Hierarchical Textual Knowledge for Enhanced Image Clustering

Supplementary Material

6. Content List

We provide additional details and results to complement the main paper. It is organized as follows:

- section 7 lists all notations and their meanings.
- section 8 provides the details of 20 datasets used.
- section 9 describes implementation details of all the compared methods and the proposed method KEC.
- section 10 presents complete results across all datasets in comparison with existing methods. We also explore further improvement with human interaction, named KEC$^+$.
- section 11 presents a more comprehensive set of ablation studies, including the analyses of different modules, parameter selection, clustering methods, pretrained model choices, and various LLMs.
- section 12 describes the constructed knowledge space.
- section 13 discusses the limitations of our work and outlines future work possibilities.

7. Symbol Definitions

We list the symbols used throughout this paper along with their meanings in Table 3. We hope this could assist readers in better understanding the items presented in our work.

8. More Details of the Datasets

We evaluate our knowledge-enhanced clustering method on 20 vision datasets. Details of each dataset are provided in Table 4. These datasets cover a wide range of vision tasks, including:

- General object classification datasets: CIFAR-10 [19], CIFAR-100 [19], STL-10 [8], ImageNet [9];
- Fine-grained object classification datasets: Food101 [1], Flowers [31], Stanford Cars [18], FGVC Aircraft [28], Oxford Pets [32];
- Handwritten digits classification dataset: MNIST [21];
- Texture classification dataset: DTD [7];
- Scene classification dataset: SUN397 [43];
- Satellite image classification datasets: EuroSAT [15], Resisc45 [5];
- German Traffic Sign Recognition Benchmark: GTSRB [37];
- The metastatic tissue classification dataset: PatchCamelyon (PCAM) [41];
- Action Recognition dataset; UCF101 [36];
- The CLEVR counting dataset [16];
- The Hateful Memes dataset [17];
- The Rendered SST2 dataset [34];

We process these datasets following the open-source code [11, 22] [1] [2]. For CLEVR, we take 2000 random samples as the training set and 500 as the testing set. For the video dataset UCF101, we take the middle frame of each video clip as the input of the pre-trained CLIP vision encoder.

9. More Implementation Details

9.1. Implementation of the Compared Methods

K-Means and other traditional clustering methods. We apply k-means clustering [27] on top of pre-trained features as a simple baseline that only uses knowledge from visual space. Following the previous work [22], we implement traditional clustering methods, such as k-means, using the FAISS [3] library for GPU acceleration (`i.e.`, faiss.Kmeans). Additionally, the relevant parameter settings are consistent with those used in TAC. We run the clustering for each dataset 20 times with 300 iterations and keep the best centroids for each iteration (`nredo` $= 20, $`niter` $= 300$). The centroids are L2 normalized after each iteration (`spherical` $=$ `True`).

Specifically, for K-means and other traditional clustering methods involved in the ablation studies (*i.e.* Spectral Clustering, Agglomerative Clustering, Bisecting K-Means), we utilized the clustering implementations found in the `cluster` module of the `sklearn` library.

Zero-shot CLIP. To validate the applicability across different scenarios, we use the same prompt list for different datasets, rather than employing a set of prompts manually designed for the characteristics of each dataset (`dataset to template` in TURTLE). We utilized the same prompt list as in the TAC open-source code, referred to as 'the simple ImageNet prompt' [4], which consists of seven prompts. The prompts are shown in Table 5.

Semantic-Enhanced Image Clustering. SIC [2] first explores utilizing the knowledge from both visual space and textual space. It attempts to generate pseudo-labels for each image according to the relationships between images and the semantics of the nouns in WordNet. We utilize the open-source code from SIC [5], employing its default parameters. When extending the datasets used, we build the *dataloader*

[1] https://github.com/XLearning-SCU/2024-ICML-TAC
[2] https://github.com/mlbio-epfl/turtle
[3] https://github.com/facebookresearch/faiss
[4] https://github.com/openai/CLIP/blob/main/notebooks/Prompt_Engineering_for_ImageNet.ipynb
[5] https://github.com/Bruce-XJChen/SIC

Table 3. The summary of the used symbols.

Symbol	Meaning
Initializing Image-Text Mapping	
$\mathcal{I}, I_i$	the collection of input images and a single input image
$\mathcal{W}, w_i$	the collection of nouns and a single noun
N_v, N_t	the number of input images and nouns
$X, \mathbf{x}_i$	the collection of visual features and the feature of image I_i
$T, \mathbf{t}_i$	the collection of noun features and the feature of noun w_i
Representative Concept Construction	
$\boldsymbol{\mu}_p$	the centroid of cluster p
$\mathcal{S}_p$	Index of the highest scoring noun in cluster p
W_p, T_p	the collection of nouns and their features of cluster p
$R_{i,j}$	the similarity between cluster i and cluster j
G, Q	the adjacency matrix and the collection of connected components of clusters
$\mathcal{C}, c_q$	the collection of concept and a single concept
$\mathcal{D}, d_q$	the collection of description of concept and a single description
ϕ_q, ψ_q	the feature of concept and its description
Discriminative Attribute Construction	
λ_1, λ_2	the nubmer of uni-concept attribute and bi-concept attribute
$\mathcal{U}_q, u_q^i$	the collection of uni-concept attribute for concept c_q and the i-th attribute
$\pi_{q,l}$	the normalized similarity between concept c_q and c_l
$\bar{l}_{q,1}, \bar{l}_{q,2}, \cdots$	the index of l for the sorted $\pi_{q,l}$
$\mathcal{P}_q$	the concept pair of concept c_q
$\mathcal{B}_{q,l}$	the bi-concept attributes for concept pair (c_q, c_l)
$\mathcal{A}_q, \xi_{q,i}$	the collection of attributes of concept c_q and the i-th attribute
Knowledge-enhanced Feature	
ζ_q	the concept feature for conecpt c_q
$\bar{\xi}_q^i$	the average of instantiated attribute features of concept c_q
$\omega_{i,q}$	the attention weight between the i-th input image and concept c_q
$\mathbf{c}_i, \mathbf{a}_i$	the concept and attribute feature of the i-th input image I_i
κ_i	the knowledge-enhanced feature of the i-th input image I_i

Table 4. Details of 20 datasets.

Dataset	Number fo Classes	Train Size	Test Size	Dataset	Number fo Classes	Train Size	Test Size
CIFAR-10	10	50,000	10,000	DTD	47	3,760	1,880
CIFAR-100	100	50,000	10,000	SUN397	397	19,850	19,850
STL-10	10	5,000	8,000	EuroSAT	10	10,000	5,000
ImageNet	1000	1,281,167	50,000	Resisc45	45	25,200	6,300
Food101	101	75,750	25,250	GTSRB	43	26,640	12,630
Flowers	102	2,040	6,149	PCAM	2	294,192	32,768
Cars	196	8,144	8,041	UCF101	101	9,537	3,783
Aircraft	100	6,667	3,333	CLEVR	8	2,000	500
Pets	37	3,680	3,669	HatefulMemes	2	8,500	500
MNIST	10	60,000	10,000	SST	2	7,792	1,821

Table 5. The prompts used for zero-shot CLIP.

Simple ImageNet Templates:
1. itap of a [class].
2. a bad photo of the [class].
3. a origami [class].
4. a photo of the large [class].
5. a [class] in a video game.
6. art of the [class].
7. a photo of the small [class].

for each newly added dataset and overwrite the *__getitem__* method accordingly.

Text-Aided Image Clustering. TAC [22] also proposes leveraging external knowledge in the textual space. Unlike SIC, TAC computes the text features of each image using noun features, without assigning explicit pseudo-labels. We utilize the open-source code from TAC [6]. We keep the default parameters. We build the *dataloader* for the extra datasets by referencing the *dataloader* in TURTLE.

TURTLE. TURTLE [11] enables unsupervised transfer from pretrained models to perform image clustering. It identifies the optimal dataset labeling by maximizing the margins of linear classifiers in the space of single or multiple pretrained models. It is compatible with any pretrained representations. We utilize the open-source code from TURTLE [7]. In the original setup, TURTLE receives image features generated from different pretrained image encoders and trains in multiple spaces. In this paper, we use features from the CLIP image encoder as one space, and the text-enhanced features from TAC and the knowledge-enhanced features from KEC as another space.

9.2. Implementation of the Proposed Method

We precompute the features of the input images for all datasets and the features of the nouns from WordNet before all the experiments, using the `batch_size` of 8192.

Code and more implementation details will be available. The comparison methods are integrated into our project using their original code. Our approach offers good adaptability and can quickly be applied to different clustering strategies via configuration file modifications.

In all experiments, we extract knowledge from `GPT-4o` and set the `temperature` to 0.1. Specifically, we utilize the `AsyncOpenAI` interface for asynchronous processing to improve efficiency, setting `max_concurrent` to 20 within Python's `async` framework.

[6] https://github.com/XLearning-SCU/2024-ICML-TAC
[7] https://github.com/mlbio-epfl/turtle

10. Main Results Across All Datasets

We present the complete numerical results in Table 6. CLIP (k-means) and TURTLE (1-space) utilize only visual space knowledge, while zero-shot CLIP incorporates ground-truth label knowledge from the textual space. SIC, TAC (no train), and KEC utilize textual space knowledge differently. Additionally, we apply two training strategies for the multi-space knowledge: TAC and TURTLE.

It is worth emphasizing again that, compared to KEC and other methods, CLIP (zero-shot) leverages additional fine-grained information, *i.e.* the ground-truth labels for each image (*e.g.*, 'Boeing 747', 'Audi A4'), which effectively simplifies the task setting. Therefore, CLIP (zero-shot) could be considered as a loosely constrained upper bound on performance. Our proposed method constructs discriminative textual knowledge and achieves superior performance over existing baselines across a wide range of datasets. In many cases, it even surpasses CLIP (zero-shot), further demonstrating the effectiveness and generality of our method.

We also observe that the extent of performance improvement of KEC varies across different datasets.
1. For higher-level or subjective categories, the improvements are less pronounced compared to other datasets. This suggests that such tasks may require a certain degree of human-machine collaboration to better construct effective textual knowledge (which will be further discussed in the Limitations and Future Work section).
2. For some fine-grained datasets, such as Aircraft and Cars, which focus on specific domains, CLIP (zero-shot) benefits from directly using ground-truth fine-grained labels, enabling it to effortlessly focus on subtle category differences. In contrast, without additional constraints, KEC adopts a more general perspective to interpret these categories, demonstrating strong generalization capability. Consequently, KEC achieves only modest performance gains or may slightly underperform compared to CLIP zero-shot, yet it still surpasses other baseline methods.

As previously mentioned, KEC possesses the ability to construct customized knowledge (by human-machine collaboration). To verify this, we conduct a simple experiment: assuming the task is to distinguish specific types of cars or aircraft, we guide KEC to selectively acquire concepts and attributes related to cars and aircraft and build hierarchical textual knowledge, which we refer to as KEC$^+$. As shown in Table 8. Without access to ground-truth labels, KEC$^+$ achieves further performance improvements through simple human-machine interaction and customization.

Table 6. Comparison results across all datasets.

Dataset	**CIFAR-10**			**CIFAR-100**			**STL-10**			**DTD**			**UCF-101**		
Metrics	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**
CLIP (k-means)	73.7	80.4	64.6	59.7	43.4	29.1	91.8	94.3	89.2	58.6	45.4	29.0	80.7	59.5	51.4
TAC (no train)	81.4	90.4	80.3	65.0	50.1	35.7	92.5	94.8	89.9	60.3	48.2	31.4	80.9	61.8	52.0
KEC (no train)	81.9	90.7	80.6	66.4	51.8	37.3	95.1	97.9	95.5	60.7	47.4	31.5	81.9	62.5	53.2
SIC	84.7	92.6	84.4	63.2	47.6	34.8	95.3	98.1	95.9	59.6	45.9	30.5	81.4	65.3	56.7
TAC	82.9	91.5	82.3	67.5	56.1	40.8	95.6	98.2	96.1	60.8	47.8	32.4	81.3	67.4	58.2
KEC_{TAC}	84.1	92.3	84.0	68.0	57.1	41.3	95.8	98.3	96.3	62.5	51.3	36.0	81.6	67.6	58.7
TURTLE (1-space)	78.6	86.5	75.1	60.8	45.0	33.1	95.8	98.4	96.4	62.9	52.9	36.7	80.9	67.1	57.1
$\text{TAC}_{\text{TURTLE}}$	83.5	91.9	83.2	62.0	46.4	34.6	95.3	98.0	95.7	63.3	52.9	36.8	81.9	69.2	59.4
$\text{KEC}_{\text{TURTLE}}$	83.7	92.1	83.5	62.9	47.6	35.3	96.1	98.5	96.7	63.1	52.8	36.7	82.8	70.0	60.7
CLIP (zero-shot)	80.7	90.0	79.3	69.9	65.0	44.7	93.8	97.0	93.7	56.1	42.6	26.6	80.3	63.7	50.2
Dataset	**ImageNet**			**Food101**			**SUN397**			**Cars**			**Aircraft**		
Metrics	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**
CLIP (k-means)	72.3	38.9	27.1	71.1	60.4	47.8	76.3	50.0	38.4	67.1	35.9	24.9	49.4	21.5	11.6
TAC (no train)	77.5	48.4	34.5	72.5	61.5	48.1	78.6	54.4	42.5	64.7	32.6	21.3	48.6	21.6	10.7
KEC (no train)	77.7	48.6	35.5	74.6	66.2	53.3	78.2	54.0	42.3	68.3	37.6	26.6	50.6	22.5	13.3
SIC	77.2	47.0	34.3	74.1	62.4	51.8	76.1	51.3	38.2	66.5	33.7	24.1	48.8	22.1	12.0
TAC	78.2	54.4	39.6	74.9	68.4	54.2	74.6	41.2	32.3	60.5	25.7	16.5	44.9	19.1	9.3
KEC_{TAC}	78.4	55.3	40.1	75.9	70.0	56.3	78.7	54.1	42.6	65.4	33.7	22.6	48.8	22.7	12.9
TURTLE (1-space)	66.4	25.6	15.6	72.9	64.3	52.5	77.8	54.7	42.9	69.4	41.4	30.2	49.6	24.0	13.6
$\text{TAC}_{\text{TURTLE}}$	65.6	23.8	14.5	71.5	62.2	49.7	76.9	53.3	40.5	69.4	41.5	30.2	48.8	23.8	13.0
$\text{KEC}_{\text{TURTLE}}$	68.7	30.6	19.9	73.8	66.9	54.2	77.7	54.6	42.1	69.7	42.4	30.4	49.8	24.5	13.7
CLIP (zero-shot)	81.0	63.6	45.3	82.7	83.3	69.8	80.3	64.4	47.5	77.4	59.3	43.3	48.8	22.0	11.3
Dataset	**Pets**			**Flowers**			**MNIST**			**Eurosat**			**Resisc45**		
Metrics	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**
CLIP (k-means)	65.0	51.3	40.5	86.5	71.5	67.7	49.4	57.8	38.6	53.7	61.9	43.7	71.8	64.1	50.5
TAC (no train)	79.2	65.8	57.8	84.5	69.4	64.8	36.9	45.5	24.7	48.8	60.9	34.5	70.7	58.3	45.6
KEC (no train)	81.2	67.8	63.3	87.3	72.8	67.5	46.9	53.7	35.2	54.6	62.7	42.8	74.3	62.4	51.5
SIC	67.7	51.6	42.6	67.7	43.1	34.0	40.8	50.1	32.5	61.5	67.2	53.5	75.7	67.7	57.0
TAC	83.6	77.9	69.1	80.4	64.2	59.6	42.7	54.6	34.0	53.0	67.8	45.7	72.9	68.0	53.0
KEC_{TAC}	84.8	79.7	71.5	85.4	71.6	66.8	45.4	56.2	36.8	56.7	69.8	48.0	75.6	70.6	57.1
TURTLE (1-space)	71.5	60.9	48.9	90.7	87.2	79.8	42.8	49.6	34.4	57.9	63.6	48.2	75.3	70.9	57.0
$\text{TAC}_{\text{TURTLE}}$	73.7	63.4	52.5	90.8	86.8	79.1	41.5	50.7	33.6	60.0	67.7	51.8	74.4	69.4	55.4
$\text{KEC}_{\text{TURTLE}}$	81.0	72.4	63.3	92.4	88.4	82.7	47.8	57.8	39.4	63.0	76.8	57.4	76.4	73.6	59.3
CLIP (zero-shot)	87.6	84.9	75.4	79.4	67.5	58.7	28.4	38.1	11.9	41.0	45.6	26.4	64.4	53.4	35.3
Dataset	**GTSRB**			**PCAM**			**CLEVR**			**HatefulMemes**			**SST**		
Metrics	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**
CLIP (k-means)	52.5	32.1	23.1	10.2	63.3	7.0	18.9	27.6	8.1	2.0	58.2	2.4	0.4	53.7	0.5
TAC (no train)	48.3	31.0	20.2	1.5	56.7	1.8	11.8	23.6	5.4	1.2	56.4	1.4	0.1	51.9	0.1
KEC (no train)	51.9	32.6	22.6	10.0	63.6	7.4	16.3	26.6	8.9	1.6	57.4	2.0	0.5	53.8	0.6
SIC	52.3	36.4	28.3	0.0	51.1	0.0	5.2	20.6	1.5	0.1	51.8	0.0	0.5	54.1	0.6
TAC	40.5	26.7	16.8	0.0	50.3	0.0	8.4	22.4	3.6	0.6	54.6	0.6	0.1	52.0	0.1
KEC_{TAC}	48.6	29.6	20.1	9.0	63.2	7.1	15.9	25.4	6.8	2.3	59.0	3.0	0.9	55.5	1.1
TURTLE (1-space)	51.1	33.9	25.1	0.0	51.3	0.0	16.0	24.6	7.0	1.1	55.6	1.8	0.2	52.4	0.4
$\text{TAC}_{\text{TURTLE}}$	48.1	29.9	22.1	0.0	50.3	0.0	16.6	25.8	7.5	1.4	56.0	2.0	0.2	52.4	0.5
$\text{KEC}_{\text{TURTLE}}$	48.8	32.4	23.5	8.9	62.7	6.8	18.8	28.6	8.0	2.0	58.4	2.6	0.8	55.2	1.0
CLIP (zero-shot)	46.2	32.3	22.3	3.1	52.1	0.2	16.4	4.4	8.0	0.5	53.4	0.3	0.0	51.0	0.0

Table 7. Ablation results across all datasets. *Con.* and *Des.* represent the name and its description of a concept, respectively. *UA* and *BA* represent the uni-concept attribute and bi-concept attribute.

Con.	Des.	UA	BA	CIFAR-10			CIFAR-100			STL-10			DTD			UCF-101		
				NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
✓				80.9	89.7	79.7	64.5	49.4	36.0	94.0	97.0	94.5	58.0	43.6	26.4	80.5	60.2	51.8
	✓			80.1	89.0	78.3	64.2	47.9	33.4	93.7	96.7	94.0	57.4	46.3	28.8	80.7	60.7	52.7
✓	✓			81.4	90.1	80.5	64.6	49.1	36.0	92.4	96.5	94.4	58.2	46.8	28.7	80.7	61.2	51.3
✓	✓	✓		81.9	90.5	80.2	66.0	50.3	36.6	95.1	97.9	95.5	59.8	46.2	30.1	81.9	60.8	52.0
✓	✓		✓	81.9	90.6	80.3	65.7	50.9	36.4	95.1	97.9	95.5	60.2	46.8	31.1	81.7	61.8	53.1
✓	✓	✓	✓	81.9	90.7	80.6	66.4	51.8	37.3	95.1	97.9	95.5	60.7	47.4	31.5	81.9	62.5	53.2

Con.	Des.	UA	BA	ImageNet			Food101			SUN397			Cars			Aircraft		
				NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
✓				75.5	45.3	33.4	73.7	65.7	52.5	77.2	52.4	40.2	67.6	36.5	24.3	49.5	22.5	12.5
	✓			74.7	45.2	33.0	73.2	63.2	50.8	77.0	52.6	40.6	66.0	35.7	23.8	48.9	21.4	10.8
✓	✓			75.8	45.3	33.1	73.9	66.0	52.4	77.6	53.0	41.6	68.3	36.4	25.8	49.1	22.5	12.2
✓	✓	✓		76.9	46.8	35.1	74.7	66.2	53.5	78.2	53.8	42.3	68.3	37.5	26.4	50.4	22.4	13.1
✓	✓		✓	77.0	47.0	35.5	74.8	66.7	53.7	78.2	53.7	42.5	68.2	37.8	26.6	50.6	22.7	13.2
✓	✓	✓	✓	77.7	48.6	35.5	74.6	66.2	53.3	78.2	54.0	42.3	68.3	37.6	26.6	50.6	22.5	13.3

Con.	Des.	UA	BA	Pets			Flowers			MNIST			Eurosat			Resisc45		
				NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
✓				78.9	66.0	61.1	84.9	68.9	62.3	45.4	50.6	32.7	52.2	61.1	41.7	71.7	63.8	50.6
	✓			78.9	64.9	60.9	85.0	68.6	62.0	45.6	50.9	31.9	51.9	60.9	40.4	72.1	64.2	50.6
✓	✓			79.0	64.7	60.5	85.5	70.8	63.6	45.9	51.7	32.6	52.0	62.0	41.5	72.2	61.8	51.2
✓	✓	✓		80.5	67.3	63.1	87.3	72.6	67.6	46.7	53.6	35.0	54.4	62.3	42.4	74.3	63.2	52.3
✓	✓		✓	80.2	66.7	62.8	86.8	72.2	65.2	45.4	49.0	29.5	54.5	62.4	42.4	74.6	62.8	52.0
✓	✓	✓	✓	81.2	67.8	63.3	87.3	72.5	67.5	46.9	53.7	35.2	54.6	62.7	42.8	74.3	62.4	51.5

Con.	Des.	UA	BA	GTSRB			PCAM			CLEVR			HatefulMemes			SST		
				NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
✓				48.9	28.6	21.2	9.0	62.6	6.4	14.2	23.2	5.5	1.6	57.2	1.9	0.5	53.9	0.6
	✓			49.6	29.7	20.5	8.1	62.7	6.5	15.4	24.2	6.3	1.4	56.8	1.7	0.6	54.3	0.7
✓	✓			49.9	30.1	21.4	9.0	62.6	6.4	15.8	25.8	8.0	1.6	57.2	1.9	0.6	54.1	0.6
✓	✓	✓		51.8	32.4	22.5	9.9	63.6	7.4	16.0	26.6	8.6	1.6	57.4	2.0	0.5	54.0	0.6
✓	✓		✓	51.9	32.4	22.7	10.0	63.7	7.5	16.3	26.6	8.9	1.7	57.6	2.1	0.5	53.8	0.5
✓	✓	✓	✓	51.9	32.6	22.6	10.0	63.6	7.4	16.3	26.6	8.9	1.6	57.4	2.0	0.5	53.8	0.6

Table 8. Results of KEC with targeted orientation.

Methods	Aircraft			Cars		
	NMI	ACC	ARI	NMI	ACC	ARI
CLIP (zero-shot)	48.8	22.0	10.7	77.4	59.3	43.3
KEC	50.6	22.5	13.3	68.3	37.6	26.6
KEC$^+$	52.1	24.2	15.8	74.8	53.7	42.6

11. More Ablation Studies

11.1. Results of ablation studies across all datasets

We present the results of ablation experiments for the proposed method across each dataset in Table 7. We gradually introduce different levels of knowledge and their components using the LLMs.

11.2. Further analysis on parameter selections

Parameters in Image-Text Mapping. During the construction of image-text mappings, we aim to associate each image with its relevant semantics (nouns in WordNet). To preserve a strong generalization capability, we intentionally avoid meticulous parameter tuning at this stage. Our method aims to distill hierarchical and discriminative knowledge from the initial textual knowledge, which contains considerable semantic redundancy. Following the settings in TAC, we set the number of clusters as $N_v/300$ (where N_v denotes the number of images), and select the TopK=5 most relevant nouns for each image.

To evaluate the sensitivity of the mapping stage in KEC to these parameters, we systematically varied them and conducted corresponding experiments. The results, shown in Table 9 and Table 10, indicate that KEC exhibits

Table 9. Results for different numbers of selected nouns.

TopK	Flowers			Pets			CLEVR		
	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
1	86.9	72.1	69.2	80.1	67.9	62.1	17.0	27.0	7.8
3	87.0	72.6	66.9	80.0	68.2	62.7	16.0	26.6	8.6
5	87.3	72.8	67.5	81.2	67.8	63.3	16.3	26.6	8.9
10	87.7	73.6	68.2	80.0	67.5	62.6	16.1	26.4	8.7

Table 10. Results for different cluster numbers in the initial Image-Text Mapping.

	Flowers			Pets			CLEVR		
	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
$N_v/50$	87.0	72.5	67.2	81.3	67.9	63.5	16.5	26.7	8.9
$N_v/100$	87.4	72.7	67.3	81.2	67.8	63.2	16.3	26.7	8.9
$N_v/300$	87.3	72.8	67.5	81.2	67.8	63.3	16.3	26.6	8.9
$N_v/500$	87.6	73.0	67.8	81.1	67.5	62.9	15.8	26.1	8.1

strong robustness under different configurations (*i.e.*, textual knowledge generated with varying degrees of semantic redundancy through different mapping strategies). Although different parameter settings may cause slight performance variations across datasets, the overall impact remains limited. Therefore, meticulously tuning parameters for each dataset is labor-intensive, marginally beneficial, and ultimately unnecessary. This further demonstrates the strong generalization capability of KEC across diverse datasets.

Weight to balance visual and textual similarity. To construct multi-modal similarity, we analyze the impact of the fusion weight α in Equation 2. Specifically, we vary α from 1.0 to 0.0, gradually increasing the contribution of textual knowledge in the similarity computation. The results, as shown in Table 11, lead to the following observations:

- When $\alpha = 1.0$ (*i.e.*, only using visual similarity), it fails to leverage any textual knowledge. As a result, the constructed text space becomes nearly identical to the visual space, yielding performance comparable to the baseline without improvement.
- As α decreases, performance initially improves and then declines, indicating that at this stage, textual knowledge relies on the guidance of visual knowledge to extract more discriminative concepts. When textual knowledge dominates, redundant or mismatched granularity of extracted nouns (either too fine-grained or overly broad) tends to disrupt semantic distillation, resulting in degraded performance. Notably, when $\alpha = 0$, *i.e.* relying entirely on textual knowledge, the results are the worst.
- To further determine the optimal setting, we conduct additional experiments with $\alpha = 0.7$ and 0.9 around 0.8. Considering the trade-off between performance and generalization across datasets, we empirically set $\alpha = 0.8$ as the default value.

Table 11. Performance analysis based on varying the weight value of visual and textual similarity. The **bold** numbers indicate the best results. In practice, we set α to 0.8.

α	Flowers			Pets			CLEVR		
	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
1.0	86.5	72.9	67.9	78.2	65.3	58.6	15.5	25.6	6.9
0.9	**87.8**	**73.1**	**68.2**	81.0	67.7	62.9	15.1	25.4	6.7
0.8	87.3	72.8	67.5	**81.2**	**67.8**	**63.3**	**16.3**	**26.6**	**8.9**
0.7	86.5	72.9	67.9	80.5	67.0	62.7	16.0	26.6	8.5
0.6	84.0	67.7	64.5	75.5	62.3	55.1	15.1	25.8	8.0
0.4	76.5	62.7	59.7	70.5	56.0	47.8	14.7	24.8	6.7
0.2	70.6	57.9	52.2	66.5	50.6	42.4	13.1	24.2	6.3
0.0	69.5	56.4	50.7	65.5	50.0	41.7	11.5	22.8	4.5

Table 12. Performance analysis based on varying the threshold value used in KEC. The **bold** numbers indicate the best results. The threshold β is empirically set to 0.8.

β	Flowers			Pets			CLEVR		
	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
0.9	86.7	71.4	65.9	79.9	64.9	60.7	15.4	24.8	6.7
0.8	87.3	**72.8**	**67.5**	**81.2**	67.8	**63.3**	**16.3**	**26.6**	**8.9**
0.7	**87.5**	71.7	67.0	80.0	**68.8**	62.8	14.7	25.4	6.5
0.6	86.3	71.2	64.9	80.5	67.2	62.1	14.5	24.2	5.9
0.5	83.5	67.2	63.8	78.8	66.1	60.9	14.0	24.4	6.1

Threshold in knowledge construction. For threshold selection, we conduct experiments by gradually lowering the threshold from 0.9. The results are shown in Table 12. As the threshold decreases, more concepts and attributes are extracted, which significantly increases computational cost and time consumption. However, this does not lead to continuous performance improvement and may even result in performance degradation under certain settings. Considering the trade-off between performance and computational efficiency across different datasets, we empirically set the threshold to 0.8 as the default value.

11.3. Compatibility with traditional methods

To demonstrate the compatibility of our method, we evaluate KEC using three additional traditional clustering algorithms beyond K-Means: spectral clustering, agglomeration clustering, and bisecting K-Means. All experiments are conducted on the concatenated features $[\mathbf{x}_i; \kappa_i]$, with the number of clusters set to match the ground-truth class count. As shown in Table 13, KEC consistently achieves strong performance across all clustering algorithms. These results highlight that the knowledge-enhanced features produced by KEC are broadly compatible with a range of standard clustering algorithms, offering flexibility and ease of

Table 13. Clustering performance of KEC using different downstream traditional clustering algorithms.

	Average		
	NMI	**ACC**	**ARI**
K-Means	58.0	56.6	**38.5**
Spectral Clustering	57.0	55.8	38.4
Agglomerative Clustering	**58.2**	**56.9**	38.2
Bisecting K-Means	57.2	56.2	37.6

integration into diverse practical scenarios.

11.4. Analysis of robustness for pretrained models

To assess the generalizability of KEC, we evaluate it under three different vision-language backbones. These backbones vary significantly in architecture and capacity. Specifically, CLIP ResNet-50 utilizes a convolutional network. CLIP ViT-B/16 introduces a Transformer-based model, and ImageBind-Huge is a recent multi-modal foundation model that jointly embeds multiple modalities, providing a stronger and more generalized representation space. As shown in Table 14, KEC consistently achieves the best performance across all backbones and metrics, significantly outperforming both the vanilla CLIP (K-Means) and TAC. These results validate the transferability of our hierarchical knowledge construction approach across a wide spectrum of model architectures and data domains.

11.5. Knowledge construction with various LLMs

We investigate the effect of using different LLMs in the knowledge construction stage of KEC. Specifically, we compare several state-of-the-art LLMs from diverse providers, including: Claude-3-7-Sonnet (Anthropic), Gemini-2.0-Flash (Google), Deepseek-V3, Qwen-Turbo, Qwen2.5-7B-Instruct, and Qwen-3-0.6B. The latter three represent models with relatively small parameters, allowing us to evaluate the feasibility of lightweight deployment. KEC is robust with open-source and small LLMs (even with a 0.6B model). As shown in Table 15, all models yield competitive performance, confirming the robustness of our hierarchical concept-attribute construction method. These results confirm that our method does not rely on any specific LLM backbone and can generalize across different model families and deployment constraints.

12. Constructed Knowledge Space

12.1. Reducing Redundancy Compared to TAC.

To better understand the effectiveness and efficiency of hierarchical knowledge construction, we compare the size of the knowledge space produced by TAC and our method (KEC) across 20 datasets. As shown in Figure 5, KEC significantly reduces the number of textual elements in the constructed knowledge, transforming hundreds or even thousands of nouns in TAC into a compact set of concepts. For example, on ImageNet, the number of textual tokens drops from 13,278 to 1,341, and similar reductions are observed across datasets such as SUN397 (from 4,691 to 686) and Aircraft (from 1,253 to 120). This demonstrates that our hierarchical knowledge construction mechanism effectively eliminates semantic redundancy and consolidates overlapping or overly fine-grained nouns into unified, higher-level concepts. Based on these representative concepts, KEC further explores discriminative attributes to assist in clustering.

12.2. Visualization of knowledge enhanced features.

To qualitatively assess the effect of knowledge-enhanced features, we visualize the feature before and after enhancement using t-SNE across six datasets: PCAM, STL10, CIFAR10, Pets, Food101, and UCF101. The left panels show the raw CLIP visual features, while the right panels show the corresponding knowledge-enhanced features generated by KEC. For datasets with a large number of classes, we visualize only the first 10 categories to improve clarity. As shown in Figure 6, the knowledge-enhanced features provided by KEC lead to better semantic separability, with tighter intra-cluster cohesion and clearer inter-cluster boundaries. This effect is especially prominent on fine-grained datasets like Pets and Food101, where textual knowledge helps refine class-specific distinctions that may not be evident from visual features alone. These results provide intuitive evidence that KEC injects structured, semantically meaningful textual guidance into the visual space, thus aiding downstream clustering performance.

12.3. Examples of the hierarchical knowledge.

To provide a clearer understanding of the knowledge constructed by our method, we present examples of hierarchical knowledge extracted from two representative datasets. As shown in Table 16, the hierarchical knowledge includes three parts: (1) **Concepts**, abstracted from semantically similar noun clusters; (2) **Uni-Concept Attributes**, which are representative and visually grounded features associated with individual concepts; and (3) **Bi-Concept Attributes**, which are contrastive properties mined between pairs of similar concepts to enhance fine-grained discrimination.

For each dataset, we showcase 20 representative concepts and 10 attributes for each type. For example, in CIFAR10, concepts such as ‘Emergency Response Vehicles’ or ‘Ecological Diversity of Amphibians and Reptiles’ are abstracted from low-level nouns. Uni-concept Attributes like ‘Prominent cargo area’ or ‘Beak shape and size’ highlight core visual traits, while Bi-concept attributes such as ‘Mode of travel’ or ‘Limb configuration’ aid in separating

Table 14. Performance of KEC under different pretrained visual-language models. We report average results across 20 datasets, along with detailed scores on three representative datasets. **Bold** numbers indicate the best performance within each setting.

Model	Method	Average			Flowers			Pets			CLEVR		
		NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
CLIP ResNet-50	**CLIP (K-Means)**	50.5	47.9	28.4	83.6	70.4	63.6	55.9	38.8	27.5	**7.9**	20.2	**2.5**
	TAC (no train)	50.9	48.4	29.1	83.3	67.0	60.4	76.0	60.1	52.1	6.8	20.0	1.8
	KEC (no train)	**51.9**	**49.3**	**30.5**	**84.1**	**71.1**	**64.2**	**78.6**	**67.9**	**58.1**	7.1	**20.6**	2.0
CLIP ViT-B/16	**CLIP (K-Means)**	58.1	56.0	38.0	86.1	72.1	68.4	69.1	53.3	44.8	12.6	26.0	5.8
	TAC (no train)	56.5	54.2	36.5	86.2	66.1	62.3	83.1	70.2	64.0	9.7	23.2	4.0
	KEC (no train)	**60.0**	**59.4**	**41.6**	**90.1**	**75.9**	**72.0**	**85.2**	**77.0**	**69.8**	**14.5**	**26.2**	**6.8**
ImageBind-Huge	**CLIP (K-Means)**	61.0	58.8	42.2	**93.8**	72.2	**72.9**	85.2	70.6	67.3	16.4	25.8	**9.5**
	TAC (no train)	62.1	59.6	43.0	89.3	70.9	70.1	85.7	68.9	65.6	15.3	24.6	7.2
	KEC (no train)	**64.8**	**62.7**	**47.5**	92.0	**72.8**	**72.9**	**88.4**	**73.1**	**71.4**	**18.2**	**26.2**	9.2

Table 15. Performance comparison of KEC with different LLMs for knowledge construction. The LLMs are from different providers and with various model sizes. The results demonstrate the robustness of KEC.

LLMs	Flowers			Pets			CLEVR		
	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
Claude-3-7-Sonnet	86.5	72.9	67.9	79.8	67.1	59.6	16.7	26.8	8.6
Gemini-2.0-Flash	86.5	72.9	67.9	81.0	67.8	62.7	17.0	26.8	9.0
Deepseek-V3	80.1	65.9	60.6	86.0	72.7	66.0	16.1	26.0	8.1
Qwen-Turbo	80.6	70.8	62.4	86.1	70.3	64.9	16.0	27.4	8.8
Qwen2.5-7B-Instruct	81.0	71.9	61.8	86.5	72.9	67.9	15.9	26.2	8.1
Qwen3-0.6B	79.8	69.1	60.2	86.0	71.1	65.7	16.1	26.7	8.2

fine-grained categories.

Similarly, in the Flowers dataset, concepts cover a wide range of botanical categories, while Uni-Concept attributes emphasize textural and morphological features (e.g., "Waxy texture" or "Serrated leaves"). Bi-Concept attributes, such as "Visual Style" or "Number of blooms," further help differentiate visually similar floral species. These structured knowledge components form the basis of our knowledge-enhanced feature representation.

This qualitative evidence underscores the semantic richness and interpretability of the constructed knowledge, which not only improves clustering performance but also contributes to the explainability of the clustering process.

13. Limitation and Future Work

While our proposed KEC achieves strong and consistent performance across diverse datasets and clustering settings, we highlight several limitations that also point toward promising future extensions:

Concept quantity may not always be larger than the target cluster number. In most cases, the number of concepts generated by our framework exceeds the target number of clusters, preventing overly coarse semantic summarization for concepts. However, in rare cases, such as CIFAR-100, only 38 concepts are formed. It is fewer than the dataset's 100 fine-grained classes. Notably, this number still exceeds the 20 coarse-grained categories defined in CIFAR-20, which uses the same image set. This observation suggests that our method tends to generate semantically meaningful concepts, even without hard constraints. In future work, incorporating user-defined granularity or constraints into the concept construction process could support more adaptive and goal-driven clustering.

Limited improvement for high-level or subjective categories. The improvement is relatively modest in datasets where categories are defined by abstract or high-level semantics, such as HatefulMemes. The labels reflect whether an image is offensive. KEC may lack the precise guidance needed to construct appropriate textual knowledge. This points to an opportunity to actively involve users in knowledge construction, for example by supplying domain-specific vocabulary, mapping terms to visual regions, or refining the concept-attribute hierarchy. Such human-in-the-loop strategies could enable more accurate alignment between textual semantics and high-level visual concepts.

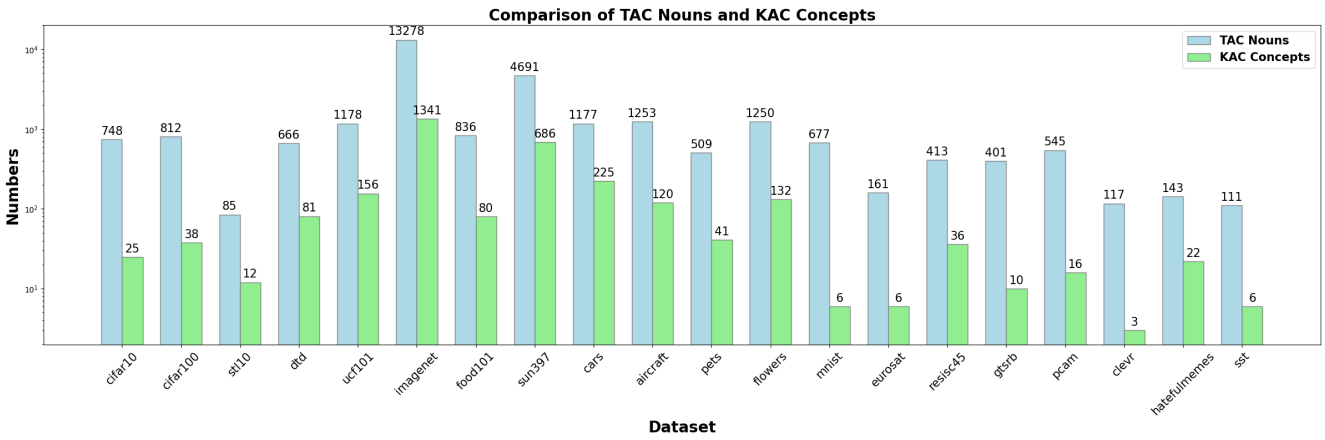

Figure 5. KEC effectively reduces semantic redundancy compared to naive use of textual knowledge.

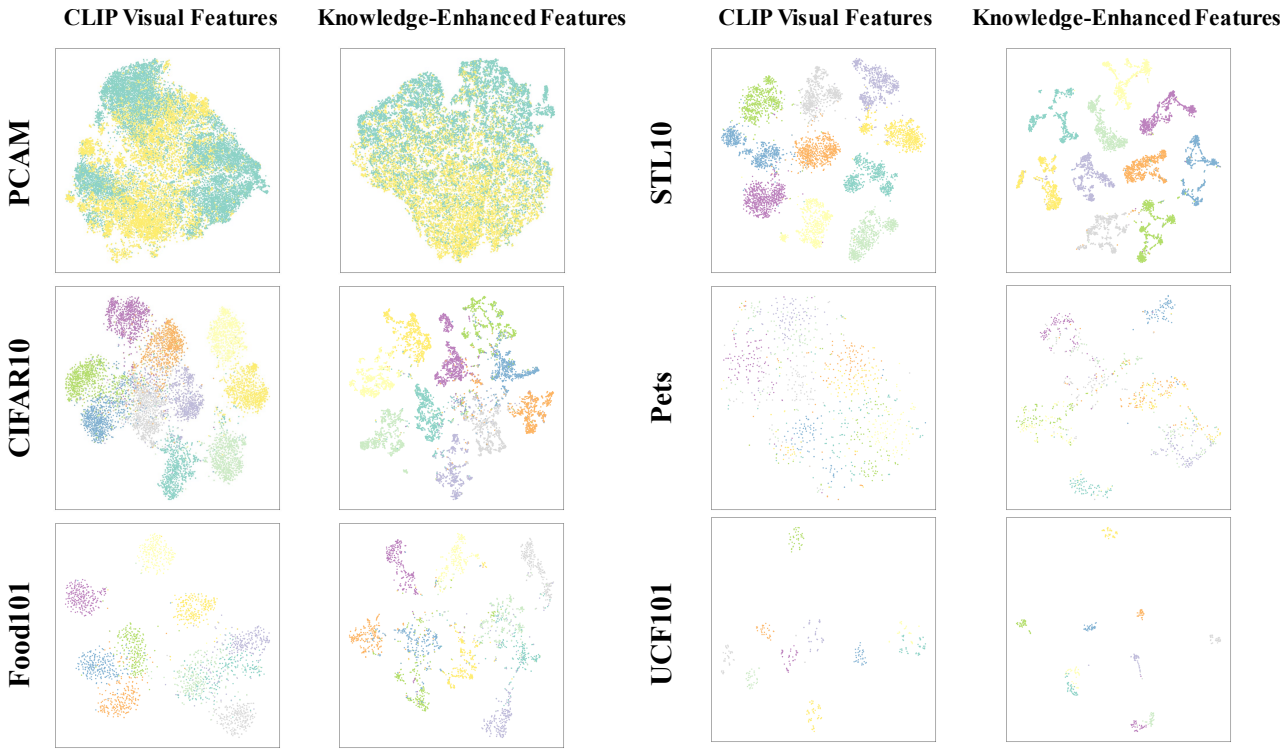

Figure 6. t-SNE visualization of features before and after knowledge enhancement across six representative datasets. The left column shows CLIP visual features, and the right column shows knowledge-enhanced features obtained by KEC. For datasets with large numbers of classes, only the first 10 classes are visualized for clarity.

Potential for Interactive and Customized Knowledge Construction. KEC is modularity and flexibility. Allowing users to define or adjust certain concepts and attributes, or to specify domain-specific requirements, could further improve clustering performance. Exploring such interactive, customized knowledge construction mechanisms represents a compelling future direction.

In summary, we present a hierarchical knowledge construction method that opens up new possibilities for leveraging textual space in image clustering. While our current system operates automatically and performs well across tasks, future extensions incorporating user input, domain adaptation, and controllable knowledge granularity could further enhance its applicability and effectiveness.

Table 16. Examples of the hierarchical knowledge constructed by KEC. We show 20 representative concepts, 10 Uni-Concept and Bi-Concept attributes, respectively. These knowledge elements are automatically derived via LLMs and serve to inject structured semantic guidance for clustering.

Dataset	**Hierarchical Knowledge**
CIFAR10	**Concepts:** Commercial and Utility Vehicles, Bird Species, Water and Land Transportation, aviation and air travel, Automobiles, Maritime Vessels, Musical Heritage, Equestrian Sports, Ecological Diversity of Amphibians and Reptiles, Pets, Feline and Animal, Antelope and Deer Species, emergency response vehicles, Equine Management and Culture, Aircraft, Dog Breeds, Maritime Vessels and Shipping, Fish Species, aviation and avian agility, Plant Ecology **Uni-Concept Attributes:** Prominent cargo area or flatbed design, Feathers covering the body, colors on animal fur or skin, Beak shape and size, Streamlined fuselage with a pointed nose, Large and swept-back wings with distinctive winglets, wheels or tracks on vehicles, Large hull structure with a streamlined shape, sails or masts or other rigging elements, Bright and high-visibility color schemes **Bi-Concept Attributes:** Feathered or fur-covered bodies, Color Variation, Mode of travel, Number of wheels, cargo or shipping containers on the vessels, Size and domestication level, representations and cultural symbols, Body structure and limb configuration, Color scheme and markings, landmark achievements
Flowers	**Concepts:** Ornamental Plants and Flowering Species, Yellow Flora, Berry Cultivation, Flowering Plants, Primulaceae and Related Plants, Fritillaria, Flora of Tropical and Subtropical Regions, Wildflowers and Mosses, Wildflowers and Mosses, Irises and Associated Flora, Irises and Associated Flora, Goldenrod Species, Squash and Related Plants, Dianthus, Hymenoptera and Related Insects, Phyllostachys, Hibiscus and Mallow Plants, Native North American Flora, Penstemon and Antirrhinum Species, Arum Lilies and Egrets **Uni-Concept Attributes:** Vibrant and varied color palettes, Thick and serrated leaves, Waxy texture, a complex and organized structure, Black and white photographic style, Furry or feathery texture, palmate or lobed leaves with a more pronounced venation pattern, ovate or elliptical leaves with a smoother margin, Large white and petal-like ray florets, Colorful and diverse petal shapes and patterns **Bi-Concept Attributes:** Flower shape and color, Colorful and intricate floral patterns, Leaf shape and arrangement, Physical Form, Number of blooms, Color and texture, Vibrancy of Colors, Visual Style, abstract or stylized expressions, like paintings or sculptures

Acknowledgement. This work was supported by the National Natural Science Foundation of China (U23B2057).

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In *ECCV (6)*, pages 446–461. Springer, 2014. 1

[2] Shaotian Cai, Liping Qiu, Xiaojun Chen, Qin Zhang, and Longteng Chen. Semantic-enhanced image clustering. In *AAAI*, pages 6869–6878. AAAI Press, 2023. 1, 3, 4, 5, 6

[3] Wei Chang, Huimin Chen, Feiping Nie, Rong Wang, and Xuelong Li. Tensorized and compressed multi-view subspace clustering via structured constraint. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):10434–10451, 2024. 3

[4] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 397–406, 2021. 2

[5] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE*, 105(10):1865–1883, 2017. 1

[6] Tianzhe Chu, Shengbang Tong, Tianjiao Ding, Xili Dai, Benjamin David Haeffele, René Vidal, and Yi Ma. Image clustering via the principle of rate reduction in the age of pretrained models. In *ICLR*. OpenReview.net, 2024. 1, 3

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613. IEEE Computer Society, 2014. 1

[8] Adam Coates, Andrew Y. Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, pages 215–223. JMLR.org, 2011. 1

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE Computer Society, 2009. 1

[10] Tianjiao Ding, Shengbang Tong, Kwan Ho Ryan Chan, Xili Dai, Yi Ma, and Benjamin D. Haeffele. Unsupervised manifold linearizing and clustering. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 5427–5438. IEEE, 2023. 3

[11] Artyom Gadetsky, Yulun Jiang, and Maria Brbic. Let go of your labels with unsupervised transfer. In *ICML*. OpenReview.net, 2024. 1, 3, 5, 6

[12] Xiwen Geng, Suyun Zhao, Yixin Yu, Borui Peng, Pan Du, Hong Chen, Cuiping Li, and Mengdie Wang. Personalized clustering via targeted representation learning. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 16790–16798. AAAI Press, 2025. 3

[13] Longkun Guo, Chaoqi Jia, Kewen Liao, Zhigang Lu, and Minhui Xue. Efficient constrained k-center clustering with background knowledge. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 20709–20717. AAAI Press, 2024. 3

[14] Wei He, Zhiyuan Huang, Xianghan Meng, Xianbiao Qi, Rong Xiao, and Chun-Guang Li. Graph cut-guided maximal coding rate reduction for learning image embedding and clustering. In *Computer Vision - ACCV 2024 - 17th Asian Conference on Computer Vision, Hanoi, Vietnam, December 8-12, 2024, Proceedings, Part X*, pages 359–376. Springer, 2024. 3

[15] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 12(7):2217–2226, 2019. 1

[16] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 1988–1997. IEEE Computer Society, 2017. 7, 1

[17] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS*, 2020. 1

[18] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, pages 554–561. IEEE Computer Society, 2013. 1

[19] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 1(4), 2009. 1

[20] Sehyun Kwon, Jaeseung Park, Minkyu Kim, Jaewoong Cho, Ernest K. Ryu, and Kangwook Lee. Image clustering conditioned on text criteria. In *ICLR*. OpenReview.net, 2024. 1, 3

[21] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998. 1

[22] Yunfan Li, Peng Hu, Dezhong Peng, Jiancheng Lv, Jianping Fan, and Xi Peng. Image clustering with external guidance. In *ICML*. OpenReview.net, 2024. 1, 3, 4, 5, 6

[23] Honglin Liu, Peng Hu, Changqing Zhang, Yunfan Li, and Xi Peng. Interactive deep clustering via value mining. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 3

[24] Weiwei Liu, Xiao-Bo Shen, and Ivor W. Tsang. Sparse embedded k-means clustering. In *NIPS*, pages 3319–3327, 2017. 1, 2

[25] Yaroslava Lochman, Carl Olsson, and Christopher Zach. Learned trajectory embedding for subspace clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 19092–19102. IEEE, 2024. 2

[26] Ke Ma, Yizhou Fang, Jean-Baptiste Weibel, Shuai Tan, Xinggang Wang, Yang Xiao, Yi Fang, and Tian Xia. Physliquid: A physics-informed dataset for estimating 3d geometry and volume of transparent deformable liquids. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7782–7790, 2026. 3

[27] J. Macqueen. Some methods for classification and analysis of multivariate observations. *Proc. Symp. Math. Statist. and Probability, 5th*, 1, 1967. 1

[28] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013. 1

[29] Ioannis Maniadis Metaxas, Georgios Tzimiropoulos, and Ioannis Patras. Divclust: Controlling diversity in deep clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 3418–3428. IEEE, 2023. 3

[30] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995. 1, 3, 6

[31] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pages 722–729. IEEE Computer Society, 2008. 7, 1

[32] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505. IEEE Computer Society, 2012. 7, 1

[33] J.C. Platt, M. Czerwinski, and B.A. Field. Phototoc: automatic clustering for browsing personal photographs. In *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*, pages 6–10 Vol.1, 2003. 1

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1, 3, 5

[35] Yazhou Ren, Jingyu Pu, Zhimeng Yang, Jie Xu, Guofeng Li, Xiaorong Pu, Philip S. Yu, and Lifang He. Deep clustering: A comprehensive survey. *IEEE Trans. Neural Networks Learn. Syst.*, 36(4):5858–5878, 2025. 1

[36] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 1

[37] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012. 1

[38] Andreas Stephan, Lukas Miklautz, Kevin Sidak, Jan Philip Wahle, Bela Gipp, Claudia Plant, and Benjamin Roth. Text-guided image clustering. In *EACL (1)*, pages 2960–2976. Association for Computational Linguistics, 2024. 1, 3

[39] Li Sun, Zhenhao Huang, Hao Peng, Yujie Wang, Chunyang Liu, and Philip S. Yu. Lsenet: Lorentz structural entropy neural network for deep graph clustering. In *ICML*. OpenReview.net, 2024. 1

[40] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alphaclip: A CLIP model focusing on wherever you want. In *CVPR*, pages 13019–13029. IEEE, 2024. 2

[41] Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *MICCAI (2)*, pages 210–218. Springer, 2018. 1

[42] Zhen Wang, Zhaoqing Li, Rong Wang, Feiping Nie, and Xuelong Li. Large graph clustering with simultaneous spectral embedding and discretization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4426–4440, 2021. 1, 2

[43] Jianxiong Xiao, Krista A. Ehinger, James Hays, Antonio Torralba, and Aude Oliva. SUN database: Exploring a large collection of scene categories. *Int. J. Comput. Vis.*, 119(1): 3–22, 2016. 1

[44] Hailong Yan, Shice Liu, Tao Wang, Xiangtao Zhang, Yijie Zhong, Jinwei Chen, Le Zhang, and Bo Li. Animeagent: Is the multi-agent via image-to-video models a good disney storytelling artist? *CoRR*, abs/2602.20664, 2026. 1

[45] Jiawei Yao, Qi Qian, and Juhua Hu. Multi-modal proxy learning towards personalized visual multiple clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14066–14075. IEEE, 2024. 3

[46] Xiao Zhang and David Yunis. Deciphering 'what' and 'where' visual pathways from spectral clustering of layer-distributed neural representations. In *CVPR*, pages 4165–4175. IEEE, 2024. 1, 3

[47] Yijie Zhong, Zhengxing Sun, Shoutong Luo, Yunhan Sun, and Yi Wang. Video supervised for 3d reconstruction from single image. *Multim. Tools Appl.*, 81(11):15061–15083, 2022. 3