Seeing and Knowing in the Wild: Open-domain Visual Entity Recognition with Large-scale Knowledge Graphs via Contrastive Learning

Hongkuan Zhou[1 2], **Lavdim Halilaj**[1], **Sebastian Monka**[1], **Stefan Schmid**[1], **Yuqicheng Zhu**[1 2], **Jingcheng Wu**[2], **Nadeem Nazer**[1 4], **Steffen Staab**[2 3]

[1]Corporate Research, Robert Bosch GmbH, Renningen, Germany
[2]University of Stuttgart, Stuttgart, Germany
[3]University of Southampton, Southampton, UK
[4]Otto-von-Guericke-University Magdeburg, Magdeburg, Germany
{hongkuan.zhou, lavdim.halilaj, sebastian.monka, stefan.schmid5, yuqicheng.zhu, fixed-term.nadeem.nazer}@de.bosch.com
{jingcheng.wu, steffen.staab}@ki.uni-stuttgart.de

Abstract

Open-domain visual entity recognition aims to identify and link entities depicted in images to a vast and evolving set of real-world concepts, such as those found in Wikidata. Unlike conventional classification tasks with fixed label sets, it operates under open-set conditions, where most target entities are unseen during training and exhibit long-tail distributions. This makes the task inherently challenging due to limited supervision, high visual ambiguity, and the need for semantic disambiguation. We propose a **Know**ledge-guided **Co**ntrastive **L**earning (KnowCoL) framework that combines both images and text descriptions into a shared semantic space grounded by structured information from Wikidata. By abstracting visual and textual inputs to a conceptual level, the model leverages entity descriptions, type hierarchies, and relational context to support zero-shot entity recognition. We evaluate our approach on the OVEN benchmark, a large-scale open-domain visual recognition dataset with Wikidata IDs as the label space. Our experiments show that using visual, textual, and structured knowledge greatly improves accuracy, especially for rare and unseen entities. Our smallest model improves the accuracy on unseen entities by 10.5% compared to the state-of-the-art, despite being 35× smaller.

Code — https://github.com/boschresearch/KnowCoL
Extended version — https://arxiv.org/abs/2510.13675

Introduction

The ability to recognize and identify visual entities in the open world is a critical milestone for scalable computer vision systems. In contrast to traditional classification tasks that depend on a fixed set of categories, open-world visual entity recognition seeks to recognize and link images to a vast and evolving universe of real-world entities, such as specific landmarks, artworks, biological species, or public figures. The recently introduced Open-domain Visual Entity recognitioN (OVEN) benchmark (Hu et al. 2023) challenges models to link images and accompanying text query (specifying the intent of the image) to the correct Wikidata entities, [1] identified by its unique QID. A few examples of the OVEN-based tasks are shown in Figure 1a.

Previous approaches addressing open-world visual entity recognition typically adopt a dual-encoder paradigm, which aligns visual representations with textual descriptions of corresponding entities. CLIP2CLIP and CLIPFusion (Hu et al. 2023) include the information of the lead image(s) of the entities into the training pipeline for a better alignment of different modalities. Recently, researchers (Hu et al. 2023; Caron et al. 2024; Xiao et al. 2024) have developed *two-step* approaches which leverage generative language models, such as PaLI (Chen et al. 2023), GIT (Wang et al. 2022), and Vicuna (Zheng et al. 2023), to generate textual labels of images and then utilize search algorithms (e.g. BM25 (Robertson, Walker, and Hancock-Beaulieu 1995)) to identify entities whose names closely match these predicted labels.

Despite the progress, this task remains fundamentally challenging. First, the label space is extremely large and long-tailed, encompassing millions of entities, many of which are rare or entirely unseen during training. Second, existing visual classifiers treat entities as isolated labels, ignoring the wealth of semantic relationships and factual knowledge that exist between entities. These limitations hinder generalization, especially in zero-shot settings, where the model must recognize entities it has never encountered during training. Third, a key drawback of the two-step generative approaches is the *information loss* incurred when converting rich visual content into simplified textual labels, leading to semantic ambiguity (Sevgili et al. 2022; Bouarroudj, Boufaïda, and Bellatreche 2022) when conducting BM25 search algorithms based on the predicted textual labels to find the final entity. Entities with similar or identical textual labels may represent fundamentally different concepts. For instance, "Mercury" can denote either the innermost planet of our Solar System or the liquid metal with chemical symbol Hg (atomic number 80), showing how simple text matching cannot distinguish these distinct meanings.

[1]We consistently use the term **entity** to refer to uniquely identifiable real-world concepts in Wikidata (e.g. Albert Einstein (Q937) and Golden Retriever (Q38686)), which differs from the typical usage of class, denoting categories within a predefined set.

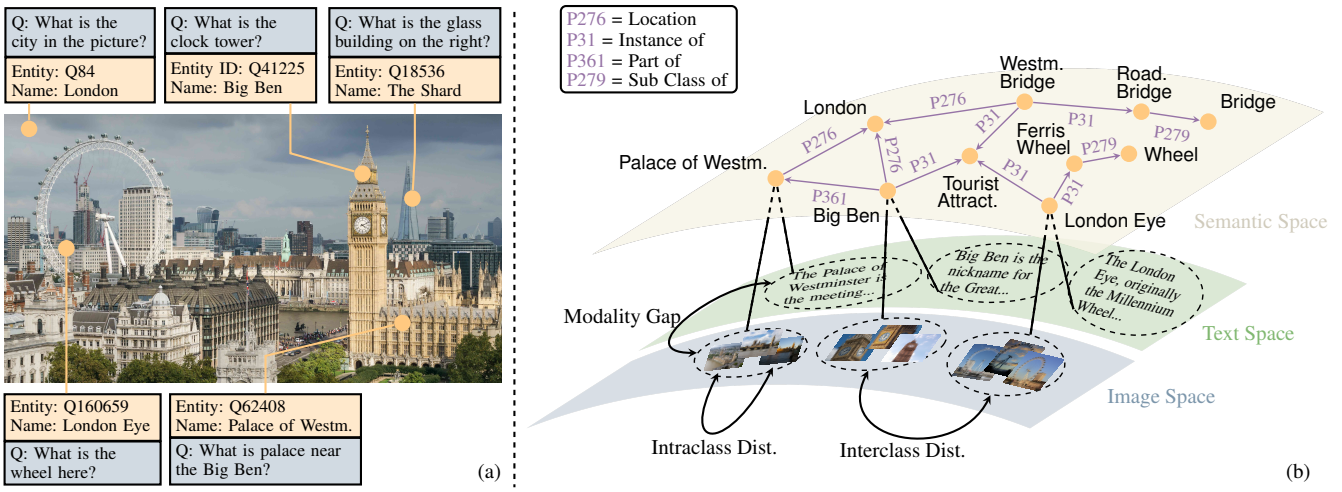


Figure 1: (a) The open-domain visual entity recognition task aims to link an image and a text query to a corresponding Wikidata entity. The text query expresses the intent to the image. (b) Both visual and textual inputs can be mapped to a shared semantic space enriched with structured knowledge, such as entity relations (e.g., instance-of, part-of, subclass-of) from Wikidata. Leveraging this additional knowledge source enables contextual reasoning and effective disambiguation of entities.

We propose that open-domain visual entity recognition should move beyond superficial recognition toward semantic-level understanding. This involves abstracting both images and textual description into a shared conceptual space with rich structural knowledge among entities (cf. Figure 1b). In this view, recognition is not simply matching visual features to entity names, but of aligning image content with structured, contextualized knowledge about the entities depicted. Available knowledge graphs (KGs), such as Wikidata, which offers a rich and structured representation of hierarchical and association relations for millions of real-world entities, can be leveraged to achieve this.

We present **Know**ledge-Guided **Co**ntrastive **L**earning (KnowCoL), an approach for open-domain visual entity recognition. Input images and candidate entities' descriptions and/or lead images are projected into a shared semantic embedding space, structured with the prior knowledge from the KG. This enables zero-shot recognition by allowing the model to generalize from seen to unseen entities based on semantic similarity, while also supporting entity disambiguation via knowledge-informed representations. KnowCoL is based on a dual-encoder paradigm, eliminating the information loss in the process of conversion in the two-step approaches. We evaluate our approach on the OVEN benchmark. Our contributions are as follows:

1. We present a dual-encoder approach for the open-domain visual entity recognition task that leverages external knowledge from both Wikidata and Wikipedia to enable a better representation of images and text descriptions.
2. We investigate the impact of different forms of external knowledge (including lead images) and different relations between entities on the recognition performance.
3. Our approach demonstrates strong zero-shot generalization on the OVEN benchmark, showing that incorporating external knowledge significantly improves entity recognition performance of unseen entities by 10.5%, even with models that are 35× smaller.

Preliminary

Contrastive Loss

Contrastive loss is a widely used objective in representation learning, particularly in contexts such as metric learning, image-text alignment, and self-supervised learning. Given two sets of embeddings $A = \{a_i\}_{i=1}^N$ and $B = \{b_i\}_{i=1}^N$, where (a_i, b_i) is a positive (matching) pair and $\forall (a_i, b_j), j \neq i$, is a negative pair, we define the contrastive loss as:

$$\ell(A, B) = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp\left(\text{sim}(a_i, b_i)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}(a_i, b_j)/\tau\right)}, \quad (1)$$

where $\text{sim}(\cdot, \cdot)$ denotes the similarity function, τ is a temperature parameter controlling the softness, and N is the set size. Common similarity functions include cosine similarity and Euclidean distance-based similarity. Specifically, the symmetric contrastive loss can be defined as

$$\ell_{\text{sym}}(A, B) = \frac{1}{2} \left(\ell\left(A, B\right) + \ell\left(B, A\right)\right), \quad (2)$$

which encourages mutual alignment, and it is widely used in various models like CLIP.

Wikidata Knowledge Graph and Wikipedia Knowledge Base

Wikidata is a collaboratively-maintained, multilingual knowledge graph that can be formalised as the directed graph $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, where $\mathcal{E}$ is the set of entities (called items and uniquely referenced by "Q-identifiers" (QIDs), e.g., Q42 for Douglas Adams), $\mathcal{R}$ is the set of relation types (called properties and referenced by "P-identifiers" (PIDs),

e.g., P31 for instance of), and $\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ is the set of knowledge triples (e_1, r, e_2) stating that subject entity e_1 is linked to object entity e_2 via property r. As of 2025, Wikidata KG contains more than 100 million entities, 12,000 properties, and 10 billion triples and is growing daily through human and bot contributions.

Wikipedia is a collaboratively edited, multilingual encyclopedia whose every article is linked to a unique Wikidata QID. This tight coupling lets each page serve as a multimodal hub: rich encyclopedic text is paired with lead images that provide both descriptive and visual context for the corresponding entity in the Wikidata KG.

Related Work

Zero-Shot Classification

Open-world recognition is closely related to zero-shot learning (ZSL), where models must recognize classes with no training images by relying on auxiliary information.

Early approaches (Farhadi et al. 2009; Lampert, Nickisch, and Harmeling 2009) introduced human-defined attributes as intermediate semantic representations for recognition of unseen classes. Later methods moved toward unsupervised semantic embeddings derived from text corpora (Socher et al. 2013; Frome et al. 2013; Norouzi et al. 2014). These dual-encoder or projection-based approaches have proven that aligning images with distributed text representations enables zero-shot recognition at scale. Vision-language pretraining has greatly advanced this paradigm. A well-known example is CLIP (Radford et al. 2021), which is trained on millions of image-text pairs to learn a joint embedding space for images and natural language. CLIP encodes images and labels into one space and ranks them by similarity to perform zero-shot classification. It has wide applications in visual inspection (Jeong et al. 2023; Sadikaj et al. 2025) and robotics (Zhou et al. 2023, 2024a; Yao et al. 2025).

Another line of zero-shot research uses large-scale generative models to bridge the gap to unseen classes. Instead of directly mapping images to text semantics, these generative approaches output the entity names based on given images and then link the entity names to the categories (two-step approach). For instance, Hu et al. (2023) use PaLI and BLIP-v2 (Li et al. 2023a) on the OVEN benchmark to make the model predict the entity name and then uses the BM25 search algorithm to find the closest label in the label space of Wikidata ids. Instead of generating entity names, GERALD (Caron et al. 2024) generates the semantic and discriminative "code" to identify entities. Auto-VER (Xiao et al. 2024) combines contrastive learning with generative models to enhance the ability to distinguish similar entities within a vast label space. Unlike dual-encoder approaches, two-step generative methods output text conditioned on images and then match them to labels in the label space. This introduces semantic ambiguity, as generated outputs may be correct lexically but vague or incomplete semantically.

Knowledge Graphs for Visual Entity Recognition

Many studies have explored incorporating knowledge into vision-and-language tasks (Monka, Halilaj, and Rettinger 2022), including visual question answering (Chang et al. 2022; Chen et al. 2021) and entity-aware image captioning (Biten et al. 2019). For visual entity recognition task, Wang, Ye, and Gupta (2018) and Kampffmeyer et al. (2019) use semantic embeddings and categorical relationships from KGs through graph convolution networks for zero-shot prediction. Li et al. (2023b) augments few-shot image recognition by introducing auxiliary semantic prior knowledge and propagating knowledge among categories via a semantic-visual mapping. KG-NN (Monka et al. 2021) combines prior knowledge encoded in KGs with visual representations to enhance generalization under distribution shifts, and KGV (Zhou et al. 2024b) demonstrates that the integration of richer multi-modal priors can further improve performance. Instead of relying on small domain-specific knowledge graphs, our approach is the first to leverage a large-scale knowledge graph for open-domain visual entity recognition, demonstrating strong scalability potential.

Methodology

Our goal is to integrate the rich, structured knowledge from Wikidata into the training pipeline. To achieve this, we align knowledge graph embeddings (KGEs), image embeddings, and text embeddings within a shared latent space, enabling the structured knowledge captured by the KGEs to be effectively injected into the latent space's representation. A framework overview can be seen in Figure 2.

Problem Definition

Let $\mathcal{E}$ be the set of Wikidata entities identified by QIDs. Define the input space as $\mathcal{X} = \mathcal{I} \times \mathcal{L}$, where $\mathcal{I}$ denotes the domain of RGB images and $\mathcal{L}$ is the linguistic (text) space. The Open-domain Visual Entity Recognition (OVEN) task seeks a function $f : \mathcal{X} \to \mathcal{E}$ which maps each image-text pair (x^p, x^t) to the unique entity $e \in \mathcal{E}$. Every entity e is associated with a Wikipedia-derived record $(t_e, I_e) \in \mathcal{K} \subseteq \mathcal{L} \times 2^{\mathcal{I}}$, where $t_e \in \mathcal{L}$ is its encyclopaedic description and $I_e \subset \mathcal{I}$ is the finite set of lead images illustrating the corresponding entity. Given a labeled training dataset $\{(x_i, e_i)\}_{i=1}^{N} \subset \mathcal{X} \times \mathcal{E}$, the goal is to learn a function f that minimizes the number of misclassifications.

Embeddings

Knowledge Graph Embeddings To model the knowledge existing in the knowledge graph $\mathcal{G}$, we adopt embedding methods to inject the knowledge into the latent space. We define knowledge graph embeddings as mappings:

$$\phi : \mathcal{E} \to \mathcal{Z}, \quad \psi : \mathcal{R} \to \mathbb{R}^{d_r},$$

where $\mathcal{Z} \subseteq \mathbb{R}^{d_e}$ is the latent space with dimension d_e. $\phi(e), \psi(r)$ are the node and relation embeddings, respectively. Common knowledge graph embedding methods define score function $f_s(e_1, r, e_2)$ to measure the plausibility of a triple $(e_1, r, e_2) \in \mathcal{T}$.

Image Embeddings Let $\mathcal{I}$ be the space of input images. We define an image mapping function as:

$$f_\theta : \mathcal{I} \to \mathcal{Z},$$

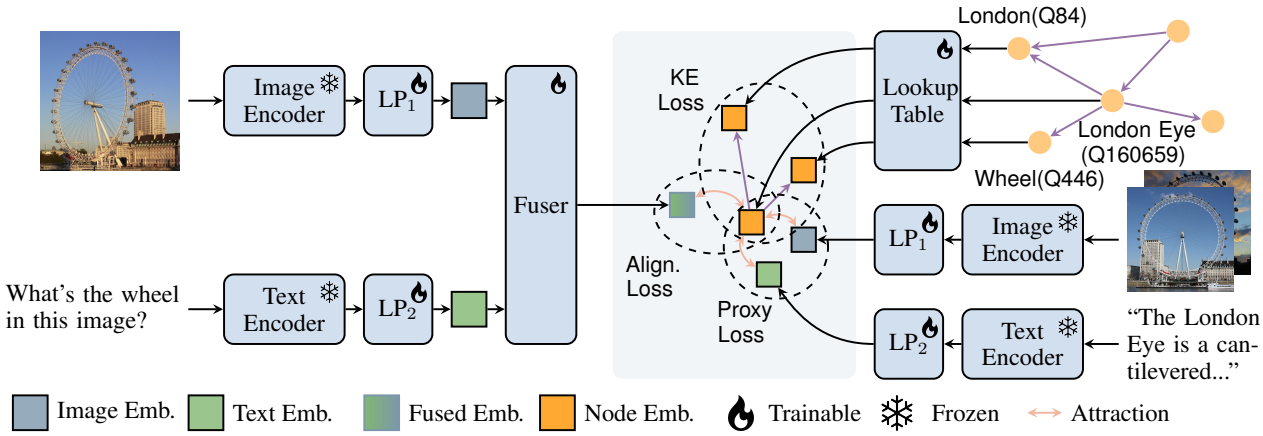


Figure 2: **Framework Overview**. Given an input image and text query (left), CLIP encoders extract image and text features. LP_1 and LP_2 present linear project layers for fine-tuning. A fuser module combines these to represent the intended entity. On the right, prior knowledge of the entity (Q160659 – London Eye)—including sample images, textual descriptions, and knowledge graph structure—is embedded using CLIP's image and text encoders, as well as knowledge graph embedding (KGE) method. A proxy loss function is employed to align the semantic space (KGEs), visual space (image embeddings), and textual space (text embeddings). Additionally, a knowledge embedding (KE) loss is applied to learn KGEs from the structured knowledge. Finally, an alignment loss is used to align the different input representation with the corresponding entity embedding.

where f_θ is represented by a deep neural network with parameters θ. Here, we leverage the pretrained frozen CLIP image encoder plus a trainable linear project layer to be f_θ.

Text Embeddings Let $\mathcal{L}$ be the linguistic (text) space of input text descriptions. We define a text mapping function:

$$f_\lambda : \mathcal{L} \to \mathcal{Z},$$

where f_λ is represented by a deep neural network with parameters λ. We leverage the pretrained *frozen* CLIP text encoder plus a *trainable* linear project layer to be f_λ.

We employ a dual-encoder approach to align the semantic representation of the input image and text query with that of the corresponding entity, which includes its lead image(s) and textual description. We structure our discussion into two parts: (1) encoding the input image and text query, and (2) encoding the multi-modal information of entities from the knowledge graph.

Input Image and Text Query Encoding To fuse the information of the input image x^p and the text query x^t, we employ an extra encoder

$$f_\gamma : (\mathcal{Z}, \mathcal{Z}) \to \mathcal{Z}$$

with parameters γ to fuse the image embedding $f_\theta(x^p)$ and text embedding $f_\lambda(x^t)$. The fused embedding can be written as

$$\mathbf{z}^{\text{input}} = f_\gamma(f_\theta(x^p), f_\lambda(x^t)) \tag{3}$$

Entity's Multi-modal Information Encoding For each entity e, t_e and I_e represent its text description and set of lead image(s), respectively. The entity text embedding can be written as

$$\mathbf{z}^{\text{entityText}} = f_\lambda(t_e), \tag{4}$$

and the image embedding can be written as

$$\mathbf{z}^{\text{entityImage}} = \begin{cases} \frac{1}{|I_e|}\sum_{a \in I_e} f_\theta(a) & \text{if } I_e \neq \emptyset \\ \mathbf{z}^{\text{entityText}} & \text{if } I_e = \emptyset \end{cases}. \tag{5}$$

If no lead image is available for a given entity, we assign $\mathbf{z}^{\text{entityText}}$ to $\mathbf{z}^{\text{entityImage}}$. This ensures a smooth fusion of visual and textual information, as discussed in the next section.

Knowledge-guided Contrastive Learning

For given input batch $\{(x_i^p, x_i^t)\}_{i=1}^{N_b}$, where x_i^p and x_i^t are the i-th input images and its corresponding text query, N_b is the batch size, the corresponding ground truth label set is denoted as $\{e_i\}_{i=1}^{N_b}$. We define three losses, namely alignment loss, proxy loss, and knowledge graph loss, for the training process. Figure 3 shows the learning visualization.

Alignment Loss The alignment loss aims to align the joint representation of input images and text queries $\{\mathbf{z}_i^{\text{input}}\}_{i=1}^{N_b}$ with their corresponding entity representations $\{\phi(e_i)\}_{i=1}^{N_b}$. Specifically, we define the alignment loss using a contrastive objective that pulls matched embeddings closer together while pushing unmatched embeddings among the batch apart. It can be defined as

$$\mathcal{L}_a = \ell_{\text{sym}}\left(\{\mathbf{z}_i^{\text{input}}\}_{i=1}^{N_b}, \{\phi(e_i)\}_{i=1}^{N_b}\right), \tag{6}$$

where $\ell_{\text{sym}}(\cdot,\cdot)$ represents the contrastive loss function define in Equation 2.

Proxy Loss The proxy loss is defined to align the node embeddings $\{\phi(e_i)\}_{i=1}^{N_b}$ with their corresponding multi-modal representations $\{\mathbf{z}_i^{\text{entity}}\}_{i=1}^{N_b}$. This objective ensures that the

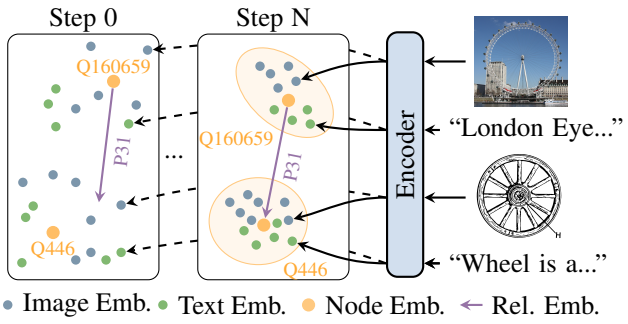


Figure 3: Node embeddings act as proxies, aligned with their corresponding image and text embeddings. Knowledge (from the KG) is captured through relational embeddings, connecting node embeddings in the latent space. Q160659, Q446, P31 represent *London Eye*, *Wheel*, and *Instance of*.

node embeddings capture semantic information from both image and text modalities. It can be defined as

$$\mathcal{L}_p = \frac{1}{2}\ell_{\text{sym}}\left(\{\phi(e_i)\}_{i=1}^{N_b}, \{\mathbf{z}_i^{\text{entityText}}\}_{i=1}^{N_b}\right) + \frac{1}{2}\ell_{\text{sym}}\left(\{\phi(e_i)\}_{i=1}^{N_b}, \{\mathbf{z}_i^{\text{entityImage}}\}_{i=1}^{N_b}\right) \tag{7}$$

We refer to this as the proxy loss because the node embeddings serve as proxies for capturing the multi-modal semantic information of the entities. Figure 3 shows that the proxy loss gathers image and text embeddings with their corresponding node embeddings.

Knowledge Graph Embedding Loss (KE Loss) The knowledge graph loss aims to capture the structural knowledge associated with entity e_i. We first extract the triplet set associated with the entity e_i by $\mathcal{T}_{e_i} = \{(h,r,t) \in \mathcal{T} \mid h = e_i \vee t = e_i\}$. For each triplet $(h,r,t) \in \mathcal{T}_{e_i}$, we define $\mathcal{T}_{(h,r,t)}$ as the set of negative samples generated by corrupting either the head or tail entity in the positive triplet:

$$\mathcal{T}_{(\text{h,r,t})} = \{(h',r,t) \mid h' \in \mathcal{E}\backslash\{h\}\} \cup \{(h,r,t') \mid t' \in \mathcal{E}\backslash\{t\}\}. \tag{8}$$

Based on that, the knowledge graph loss can be defined as

$$\mathcal{L}_{\text{KE}} = \sum_{i=1}^{N_b} \frac{1}{|\mathcal{T}_{e_i}|} \sum_{(h,r,t)\in\mathcal{T}_{e_i}} \frac{1}{|\mathcal{T}_{(h,r,t)}|} \left[-\log \frac{\exp(f_s(h,r,t)/\tau)}{\sum_{(h',r,t')\in\mathcal{T}_{(h,r,t)}} \exp(f_s(h',r,t')/\tau)}\right], \tag{9}$$

where τ is a temperature and $f_s(\cdot)$ is the score function

$$f_s(h,r,t) = \frac{(\phi(h)+\psi(r))^\top \phi(t)}{\|\phi(h)+\psi(r)\|_2 \, \|\phi(t)\|_2}, \tag{10}$$

which is the cosine similarity between $\phi(h)+\psi(r)$ and $\phi(t)$. Here, we follow a translation-based KGE approach with a cosine similarity distance metric. The final loss for optimization can be defined as

$$\mathcal{L} = \mathcal{L}_a + \beta_1 \mathcal{L}_p + \beta_2 \mathcal{L}_{\text{KE}}, \tag{11}$$

where β_1 and β_2 are two hyperparameters for balancing.

Inference

Given an input image x^p and a text query x^t, we first obtain the input embedding $\mathbf{z}^{\text{input}}$ as defined in Equation 3. The corresponding entity answer e^* is then identified through the following inference step:

$$e^* = \arg\max_{e_j \in \mathcal{E}} \text{sim}\left(\mathbf{z}^{\text{input}}, \frac{1}{2}\left(\mathbf{z}_j^{\text{entityText}} + \mathbf{z}_j^{\text{entityImage}}\right)\right), \tag{12}$$

where $\mathbf{z}_j^{\text{entityText}}$ and $\mathbf{z}_j^{\text{entityImage}}$ denote the textual and visual embeddings of entity e_j in Equations 4 and 5.

Experiments

In our experiments, we investigate whether incorporating knowledge from large-scale KGs and Knowledge Base (KB) can enhance open-domain visual entity recognition. We analyze the impact of different types of prior knowledge, fusion methods, and KGE techniques. In addition, ablation studies on various hyperparameter values are also provided.

Experiment Settings

OVEN Dataset The OVEN dataset contains 6,063,945 training samples, which combines 14 image recognition datasets. Models are evaluated on the test splits, and performance is reported using the harmonic mean (HM) of top-1 accuracy for two types of entities: seen entities, which appear in the OVEN training set, and unseen entities, which are not present during training. Specifically, the test split includes 15,888 entities (8,355 seen and 7,533 unseen). Our evaluation is conducted on the OVEN benchmark, comprising 14 diverse image recognition datasets, thus indicating the generalizability of our approach.

Training details We extract a subgraph from Wikidata containing 32,122 entities and 501 relation types. For training, we use the AdamW optimizer with a learning rate of 0.001, batch size of 4096, and weight decay of 0.0001. Our approach includes three model variants based on different OpenCLIP image encoders: ViT-L/14, ViT-H/14, and ViT-bigG/14. The temperature τ for contrastive learning is fixed at 0.07 during fine-tuning. The hyperparameters β_1 and β_2 are both set to 1. We fuse information from the input image and text query (Equation (3)), using a simple *addition* operation. More details are in the extended version (Appendix).

Baselines

Dual Encoders Dual-encoder approaches consist of two encoders, each encoding a dedicated modality (e.g., images and text) into the *same* latent space. The visual recognition is conducted by searching the nearest neighbors of prototypes in the latent space for a given input image. Representative models include CLIP, CLIPfusion, and CLIP2CLIP.

Auto-regressive Captioning Another approach uses auto-regressive models to generate captions for the given images. Representative vision-language models (VLMs) such as PaLI, GIT-Large, GER-ALD, and Auto-VER follow this paradigm. These models generate a descriptive caption or name for the visual entity and then match it to the closest label in the predefined label space for final prediction.

Paradigm	Model	Venues	Parameters (B)	Pre-train Dataset	Seen	Unseen	HM
Dual-Encoder	CLIP (Radford et al. 2021)	ICML2021	0.4	OpenAI	5.6	4.9	5.2
	CLIPFusion (Hu et al. 2023)	ICCV2023	0.9	OpenAI	33.6	4.8	8.4
	CLIP2CLIP (Hu et al. 2023)	ICCV2023	0.9	OpenAI	12.6	10.5	11.4
Two-Step Generative	BLIP-v2 (Li et al. 2023a)	ICML2023	12.2	-	8.6	3.4	4.9
	PaLI-3B (Hu et al. 2023)	ICCV2023	3.0	WebLI	19.1	6.0	9.1
	PaLI-17B (Hu et al. 2023)	ICCV2023	17.0	WebLI	28.3	11.2	16.0
	GIT-Large (Caron et al. 2024)	CVPR2024	0.4	WebLI	17.6	4.3	7.0
	GER-ALD (Caron et al. 2024)	CVPR2024	0.4	LAION	29.1	16.3	20.9
	Auto-VER-7B (Xiao et al. 2024)	ECCV2024	7.0	-	62.8	16.0	25.5
	Auto-VER-14B (Xiao et al. 2024)	ECCV2024	14.0	-	**65.0**	18.6	28.9
Knowledge-Guided Dual-Encoder	KnowCoL-bigG(ours)	-	2.0	LAION	41.8	**36.1**	**38.8**

Table 1: Comparison with current state-of-the-art approaches on OVEN entity test split. We evaluate the harmonic mean (HM) of the seen and unseen splits (top1 accuracy) after fine tuning on OVEN training set. It serves as our main metric. The numbers of baselines are taken from papers (Caron et al. 2024; Hu et al. 2023; Xiao et al. 2024). The total parameters and pre-training datasets of the models are given for comparison.

Comparison with the state of the art We compare our proposed KnowCoL model with both dual-encoder and auto-regressive captioning baselines. As shown in Table 1, our model achieves significantly stronger performance on both seen and unseen entities. KnowCoL-bigG achieves the highest harmonic mean (HM) of 38.8%, outperforming large-scale generative models such as Auto-VER-14B and PaLI-17B despite using 7× fewer parameters. We identify that the Auto-VER achieves impressive results for seen entities, but the performance significantly drops for unseen entities, indicating overfitting. Compared to theirs, KnowCoL achieves a balanced result on both seen and unseen entities, representing a strong generalization ability by including structure, textual, and visual prior knowledge.

Analysis and Ablation Studies

Here, we analyse the impact of model size, prior knowledge types, fusion strategy of the text query and input image, various KGE methods, and hyperparameters - β_1, β_2, latent space dimension d_e, and temperature τ. The detailed setting can be found in the Appendix section, experiment setting.

Model	Backbone	Para.(B)	Seen	Unseen	HM
CLIP	ViT-L-14	0.4	5.6	4.9	5.2
CLIPFusion	ViT-L-14	0.9	33.6	4.8	8.4
CLIP2CLIP	ViT-L-14	0.9	12.6	10.5	11.4
KnowCoL-L	ViT-L-14	0.4	**34.3**	**29.1**	**31.5**
KnowCoL-H	ViT-H-14	0.9	38.5	33.4	35.8
KnowCoL-bigG	ViT-bigG-14	2.0	**41.8**	**36.1**	**38.8**

Table 2: Comparison of model sizes. ViT-H-14 and ViT-bigG-14 are used for our KnowCoL-H and KnowCoL-bigG models. ViT-L-14 is used for CLIP, CLIPFusion, CLIP2CLIP, and KnowCoL-L. Note that CLIPFusion and CLIP2CLIP utilize two ViT-L-14 backbones.

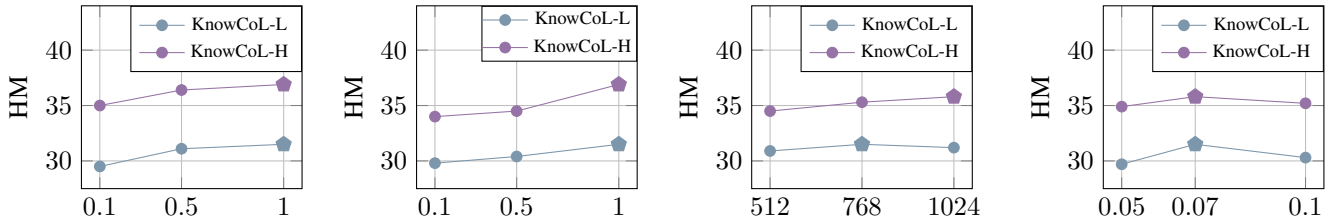
Impact of Model Size As Table 2 demonstrates, we evaluate our model with different CLIP backbones. The backbones ViT-L-14, ViT-H-14, and ViT-bigG-14 are used for KnowCoL-L, KnowCoL-H, and KnowCoL-bigG, respectively. The performance across both seen and unseen entities consistently improves as model size increases.

We highlight that KnowCoL-L, CLIP, CLIPFusion, and CLIP2CLIP use the same backbone ViT-L-14, but KnowCoL-L significantly outperforms all. While CLIPFusion and CLIP2CLIP achieve harmonic means of 8.4% and 11.4%, respectively, KnowCoL-L reaches 31.5%, showing about 2.8 times improvement. This demonstrates that external knowledge significantly improves zero-shot generalization without increasing model size.

Model	Seen	Unseen	HM
KnowCoL-L (ViT-L-14)	**34.3**	**29.1**	**31.5**
w/o Lead images	32.5	27.9	30.0
w/o KG Hierarchical Knowledge	32.1	25.3	28.3

Table 3: Impact of different types of knowledge. We investigate the impact of structural knowledge, namely the sample images in the KB and the hierarchical relations in the KG

Comparison of Knowledge Types Table 3 demonstrates the importance of different knowledge types in the KnowCoL-L model. The complete KnowCoL-L model achieves the best score of 31.5% HM. Removing lead images (w/o Lead images) reduces performance to 32.5% (seen), 27.9% (unseen), and 30.0% (HM), highlighting the role of visual context. Omitting hierarchical KG knowledge of 'Instance of', 'subclass of', and 'parent taxon' further lowers scores to 32.1% (seen), 25.3% (unseen), and 28.3% (HM), significantly impairing unseen entity recognition.



(a) Hyperparameter β_1 (b) Hyperparameter β_2 (c) Latent Space Dim d_e (d) Temperature τ

Figure 4: Ablation Studies for hyperparameter β_1, β_2, latent space dimension d_e, and temperature τ. HM indicates the harmonic mean of accuracies of seen and unseen entities. ⬟- represents the default setting of KnowCoL approach.

Fusion Method	Layer Number	Seen	Unseen	HM
Addition	-	34.3	**29.1**	**31.5**
Concat. + MLP	1	37.5	26.8	31.2
Concat. + MLP	2	45.4	21.0	28.7
TE	2	54.4	16.6	25.5
TE + IP	2	**56.7**	17.9	27.2

Table 4: Comparison of fusion functions for combining image and text embeddings of KnowCoL-L. Our default fusion method is addition. 'TE' refers to the Transformer Encoder used in (Hu et al. 2023). 'TE + IP' is inspired by LLaVA (Liu et al. 2023) which incorporates local ViT patch tokens into the fusion process.

Comparison of Fusion Functions As Table 4 shows, the default Addition method provides balanced performance, achieving the best harmonic mean 31.5%. By using the multi-layer perceptron (MLP) method, the performance for seen entities improves up to 37.5%, while for unseen entities decreases to 26.8%. Using transformer encoder (TE) methods with or without local image patch embeddings significantly boost the accuracy on seen entities (54.4% and 56.7%), but degrade unseen entity recognition (16.6% and 17.9%). The trend shows that more complex fusion methods, containing more layers or a more complex structure, lead to overfitting on seen entities and sacrificing zero-shot generalization to unseen entities. Additionally, by introducing local image patch tokens (TE + IP), performance on both seen (56.7%) and unseen entities (17.9%) improves compared to the TE-only method. This improvement likely occurs because local image patch tokens provide fine-grained visual details and enhance recognition ability.

Comparison of KGE Methods Table 5 compares various knowledge graph embedding methods and distance measures. Our default choice, TransE (Bordes et al. 2013) with cosine similarity, achieves the best overall performance. This setup aligns well with CLIP, which also uses cosine similarity in its contrastive learning. When replacing the cosine similarity with Euclidean distance, the HM drops from 31.5% to 31.0%. TransH (Wang et al. 2014) introduces more representational flexibility by allowing entity embeddings to vary across different relations. This additional modeling capacity slightly improved the seen performance to 34.4% but comes at the cost of lower unseen accuracy 28.6%. The DistMult method (Yang et al. 2015) shows the lowest performance, suggesting it is less effective for integrating structural knowledge from the KG into the visual-textual space.

KGE Methods	Distance Measure	Seen	Unseen	HM
TransE	Cosine Similarity	34.3	**29.1**	**31.5**
TransE	Euclidean Dis.	33.7	28.5	31.0
TransH	Cosine Similarity	**34.4**	28.6	31.2
DistMult	-	33.8	28.5	30.9

Table 5: Comparison of KGE methods for incorporating structural knowledge. Our default approach uses TransE combined with a cosine similarity metric. We compare this against alternative configurations, including TransE with Euclidean distance, TransH, and DistMult.

Hyperparameters As Figure 4 shows, we conduct experiments varying the hyperparameters β_1, β_2, latent space dimension d_e, and contrastive learning temperature τ. Figure 4a shows that increasing the weight of the proxy loss β_1, aligning node embeddings with multi-modal prior knowledge, enhances model performance. Figure 4b demonstrates that increasing β_2, the weight of KE loss incorporating structured prior knowledge from Wikidata, also improves performance. Additionally, we identify that the larger model (KnowCoL-H) with more prior knowledge can benefit more from the inclusion of structured prior knowledge. Figure 4c illustrates that larger models require higher-dimensional latent spaces to represent richer features. The temperature 0.07 performs best in our task, as shown in Figure 4d.

Conclusion and Future Work

We propose KnowCoL, a novel approach to improve open-domain visual entity recognition by incorporating multi-modal prior knowledge, namely structured knowledge, textual, and visual priors. KnowCoL significantly enhances visual entity recognition performance, especially for zero-shot recognition of unseen entities. Future work includes exploring non-Euclidean spaces like hyperbolic and hyperspherical spaces to better capture the hierarchical relationships and semantic structures inherent in knowledge graphs.

Acknowledgments

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Hongkuan Zhou, Yuqicheng Zhu, and Jingcheng Wu. This work was partially funded by the European Union's Horizon RIA research and innovation programme under grant agreement No. 101092908 (SMARTEDGE). Jingcheng Wu has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - SFB 1574 - Project number 471687386. The authors also gratefully acknowledge the computing time provided on the high-performance computer HoreKa by the National High-Performance Computing Center at KIT (NHR@KIT). This center is jointly supported by the Federal Ministry of Education and Research and the Ministry of Science, Research, and the Arts of Baden-Württemberg, as part of the National High-Performance Computing (NHR) joint funding program (https://www.nhr-verein.de/en/our-partners). HoreKa is partly funded by the German Research Foundation (DFG).

References

Biten, A. F.; Gómez, L.; Rusiñol, M.; and Karatzas, D. 2019. Good News, Everyone! Context Driven Entity-Aware Captioning for News Images. In *CVPR*, 12466–12475. Computer Vision Foundation / IEEE.

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

Bouarroudj, W.; Boufaïda, Z.; and Bellatreche, L. 2022. Named entity disambiguation in short texts over knowledge graphs. *Knowl. Inf. Syst.*, 64(2): 325–351.

Caron, M.; Iscen, A.; Fathi, A.; and Schmid, C. 2024. A Generative Approach for Wikipedia-Scale Visual Entity Recognition. In *CVPR*, 17313–17322. IEEE.

Chang, Y.; Cao, G.; Narang, M.; Gao, J.; Suzuki, H.; and Bisk, Y. 2022. WebQA: Multihop and Multimodal QA. In *CVPR*, 16474–16483. IEEE.

Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A. J.; Padlewski, P.; Salz, D.; Goodman, S.; Grycner, A.; Mustafa, B.; Beyer, L.; Kolesnikov, A.; Puigcerver, J.; Ding, N.; Rong, K.; Akbari, H.; Mishra, G.; Xue, L.; Thapliyal, A. V.; Bradbury, J.; and Kuo, W. 2023. PaLI: A Jointly-Scaled Multilingual Language-Image Model. In *ICLR*. OpenReview.net.

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

Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. A. 2009. Describing objects by their attributes. In *CVPR*, 1778–1785. IEEE Computer Society.

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

Hu, H.; Luan, Y.; Chen, Y.; Khandelwal, U.; Joshi, M.; Lee, K.; Toutanova, K.; and Chang, M. 2023. Open-domain Visual Entity Recognition: Towards Recognizing Millions of Wikipedia Entities. In *ICCV*, 12031–12041. IEEE.

Jeong, J.; Zou, Y.; Kim, T.; Zhang, D.; Ravichandran, A.; and Dabeer, O. 2023. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19606–19616.

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

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 951–958. IEEE Computer Society.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. C. H. 2023a. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, 19730–19742. PMLR.

Li, Z.; Tang, H.; Peng, Z.; Qi, G.-J.; and Tang, J. 2023b. Knowledge-Guided Semantic Transfer Network for Few-Shot Image Recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 1–15.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. arXiv:2304.08485.

Monka, S.; Halilaj, L.; and Rettinger, A. 2022. A survey on visual transfer learning using knowledge graphs. *Semantic Web*, 13(3): 477–510.

Monka, S.; Halilaj, L.; Schmid, S.; and Rettinger, A. 2021. Learning Visual Models Using a Knowledge Graph as a Trainer. In Hotho, A.; Blomqvist, E.; Dietze, S.; Fokoue, A.; Ding, Y.; Barnaghi, P. M.; Haller, A.; Dragoni, M.; and Alani, H., eds., *The Semantic Web - ISWC 2021 - 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24-28, 2021, Proceedings*, volume 12922 of *Lecture Notes in Computer Science*, 357–373. Springer.

Norouzi, M.; Mikolov, T.; Bengio, S.; Singer, Y.; Shlens, J.; Frome, A.; Corrado, G.; and Dean, J. 2014. Zero-Shot Learning by Convex Combination of Semantic Embeddings. In *ICLR*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Robertson, S. E.; Walker, S.; and Hancock-Beaulieu, M. 1995. Large Test Collection Experiments on an Operational, Interactive System: Okapi at TREC. *Inf. Process. Manag.*, 31(3): 345–360.

Sadikaj, Y.; Zhou, H.; Halilaj, L.; Schmid, S.; Staab, S.; and Plant, C. 2025. MultiADS: Defect-aware Supervision for Multi-type Anomaly Detection and Segmentation in Zero-Shot Learning. *CoRR*, abs/2504.06740.

Sevgili, Ö.; Shelmanov, A.; Arkhipov, M. Y.; Panchenko, A.; and Biemann, C. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3): 527–570.

Socher, R.; Ganjoo, M.; Manning, C. D.; and Ng, A. Y. 2013. Zero-Shot Learning Through Cross-Modal Transfer. In *NIPS*, 935–943.

Wang, J.; Yang, Z.; Hu, X.; Li, L.; Lin, K.; Gan, Z.; Liu, Z.; Liu, C.; and Wang, L. 2022. GIT: A Generative Image-to-text Transformer for Vision and Language. *Trans. Mach. Learn. Res.*, 2022.

Wang, X.; Ye, Y.; and Gupta, A. 2018. Zero-Shot Recognition via Semantic Embeddings and Knowledge Graphs. In *CVPR*, 6857–6866. Computer Vision Foundation / IEEE Computer Society.

Wang, Z.; Zhang, J.; Feng, J.; and Chen, Z. 2014. Knowledge Graph Embedding by Translating on Hyperplanes. In *AAAI*, 1112–1119. AAAI Press.

Xiao, Z.; Gong, M.; Cascante-Bonilla, P.; Zhang, X.; Wu, J.; and Ordonez, V. 2024. Grounding Language Models for Visual Entity Recognition. In *ECCV (11)*, volume 15069 of *Lecture Notes in Computer Science*, 393–411. Springer.

Yang, B.; Yih, W.; He, X.; Gao, J.; and Deng, L. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *ICLR (Poster)*.

Yao, X.; Blei, T.; Meng, Y.; Zhang, Y.; Zhou, H.; Bing, Z.; Huang, K.; Sun, F.; and Knoll, A. 2025. Long-Horizon Language-Conditioned Imitation Learning for Robotic Manipulation. *IEEE/ASME Transactions on Mechatronics*, 1–12.

Zheng, L.; Chiang, W.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *NeurIPS*.

Zhou, H.; Bing, Z.; Yao, X.; Su, X.; Yang, C.; Huang, K.; and Knoll, A. 2024a. Language-Conditioned Imitation Learning With Base Skill Priors Under Unstructured Data. *IEEE Robotics and Automation Letters*, 9(11): 9805–9812.

Zhou, H.; Halilaj, L.; Monka, S.; Schmid, S.; Zhu, Y.; Xiong, B.; and Staab, S. 2024b. Visual Representation Learning Guided By Multi-modal Prior Knowledge. *CoRR*, abs/2410.15981.

Zhou, H.; Yao, X.; Mees, O.; Meng, Y.; Xiao, T.; Bisk, Y.; Oh, J.; Johns, E.; Shridhar, M.; Shah, D.; Thomason, J.; Huang, K.; Chai, J.; Bing, Z.; and Knoll, A. 2023. Bridging Language and Action: A Survey of Language-Conditioned Robot Manipulation. *CoRR*, abs/2312.10807.