented in this research is available upon request from the corresponding author. The public datasets Fliker30K, Fliker8K, and MSCOCO can be accessed from: https://www.kaggle.com/datasets/utariq9/fliker30k-and-fliker8k.

REFERENCES

[1] S. K. Mishra, S. Joshi, and V. Gopalakrishnan, "Re-thinking text clustering for images with text," in *Proc. Int. Conf. Document Anal. Recognit.* Cham, Switzerland: Springer, Aug. 2023, pp. 280–294.

[2] A. Y. Muaad, S. Raza, M. B. B. Heyat, and A. Alabrah, "An intelligent COVID-19-related Arabic text detection framework based on transfer learning using context representation," *Int. J. Intell. Syst.*, vol. 2024, pp. 1–15, May 2024, doi: 10.1155/2024/8014111.

[3] T. Tayir and L. Li, "Unsupervised multimodal machine translation for low-resource distant language pairs," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 23, no. 4, pp. 1–22, Apr. 2024, doi: 10.1145/3652161.

[4] L. Ali, Z. He, W. Cao, H. T. Rauf, Y. Imrana, and M. B. B. Heyat, "MMDD-ensemble: A multimodal data–driven ensemble approach for Parkinson's disease detection," *Frontiers Neurosci.*, vol. 15, pp. 1–11, Nov. 2021, doi: 10.3389/fnins.2021.754058.

[5] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, Jan. 2014, doi: 10.1007/s11263-013-0658-4.

[6] Z. Guo, R. Zhang, L. Qiu, X. Ma, X. Miao, X. He, and B. Cui, "CALIP: Zero-shot enhancement of CLIP with parameter-free attention," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 1, Jun. 2023, pp. 746–754.

[7] M. Ali and S. Khan, "CLIP-decoder: ZeroShot multilabel classification using multimodal CLIP aligned representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, vol. 1, Oct. 2023, pp. 4677–4681, doi: 10.1109/ICCVW60793.2023.00505.

[8] A. Sedaghat and H. Ebadi, "Very high resolution image matching based on local features and k-means clustering," *Photogramm. Rec.*, vol. 30, no. 150, pp. 166–186, Jun. 2015, doi: 10.1111/phor.12101.

[9] L. Yin, L. Wang, S. Lu, R. Wang, H. Ren, A. AlSanad, S. A. AlQahtani, Z. Yin, X. Li, and W. Zheng, "AFBNet: A lightweight adaptive feature fusion module for super-resolution algorithms," *Comput. Model. Eng. Sci.*, vol. 140, no. 3, pp. 2315–2347, 2024, doi: 10.32604/cmes.2024.050853.

[10] H. J. Kim, N. Adluru, B. B. Bendlin, S. C. Johnson, B. C. Vemuri, and V. Singh, "Canonical correlation analysis on Riemannian manifolds and its applications," in *Proc. ECCV*, vol. 8690, 2014, pp. 251–267, doi: 10.1007/978-3-319-10605-2_17.

[11] B. Wu, R. Cheng, P. Zhang, T. Gao, P. Vajda, and J. E. Gonzalez, "Data efficient language-supervised zero-shot recognition with optimal transport distillation," in *Proc. 10th Int. Conf. Learn. Represent.*, Dec. 2021, pp. 13–32.

[12] J. V. B. Benifa, C. Chola, A. Y. Muaad, M. A. B. Hayat, M. B. B. Heyat, R. Mehrotra, F. Akhtar, H. S. Hussein, D. L. R. Vargas, Á. K. Castilla, I. D. L. T. Díez, and S. Khan, "FMDNet: An efficient system for face mask detection based on lightweight model during COVID-19 pandemic in public areas," *Sensors*, vol. 23, no. 13, p. 6090, Jul. 2023, doi: 10.3390/s23136090.

[13] M. S. Iqbal, R. Abbasi, M. B. Bin Heyat, F. Akhtar, A. S. Abdelgeliel, S. Albogami, E. Fayad, and M. A. Iqbal, "Recognition of mRNA N4 acetylcytidine (ac4C) by using non-deep vs. Deep learning," *Appl. Sci.*, vol. 12, no. 3, p. 1344, Jan. 2022, doi: 10.3390/app12031344.

[14] J. Li, Z. Ling, L. Niu, and L. Zhang, "Zero-shot sketch-based image retrieval with structure-aware asymmetric disentanglement," *Comput. Vis. Image Understand.*, vol. 218, Apr. 2022, Art. no. 103412, doi: 10.1016/j.cviu.2022.103412.

[15] F. Akhtar, M. B. B. Heyat, S. Parveen, P. Singh, M. F. U. Hassan, S. Parveen, M. A. B. Hayat, E. Sayeed, A. Ali, J. P. Li, and M. Sawan, "Early coronary heart disease deciphered via support vector machines: Insights from experiments," in *Proc. 20th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP)*, Dec. 2023, pp. 1–7, doi: 10.1109/iccwamtip60502.2023.10387051.

[16] M. B. Bin Heyat, D. Lai, K. Wu, F. Akhtar, A. Sultana, S. Tumrani, B. N. Teelhawod, R. Abbasi, M. Amjad Kamal, and A. Y. Muaad, "Role of oxidative stress and inflammation in insomnia sleep disorder and cardiovascular diseases: Herbal antioxidants and anti-inflammatory coupled with insomnia detection using machine learning," *Current Pharmaceutical Design*, vol. 28, no. 45, pp. 3618–3636, Dec. 2022, doi: 10.2174/1381612829666221201161636.

[17] M. B. B. Heyat, F. Akhtar, S. J. Abbas, M. Al-Sarem, A. Alqarafi, A. Stalin, R. Abbasi, A. Y. Muaad, D. Lai, and K. Wu, "Wearable flexible electronics based cardiac electrode for researcher mental stress detection system using machine learning models on single lead electrocardiogram signal," *Biosensors*, vol. 12, no. 6, p. 427, Jun. 2022, doi: 10.3390/bios12060427.

[18] Sumbul, A. Sultana, M. B. B. Heyat, K. Rahman, F. Akhtar, S. Parveen, M. B. Urbano, V. Lipari, I. De la Torre Díez, A. A. Khan, and A. Malik, "Efficacy and classification of sesamum indicum linn seeds with Rosa damascena mill oil in uncomplicated pelvic inflammatory disease using machine learning," *Frontiers Chem.*, vol. 12, pp.1361–1385, 2024.

[19] H. Pham, Z. Dai, G. Ghiasi, K. Kawaguchi, H. Liu, A. W. Yu, J. Yu, Y.-T. Chen, M.-T. Luong, Y. Wu, M. Tan, and Q. V. Le, "Combined scaling for zero-shot transfer learning," *Neurocomputing*, vol. 555, Oct. 2023, Art. no. 126658, doi: 10.1016/j.neucom.2023.126658.

[20] Y. Tewel, Y. Shalev, I. Schwartz, and L. Wolf, "ZeroCap: Zero-shot image-to-text generation for visual-semantic arithmetic," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17897–17907, doi: 10.1109/CVPR52688.2022.01739.

[21] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, "Latent embeddings for zero-shot classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 69–77, doi: 10.1109/CVPR.2016.15.

[22] Q. Zhang, Y. Zhu, M. Yang, G. Jin, Y. Zhu, and Q. Chen, "Cross-to-merge training with class balance strategy for learning with noisy labels," *Expert Syst. Appl.*, vol. 249, Sep. 2024, Art. no. 123846, doi: 10.1016/j.eswa.2024.123846.

[23] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 9694–9705.

[24] F. Peng and K. Li, "Deep image clustering based on label similarity and maximizing mutual information across views," *Appl. Sci.*, vol. 13, no. 1, p. 674, Jan. 2023, doi: 10.3390/app13010674.

[25] Q. Zhang, G. Jin, Y. Zhu, H. Wei, and Q. Chen, "BPT-PLR: A balanced partitioning and training framework with pseudo-label relaxed contrastive loss for noisy label learning," *Entropy*, vol. 26, no. 7, p. 589, Jul. 2024, doi: 10.3390/e26070589.

[26] Y. S. Chuang, R. Dangovski, H. Luo, Y. Zhang, S. Chang, M. Soljacic, S.-W. Li, S. Yih, Y. Kim, and J. Glass, "DiffCSE: Difference-based contrastive learning for sentence embeddings,"2022, *arXiv:2204.10298*.

[27] L. Yin, L. Wang, S. Lu, R. Wang, Y. Yang, B. Yang, S. Liu, A. AlSanad, S. A. AlQahtani, Z. Yin, X. Li, X. Chen, and W. Zheng, "Convolution-transformer for image feature extraction," *Comput. Model. Eng. Sci.*, vol. 141, no. 1, pp. 87–106, 2024, doi: 10.32604/cmes.2024.051083.

[28] V. Viswanathan, K. Gashteovski, C. Lawrence, T. Wu, and G. Neubig, "Large language models enable few-shot clustering," 2023, *arXiv:2307.00524*.

[29] E. Lyou, D. Lee, J. Kim, and J. Lee, "Modality-aware representation learning for zero-shot sketch-based image retrieval," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 5634–5643.

[30] N. Park, D. Chae, J. Shim, S. Kim, E.-S. Kim, and J. Kim, "Clustering-based image-text graph matching for domain generalization," 2023, *arXiv:2310.02692*.

[31] P. Li, P. Chen, Y. Xie, and D. Zhang, "Bi-modal learning with channel-wise attention for multi-label image classification," *IEEE Access*, vol. 8, pp. 9965–9977, 2020, doi: 10.1109/ACCESS.2020.2964599.

[32] M. S. Iqbal, M. B. B. Heyat, S. Parveen, M. A. B. Hayat, M. Roshanzamir, R. Alizadehsani, F. Akhtar, E. Sayeed, S. Hussain, H. S. Hussein, and M. Sawan, "Progress and trends in neurological disorders research based on deep learning," *Computerized Med. Imag. Graph.*, vol. 116, Sep. 2024, Art. no. 102400, doi: 10.1016/j.compmedimag.2024.102400.

[33] J. Zhou, L. Dong, Z. Gan, L. Wang, and F. Wei, "Non-contrastive learning meets language-image pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11028–11038, doi: 10.1109/cvpr52729.2023.01061.

[34] H. Ullah, M. B. B. Heyat, F. Akhtar, A. Y. Muaad, C. C. Ukwuoma, M. Bilal, M. H. Miraz, M. A. S. Bhuiyan, K. Wu, R. Damaševičius, T. Pan, M. Gao, Y. Lin, and D. Lai, "An automatic premature ventricular contraction recognition system based on imbalanced dataset and pre-trained residual network using transfer learning on ECG signal," *Diagnostics*, vol. 13, no. 1, p. 87, Dec. 2022, doi: 10.3390/diagnostics13010087.

[35] A. Vulli, P. N. Srinivasu, M. S. K. Sashank, J. Shafi, J. Choi, and M. F. Ijaz, "Fine-tuned DenseNet-169 for breast cancer metastasis prediction using FastAI and 1-Cycle policy," *Sensors*, vol. 22, no. 8, p. 2988, Apr. 2022, doi: 10.3390/s22082988.

[36] D. Kong, K. Kong, and S.-J. Kang, "Image clustering using generated text centroids," *Signal Process., Image Commun.*, vol. 125, 2024, Art. no. 117128.

[37] A. Petukhova, J. P. Matos-Carvalho, and N. Fachada, "Text clustering with LLM embeddings," 2024, *arXiv:2403.15112*.

[38] J. U. Allingham, J. Ren, M. W. Dusenberry, X. Gu, Y. Cui, D. Tran, J. Z. Liu, and B. Lakshminarayanan, "A simple zero-shot prompt weighting technique to improve prompt ensembling in text-image models," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2023, pp. 547–568.

[39] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, doi: 10.18653/v1/2021.emnlp-main.552.

[40] D. Lai, X. Zhang, Y. Zhang, and M. B. Bin Heyat, "Convolutional neural network based detection of atrial fibrillation combing R–R intervals and F-wave frequency spectrum," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 4897–4900, doi: 10.1109/EMBC.2019.8856342.
[41] M. Mayank, S. Sharma, and R. Sharma, "DEAP-FAKED: Knowledge graph based approach for fake news detection," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Nov. 2022, pp. 47–51, doi: 10.1109/ASONAM55673.2022.10068653.
[42] C. C. Ukwuoma, G. C. Urama, Z. Qin, M. B. B. Heyat, H. M. Khan, F. Akhtar, M. S. Masadeh, C. S. Ibegbulam, F. L. Delali, and O. AlShorman, "Boosting breast cancer classification from microscopic images using attention mechanism," in *Proc. Int. Conf. Decis. Aid Sci. Appl. (DASA)*, Mar. 2022, pp. 258–264, doi: 10.1109/DASA54658.2022.9765013.
[43] A. M. Butnaru and R. T. Ionescu, "From image to text classification: A novel approach based on clustering word embeddings," *Proc. Comput. Sci.*, vol. 112, pp. 1783–1792, Jan. 2017, doi: 10.1016/j.procs.2017.08.211.
[44] A. Tabassum, R. Kannan, J. Yin, S.-H. Lim, G. Cong, and S. M. Hasan, 2024, "Knowledge graph embedding using large language models for COVID-19," doi: 10.13139/ORNLNCCS/2229136.
[45] D. Shu, T. Chen, M. Jin, C. Zhang, M. Du, and Y. Zhang, "Knowledge graph large language model (KG-LLM) for link prediction," 2024, *arXiv:2403.07311*.
[46] Q. Zhu, L. Hu, and R. Wang, "Image clustering algorithm based on predefined evenly-distributed class centroids and composite cosine distance," *Entropy*, vol. 24, no. 11, p. 1533, Oct. 2022, doi: 10.3390/e24111533.
[47] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, "VideoCLIP: Contrastive pre-training for zero-shot video-text understanding," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Stroudsburg, PA, USA, 2021, pp. 6787–6800, doi: 10.18653/v1/2021.emnlp-main.544.
[48] R. Patil, S. Boit, V. Gudivada, and J. Nandigam, "A survey of text representation and embedding techniques in NLP," *IEEE Access*, vol. 11, pp. 36120–36146, 2023, doi: 10.1109/ACCESS.2023.3266377.
[49] R. Liu, Q. Zhong, M. Cui, H. Mai, Q. Zhang, S. Xu, X. Liu, and Y. Du, "The short text matching model enhanced with knowledge via contrastive learning," 2023, *arXiv:2304.03898*.
[50] S. Cheng, N. Zhang, B. Tian, X. Chen, Q. Liu, and H. Chen, "Editing language model-based knowledge graph embeddings," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 16, Mar. 2024, pp. 17835–17843.
[51] P. Huang, J. Han, D. Cheng, and D. Zhang, "Robust region feature synthesizer for zero-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7612–7621, doi: 10.1109/CVPR52688.2022.00747.
[52] C. Geng, S.-J. Huang, and S. Chen, "Recent advances in open set recognition: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3614–3631, Oct. 2021, doi: 10.1109/TPAMI.2020.2981604.
[53] Q. Chen, W. Wang, K. Huang, and F. Coenen, "Zero-shot text classification via knowledge graph embedding for social media data," *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9205–9213, Jun. 2022, doi: 10.1109/JIOT.2021.3093065.
[54] A. Sheth, "Transforming big data into smart data: Deriving value via harnessing volume, variety & velocity using semantics and semantic web," in *Proc. IEEE 30th Int. Conf. Data Eng. (ICDE)*. IEEE Computer Society, Mar. 2014, p. 2.
[55] A. Sadhu, K. Chen, and R. Nevatia, "Zero-shot grounding of objects from natural language queries," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4693–4702, doi: 10.1109/ICCV.2019.00479.
[56] J. Z. Li, D. Zhe, G. Lijuan, W. Furu, and W. Microsoft. *Non-Condsftrastive Learning Meets Language-Image Pre-Training*. Accessed: Mar. 12, 2024. [Online]. Available: https://github.com/shallowtoil/xclip
[57] R. Zhang, Y.-S. Wang, and Y. Yang, "Generation-driven contrastive self-training for zero-shot text classification with instruction-following LLM," 2023, *arXiv:2304.11872*.
[58] Z. Akata, M. Malinowski, M. Fritz, and B. Schiele, "Multi-cue zero-shot learning with strong supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 59–68, doi: 10.1109/CVPR.2016.14.
[59] J. Jeong, K. Tian, A. Li, S. Hartung, S. Adithan, F. Behzadi, J. Calle, D. Osayande, M. Pohlen, and P. Rajpurkar, "Multimodal image-text matching improves retrieval-based chest X-ray report generation," in *Proc. Med. Imag. Deep Learn.*, Jan. 2023, pp. 978–990.
[60] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer, "LiT: Zero-shot transfer with locked-image text tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18102–18112.
[61] J. Yuan, S. Zhu, S. Huang, H. Zhang, Y. Xiao, Z. Li, and M. Wang, "Discriminative style learning for cross-domain image captioning," *IEEE Trans. Image Process.*, vol. 31, pp. 1723–1736, 2022, doi: 10.1109/TIP.2022.3145158.
[62] Y. Wei, Q. Huang, Y. Zhang, and J. Kwok, "KICGPT: Large language model with knowledge in context for knowledge graph completion," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2023, pp. 8667–8683, doi: 10.18653/v1/2023.findings-emnlp.580.

UMAIR TARIQ received the B.S. degree in computer science from the University of Azad Jammu and Kashmir, Pakistan, and the M.S. degree in computer science from IQRA University Islamabad Campus, Pakistan. He is currently a Ph.D. Research Fellow with Beijing University of Posts and Telecommunication, Beijing, China. His research interests include embedded computing, natural language processing, embedded artificial intelligence, image-text matching, image-text retrieval, multi-model, and machine learning.

ZONGHAI HU received the B.S. degree from Peking University, in 1996, and the Ph.D. degree from Columbia University, in 2001. He is currently a Professor of electronic engineering with Beijing University of Posts and Telecommunications. His current research interests include information materials and devices and intelligent information processing.

KHAWAJA TAUSEEF TASNEEM received the bachelor's degree in computer science from the University of Peshawar, Pakistan, in 2003, the M.S. degree in electronic engineering from Muhammad Ali Jinnah University, Pakistan, in 2006, and the Ph.D. degree in electrical and electronic engineering from the University of Canterbury, New Zealand, in 2013. From 2013 to 2016, he was an Assistant Professor with Iqra University Islamabad Campus, Pakistan. In 2016, he joined Saudi Electronic University, Saudi Arabia, where he is currently an Assistant Professor with the College of Computing and Informatics. His research interests include artificial intelligence, wireless communications, statistical signal processing, and data mining

MD BELAL BIN HEYAT received the B.Tech. degree in electronics and instrumentation (EI) and the M.Tech. degree in electronics circuits and systems (ECS) from Integral University, Lucknow, Uttar Pradesh, India, in 2014 and 2016, respectively, and the Ph.D. degree in electronic science and technology from the University of Electronic Science and Technology of China, Chengdu, Sichuan, China.

He was a Research Associate and the member of the UESTC Country League and the Country Representative of India, during the Ph.D. degree. He was a Postdoctoral Fellow with the IoT Research Center, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, Guangdong, China, from 2021 to 2023. Currently, he is with the CenBRAIN Neurotech Center of Excellence, Westlake University, as a Postdoctoral Fellow. He is also a Visiting Postdoctoral Researcher with CVEST, IIIT Hyderabad, India, and a Faculty Member with the Department of Science and Engineering, Novel Global Community Educational Foundation, NSW, Australia. He has published more than 70 articles in reputed international journals and conferences. His research interests include detection, sleep disorders, neurological disorders, psycho-neurological disorders, cardiovascular diseases, signal processing, and medical machine learning. He received nine awards. He has been serving as a Guest Editor for three journals, including *Life*, *Journal of Integrative Neuroscience*, and *Applied Sciences*, Additionally, he is a Reviewer of over five publishers, including Oxford press, IEEE, Elsevier, Hindawi, Wiley, Springer, and MDPI.

MUHAMMAD SHAHID IQBAL received the Ph.D. degree from the School of Computer Science and Application Technology, Anhui University, Hefei, China, in 2019. He is currently a Research Associate with the School of Computer Science and Technology, Anhui University, and an Assistant Professor with the Department of Computer Science and Information Technology, Women University of Azad Jammu and Kashmir, Azad Jammu and Kashmir, Pakistan. His research interests include AI in biology, machine learning, deep learning, and computational and digital pathology.

KAMRAN AZIZ received the M.S. degree in computer science and technology from Nanjing University of Information Technology, Nanjing, China. He is currently pursuing the Ph.D. degree in cyberspace security with the School of Cyber Science and Engineering, Wuhan University, China. He is an Expert in natural language processing (NLP) and focuses on cutting-edge applications, such as fake news detection, named entity recognition, sentiment analysis, and data summarization and augmentation. His research interest includes aims to enhance the reliability and efficiency of information processing in digital media.