Large Language Model Informed Patent Image Retrieval

Hao-Cheng Lo[1,2], Jung-Mei Chu[1,2], Jieh Hsiang[1,2] and Chun-Chieh Cho[2]

[1]*National Taiwan University*

[2]*JCIPRNET*

Abstract

In patent prosecution, image-based retrieval systems for identifying similarities between current patent images and prior art are pivotal to ensure the novelty and non-obviousness of patent applications. Despite their growing popularity in recent years, existing attempts, while effective at recognizing images within the same patent, fail to deliver practical value due to their limited generalizability in retrieving relevant prior art. Moreover, this task inherently involves the challenges posed by the abstract visual features of patent images, the skewed distribution of image classifications, and the semantic information of image descriptions. Therefore, we propose a language-informed, distribution-aware multimodal approach to patent image feature learning, which enriches the semantic understanding of patent image by integrating Large Language Models and improves the performance of underrepresented classes with our proposed distribution-aware contrastive losses. Extensive experiments on DeepPatent2 dataset show that our proposed method achieves state-of-the-art or comparable performance in image-based patent retrieval with mAP $+53.3\%$, Recall@10 $+41.8\%$, and MRR@10 $+51.9\%$. Furthermore, through an in-depth user analysis, we explore our model in aiding patent professionals in their image retrieval efforts, highlighting the model's real-world applicability and effectiveness.

1. Introduction

Prior art search aims to identify similarities between new inventions and existing technologies, thus ensuring the inventions satisfy novelty and non-obviousness requirements during patent drafting, examination, and infringement analysis [2]. Traditionally focused on metadata and textual information [3], researchers have increasingly turned to image-based patent retrieval to overcome the limitations (e.g., the complexity of legal and technical patent language) of textual analysis [4, 5, 6, 7], given that patent images provide a clearer, more intuitive understanding of inventions (e.g., vehicle, design, and fashion), enabling faster and deeper insights compared to text alone [8].

Patent images, designed to convey technical and scientific information, exhibit distinctive features that set them apart from natural and sketch images. Firstly, they often lack the background context, color, texture, and intensity variability found in natural images, characterized instead by their abstractness and sparseness. Secondly, unlike sketch images, patent images provide detailed and high-quality visualizations from multiple viewpoints. This specificity results in commercial search engines like Google facing difficulties in accurately retrieving relevant patent images from drawing queries [5, 7], thereby rendering image-based patent retrieval a significant and ongoing challenge.

Research on image-based patent retrieval, though limited, can be broadly categorized into two: (*i*) Low-level vision-based methods, which employee basic visual features such as visual words [9], shape and contour [10, 11], relational skeletons [12], and adaptive hierarchical density histograms [13] to describe patent images for retrieval. These methods, however, falter in large-scale applications [7]. (*ii*) Learning-based methods have gained traction in recent years. For instance, one early work, using object detection and multi-task framework for patent classification, simultaneously performs image-based retrieval [4]. With the emergence of DeepPatent dataset [5] and the ECCV 2022 DIRA Workshop Image Retrieval Challenge has prompted exploration into various network architectures, loss functions, and Re-ID techniques to improve retrieval systems [6, 7].

Despite these efforts, past studies have often overlooked the real-world workflow of patent attorneys conducting prior art searches with images. In practice, patent attorneys evaluate not only the visual similarity between current images and those of prior art but also consider the images' descriptions and their associated patent classifications. This oversight leads to several critical gaps and our corresponding contributions: (*i*) Given the importance of textual content, we adopt a visual language model (VLM) [14] without following pretrain-finetune paradigm. Furthermore, recognizing the limited semantics in patent images' textual content (i.e., primarily object names and perspectives), we, inspired by past prompting engineering [15, 16], propose using large language models (LLMs) [17] to generate detailed, alias-containing, free-form descriptions. (*ii*) To incorporate patent classification and address its long-tail distribution (1), beyond the InfoNCE loss, we introduce multiple coarse-grained losses with uncertainty factors tailored for long-tail data into our
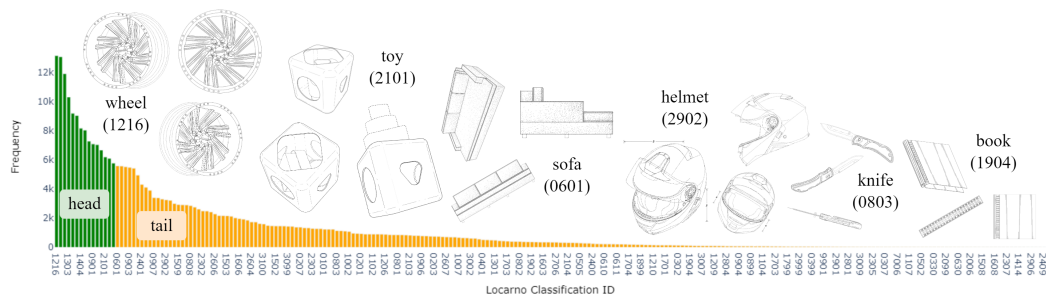
Figure 1: The long-tail distribution of patent images within DeepPatent2 [1], plotting Locarno classification IDs against their frequency. The graph differs between the top 40% of classes (head) and the bottom 60% (tail), showing patent image examples from different perspectives with their object names under various classifications.

VLM. This strategy aims to ensure that patent image representations capture class information while remaining sensitive to the distribution [18]. (*iii*) Previous works have treated image-based patent retrieval tasks as Re-ID tasks, which do not fully align with industrial needs. Typically, searches are conducted on large databases and retrieval is carried out both before and after a patent is granted, primarily in two scenarios: novelty detection or prior art search, where a current invention is compared against past inventions to identify similarities; and infringement search, which aims to identify subsequent inventions that might infringe upon the granted patent [3]. Accordingly, we train and validate our model on a larger dataset and ensure that the retrieval metrics align with these temporal concerns. (*iv*) To further understand the practical value of our image-based patent retrieval system, we conducted blind user studies [19]. The goal is to directly evaluate and compare the satisfaction, usability, and performance of our method against existing methods in a real-world setting, showing the practical significance of our approach. Hence, we present four-fold distinctive contributions to better meet the industrial demands:

- We introduce a language-informed, distribution-aware multimodal approach to patent image feature learning, which is both simple and robust. This method enhances images with corresponding semantic information, augmented via LLMs.
- We propose tailored losses specifically designed for the long-tail distribution of patent classifications. This strategy significantly boosts the robustness and accuracy of patent image representations, particularly in scenarios sensitive to class distinctions, leading our method achieve state-of-the-art results.
- Our model is validated on a large dataset with

metrics specifically tailored for novelty detection, ensuring it meets broad industrial needs.

- We employ a multi-paradigm approach, validating the system's effectiveness not only through technical retrieval metrics but also by accentuating its practical utility through user studies.

2. Related Work

Existing learning-based works on image-based patent retrieval systems can be categorized into two approaches: The first approach is intuitive, starting with the identification of objects within patent images, then training a classifier to associate these identified objects with their respective International Patent Classification (IPC) classes, and extracting vectors from the network for retrieval [4, 9, 20]. However, this method faces two limitations: (*i*) It relies on objects that the original detector has been pre-trained to recognize, resulting in the exclusion of unidentifiable patent images and thus limiting its applicability for large-scale applications. (*ii*) Although it considers IPC, IPC provides a rather coarse classification to the entire patent, which fails to accurately reflect the specific class of a certain image.

The second approach developed with the release of the large-scale DeepPatent dataset [5], where a series of studies have treated learning patent image representation as a Re-ID (i.e., Patent ID) problem [21]. Employing various CNN backbones such as EfficientNet [7, 6], ResNet [5], ViT [6], and SwinTransformer [6], these studies aim to embed patent drawings into a common feature space using contrastive loss functions (e.g., triplet loss [5], ArcFace [6, 7]), clustering identical ID images closely and separating different ID images. Although these studies have shown excellent Re-ID capabilities, they have over-
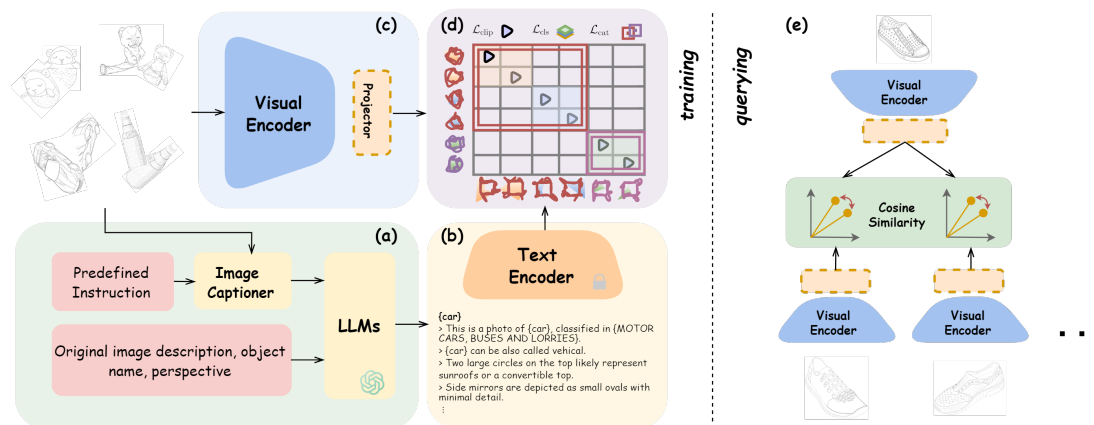
Figure 2: Model Architecture. (a) Generation of diverse, alias-containing, fine-grained descriptions for each patent image using a captioner and LLMs. (b) Text feature extraction from the enriched text via a frozen text encoder. (c) Visual feature extraction through a trainable visual encoder; a projector is employed when a mismatch between text and visual features. (d) Proposed distribution-aware contrastive losses. (e) Query patent images are converted into embeddings for retrieval based on cosine similarity.

looked several critical aspects: (*i*) Re-ID primarily focuses on retrieving images within the same patent, which does not align with the patent industry's workflow (i.e., retrieving images from different patents). (*ii*) The Re-ID approach is susceptible to overfitting, which reduces its generalizability and accuracy in retrieving similar cases across different patents [22]. (*iii*) These methods often ignore other patent-specific information, such as image descriptions and Locarno classification, which are crucial in practical patent work.

With the rise of VLM and multimodal learning, the retrieval of natural color images has seen significant improvements [23, 24, 25, 26]. Likewise, multimodal methods have become increasingly prevalent in retrieval strategies for sketch images, which are similar, if not identical, to patent images. For example, sketch-based image retrieval involves retrieving natural images using sketch representations [27, 28, 29, 30]. They mainly utilize the associations with natural images to achieve such effective results. While patent images lack the stroke information found in sketches and are challenging to associate with natural images due to their nature of novelty and multiple perspectives, the textual information and well-defined classification in patents provide a solid foundation for employing a VLM approach. Therefore, we explore the potential of applying VLM to patent image retrieval, an area currently underappreciated, leveraging the rich auxiliary information in patents [1].

3. Method

3.1. Model Overview

Our objective is to leverage the powerful capabilities of pre-trained LLMs to facilitate feature learning for patent images, thus achieving semantically rich image representations and efficient patent image retrieval. To this end, we introduce a one-stage framework, distinct from the conventional pretrain-finetune approaches employed by prior studies. As depicted in 2, our model comprises two main components: visual feature extraction and text feature extraction. In the visual component of our model, we preprocess the input patent images using data augmentation techniques tailored for patent images, such as flipping, random cropping [31], random erasing [32], and gridmask [33]. We then employ CNN-based backbones to extract visual features from these augmented images. For cases where the output feature dimensions from certain backbones are too long or too short, we utilize projectors composed of MLPs to align the dimensions of visual features with those of text features for subsequent contrastive training (2 (c)).

In the textual component, our process begins with employing an image captioner, which, given a patent image and a predefined prompt, generates a sophisticated description of the image [34, 35, 24]. This generated description is then combined with other pertinent information about the image, such as its Locarno classification, original image description, object name, and perspective. This composite input is fed into LLMs to produce diversified,

alias-caontaining, fine-grained text descriptions, thereby enriching the semantic understanding of the patent image (see 2 (a) and 3.3). Following, we employ a frozen text encoder (i.e., text encoder in CLIP), to extract textual features of the descriptions (2 (b)).

To ensure our contrastive loss is distribution-aware, we introduce three types of loss functions. Firstly, for the conventional VLM contrastive loss $\mathcal{L}_{\text{clip}}$, we treat the pairing of a fine-grained image with its corresponding description as a positive match. For the coarse-grained approach, inspired by [18, 36], we define two scenarios: class-wise, where an image and a sentence from the same *class* are considered a positive pair ($\mathcal{L}_{\text{cls}}$); and category-wise, where an image and a sentence from the same *category* (e.g., head or tail categories) are seen as a positive pair ($\mathcal{L}_{\text{cat}}$). These losses are combined to update the visual encoder during the training phase (see 2 (d) and 3.2).

In the query phase (2 (e)), each patent image is transformed into embeddings with the visual encoder and then stored in a vector database. Retrieval of other images is done by comparing these embeddings via cosine similarity, where embeddings closer in distance are ranked higher, and those farther apart are ranked lower.

3.2. Distribution-Aware Contastive Loss

As previously mentioned, we formulate our task as a VLM contrastive learning paradigm. The traditional instance-based training objective for a single image can be described as follows (i.e., InfoNCE loss, 1):

$$\mathcal{L}_{\text{clip}} = -\log\left(\frac{\exp(\mathbf{t}_+^T \cdot \mathbf{v}/\tau)}{\sum_{i=1}^{B} \exp(\mathbf{t}_i^T \cdot \mathbf{v}/\tau)}\right), \quad (1)$$

where $(\mathbf{t}_1^T, \mathbf{t}_2^T, ..., \mathbf{t}_B^T)$ denotes B text features extracted by the text encoder and $\mathbf{v}$ denotes the learned feature of a patent image. The term $\mathbf{t}^T \cdot \mathbf{v}$ represents the cosine similarity score between the patent image and texts and τ is a learnable temperature coefficient. This objective is to maximize $\mathbf{t}_+^T \cdot \mathbf{v}$, which indicates the feature similarity between the patent image and the corresponding textual information.

However, relying solely on instance-based contrastive loss falls short in capturing class or category information, potentially leading to suboptimal performance in datasets with skewed distributions [36]. Therefore, we propose *class*-based and *category*-based coarse-grained losses (i.e., $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{cat}}$, see 2 (d)). These losses can be described in a similar form as follows (2):

$$-\frac{1}{|\mathbf{V}_i^+|} \sum_{\mathbf{V}_j \in \mathbf{V}_i^+} \log \frac{\exp(\mathbf{t}_i^T \cdot \mathbf{v}_j/\tau)}{\sum_{\mathbf{V}_k \in \mathbf{V}} \exp(\mathbf{t}_i^T \cdot \mathbf{v}_k/\tau)}$$

$$-\frac{1}{|\mathbf{T}_i^+|} \sum_{\mathbf{T}_j \in \mathbf{T}_i^+} \log \frac{\exp(\mathbf{t}_j^T \cdot \mathbf{v}_i/\tau)}{\sum_{\mathbf{T}_k \in \mathbf{T}} \exp(\mathbf{t}_k^T \cdot \mathbf{v}_j/\tau)}, \quad (2)$$

where V represents a batch of patent images, and T denotes the corresponding set of text sentences. The subset T_i^+ comprises texts that share the same class (in the case of $\mathcal{L}_{\text{cls}}$) or category (for $\mathcal{L}_{\text{cat}}$) with the image V_i. Similarly, V_i^+ includes all images that share the same class or category with the text T_i. By doing so, our model gains an understanding of the class and category information, enabling it to acquire robust representations even for those in the tail classes. Furthermore, since the text description for each image sample varies with each iteration, and combined with the class or category loss, the one-to-one pairing relationship between images and texts is less rigid. This variability acts as an additional regularization mechanism, preventing the model from adhering to fixed, trivial correlations within specific image-text pairs.

Considering each loss's stability varies, we move away from linear loss combination towards a method based on homoscedastic uncertainty [37, 38], learnable through probabilistic deep learning. This type of uncertainty, independent of input data, reflects the task's intrinsic uncertainty. The loss includes residual regression and uncertainty regularization components. The implicitly learned variance $\hat{s}$ moderates the residual regression, while regularization prevents the network from predicting infinite uncertainty. Hence, The overall loss can be written as in 3.

$$\begin{aligned}
\mathcal{L} = &\mathcal{L}_{\text{clip}} \exp(-\hat{s}_{\text{clip}}) + \hat{s}_{\text{clip}} \\
&+ \mathcal{L}_{\text{cls}} \exp(-\hat{s}_{\text{cls}}) + \hat{s}_{\text{cls}} \\
&+ \mathcal{L}_{\text{cat}} \exp(-\hat{s}_{\text{cat}}) + \hat{s}_{\text{cat}}
\end{aligned} \quad (3)$$

where $\hat{s}$ is learnable homoscedastic uncertainty. We find this loss is robust to our task.

3.3. Text Enrichment

Converting object names, perspectives, class names, and the patent image's original descriptions into text for input into a text encoder is a plausible way to generate text embeddings for supervisory purposes. However, this method encounters several limitations. Firstly, the most comprehensive descriptions provided by patents are typically succinct and straightforward, such as *FIG. 3 is a front elevational view of the light device*, leading to unclear and sparse information about the image. Additionally, the similarity and potential overlap among different classes (e.g., *automobiles*, *motor cars*, and *toy cars*) can obscure the distinction of nuanced concepts. To overcome these challenges and derive more discriminative text features, we employee captioners [34] and LLMs [17] for producing detailed, enriched descriptions that enhance the semantic understanding of the images.

Specifically, we firstly employ captioners to generate descriptions of images in a manner that would capture aspects patent attorneys focus on. We provide the captioners with images alongside a set of predefined instructions

Table 1
Evaluation results across models using metrics such as mAP, Recall@K, and MRR@K, under different settings. Highest values are highlighted in bold.

		Head Classes				Tail Classes				All Classes			
		mAP	R@5	R@10	M@10	mAP	R@5	R@10	M@10	mAP	R@5	R@10	M@10
Baseline	PatentNet [5]	22.1	13.4	26.2	35.7	12.9	3.1	11.6	25.2	15.8	8.0	16.8	29.4
	EfficientNetB0+ArcFace [7]	31.5	19.8	34.9	43.9	20.1	8.7	20.3	33.1	24.1	13.6	26.7	37.4
	SwinV2-B+ArcFace [6]	32.0	19.4	34.2	45.2	20.3	6.3	21.4	32.7	25.1	11.6	27.3	37.7
Ours (backbone)	ResNet50	31.9	35.1	48.9	45.7	22.4	20.9	38.2	36.3	25.6	26.6	42.2	40.1
	EfficientNetB-0	64.2	43.5	58.5	76.4	43.0	31.2	44.0	55.7	52.4	36.9	50.5	63.1
	ViT-B-32	**78.0**	51.7	65.3	90.3	**61.6**	36.7	**53.5**	75.4	**69.1**	43.4	**58.6**	**81.3**
	SwinV2-B	77.2	**52.5**	**66.2**	90.8	61.3	**42.5**	52.5	75.0	67.8	**46.0**	57.7	80.6

that guide the description process. For example, instructions might include: *Describe the distinct visual elements present in the design, such as shapes, contours, texture, and the arrangement of various components.* The outcomes of this process, merged with pre-existing auxiliary information and predetermined instruction templates, are then fed into LLMs. For example, we employ templates such as *This is a photo of* {Object Name}, *classified as* {Class}, *This image features* {Details}, and {Object Name} *can also be referred to as* {Synonym}, to generate enriched text. Ultimately, this approach yields around 20 detailed text descriptions per image, designed to mine the semantic nuances within the text feature space.

Additionally, our research revealed that utilizing more specific object names significantly enhances feature learning. For example, the broad class of *Emergency equipment*, which encompasses distinct items such as *lighting fixture*, *horticulture grow light*, and *lighting device*. Therefore, in our text generation process, we prioritize these detailed object names over the more generic class names, different from the previous approach that typically leans on class categorization [39, 40].

3.4. Metrics for Patent Retrieval

As mentioned before, in image-based patent retrieval, particularly for novelty detection and related work searches, patent professionals explore databases of patents, applications, and scientific literature to determine an invention's uniqueness. This process, crucial both before and after a patent application is filed, focuses on identifying if similar prior inventions exist. Accordingly, when evaluating retrieval metrics, it's necessary to account for the temporal factor, ensuring that only prior art—rather than contemporaneous or subsequent inventions—is considered for retrieval [3]. Considering a database D where each data point is represented by a tuple $(\mathbf{v}, t)$, with $\mathbf{v}$ being the image embedding and t the granted time of the patent the image belongs to. For each query image $\mathbf{v}_q$ with a granted time t_q, define $D'_q \subseteq D$ containing

images granted before t_q. Hence, the following retrieval metrics should be calculated over D'_q given $\mathbf{v}_q$: (*i*) We use mean Average Precision (mAP), a metric obtained by averaging AP scores across all classes. (*ii*) Following previous works, the standard evaluation protocol [41] is to report the recall at rank K (Recall @ K or R@K) at different ranks (5, 10). (*iii*) We calculate the Mean Reciprocal Rank @ K with temporal concern (MRR@K or M@K), which averages the reciprocal of the rank for the first correctly predicted patent image within the top 10 rankings across all test samples.

Table 2
Evaluation results of an ablation study on various components within the model architecture.

	Head Classes		Tail Classes		All Classes	
	mAP	R@10	mAP	R@10	mAP	R@10
Baseline	32.3	34.7	21.9	30.2	26.0	32.8
$\mathcal{L}_{cls}$ and $\mathcal{L}_{cat}$	47.4	49.2	47.0	48.4	47.9	49.7
Text Generation	70.2	59.7	52.4	53.6	58.7	56.3
Captioner	78.0	65.3	61.6	53.5	69.1	58.6

4. Experiments

4.1. Implementation Details

In our experiments, we employed PyTorch [42] and utilized clusters of NVIDIA A100 GPUs. For the VLM, we explored various ViT variants [31], ResNet50 [43], EfficientNetB-0 [44], and SwinV2-B [45, 46] as backbones for the visual encoder. The text encoder was adapted from the original CLIP model [14], remaining fixed throughout the experiments. For captioner, we leveraged open-source BLIP-2 [35] and GPT-4V [47]. Regarding LLMs, our focus was on GPT-4 [47], though we also experimented with other LLMs like GPT-3.5-Turbo [17] and LLaMA-2 [48].

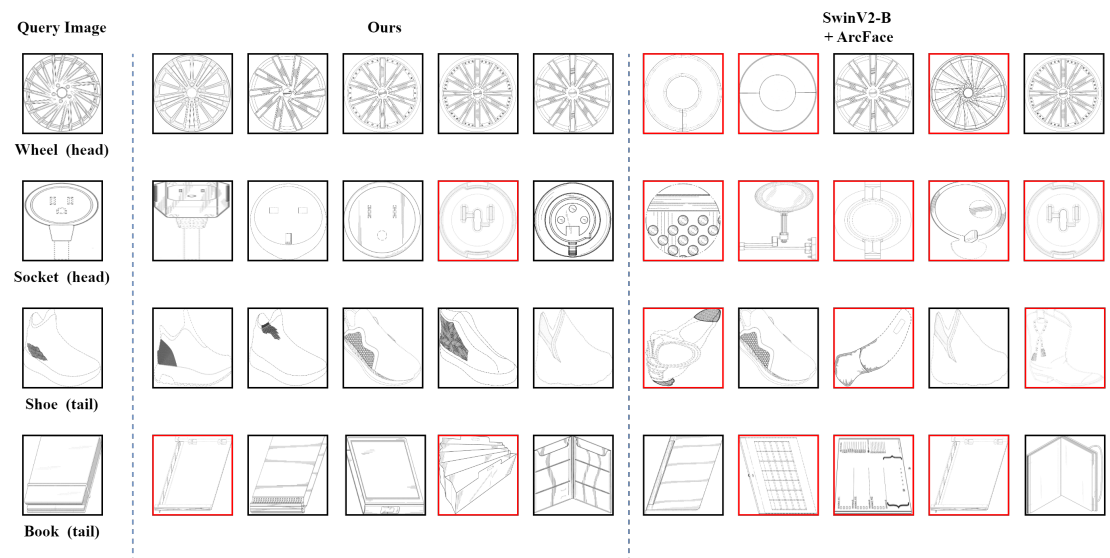For our experiments, we employed the DeepPatent2

Figure 3: Qualitative Results of the Image-Based Patent Image Retrieval System. The leftmost image represents the query image, annotated with its object name and classification as either a head or tail class. The middle section displays the retrieval results from our method, where images framed in black indicate a match with the query image's class, and those framed in red indicate a mismatch. The right section shows the comparative results using the previous state-of-the-art method (SwinV2-B + ArcFace).

dataset. Previous research has mainly adopted the original DeepPatent dataset [5]; however, this dataset suffers from a narrow collection span, lacks image-related metadata, and does not segment sub-images. These limitations can introduce substantial noise in inter-image relationships. Fortunately, the DeepPatent2 [1] dataset addresses these issues effectively. To maintain comparability with previous methods, we utilized DeepPatent2 data from 2016 to 2019, consisting of 822,792 records with 407 Locarno classes. Of these, 90% were used for the training set, and 10% for the validation set. For our query dataset, we used 252,296 records from the year 2020.

Our baseline models are our replicating state-of-the-arts on the same dataset to ensure comparability: these include PatentNet [5], SwinV2-B+ArcFace [6], and EfficientNet+ArcFace [7]. Our primary experiments focused on a variety of visual encoder backbones, such as ResNet50, EfficientNetB-0, ViT-B-32, and SwinV2-B. Due to space constraints in this paper, detailed results from experiments involving the captioner and LLMs will be presented in the full manuscript. For preliminary insights, we utilized GPT-4 for both the captioner and LLMs. In our ablation study, we explore the effects of distribution awareness losses (i.e., $\mathcal{L}_{cls}$ and $\mathcal{L}_{cat}$), the text generation component, and the captioner.

4.2. Experimental Results

Table 1 presents the quantitative results on the DeepPatent2 query set. Overall, our approach significantly outperforms the state-of-the-art, achieving up to a 53% improvement in the mAP metric, 38% in Recall @ 5, 41.8% in Recall @ 10, and 51.9% in MRR @ 10. Notably, both ViT-B-32 and SwinV2-B show comparable performance, with each excelling under different metrics or scenarios. For instance, ViT-B-32 performs better in tail classes, while SwinV2-B shows strength in head classes, suggesting an interaction between data distribution and model architecture. With these results, we have achieved the state-of-the-art in this task. Given the strong performance across all classes with ViT as the backbone, we will base our ablation study on this model to further investigate the impact of various model components on performance.

Table 2 presents the evaluation results of an ablation study on four components, starting with a baseline model, which is a standard CLIP model. The subsequent rows represent enhancements to this baseline. Firstly, by incorporating a distribution-aware contrastive loss, we observe significant performance gains in both head and tail classes, with tail classes experiencing more substantial improvements (approximately 20% in mAP and Recall @ 10), and head classes seeing a 10% increase in these metrics. Next, by incorporating text generation function-
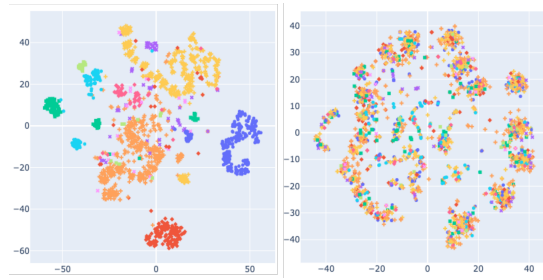
Figure 4: t-SNE Visualization of Image Embeddings. This figure presents a two-dimensional t-SNE projection of randomly sampled 2,000 image embeddings, with each axis representing one dimension. Different colors and shapes in the plot indicate distinct classes. The left subplot illustrates the results from our model, while the right subplot displays the results using the previous state-of-the-art, SwinV2-B+ArcFace.

ality, which is guided by the LLM, the model learns richer semantic relationships between images. The addition of this feature leads to a 10-20% improvement across various metrics. Finally, by integrating a captioner module, the key details of the design inventions in the images are directly expressed by the captioner, further highlighting the semantic focal points of the images and enhancing the retrieval performance.

4.3. Qualitative Results

Qualitative results (see 3) indicate that our approach retrieves better results than the previous state-of-the-art, evident in both head and tail classes. Focusing solely on our model, it is apparent that it underperforms in tail classes. Additionally, our system not only retrieves the correct class given an image but also finds images that are visually similar to the query image. Furthermore, we delve into the errors made by the previous state-of-the-art. For example, when presented with an image labeled *shoe*, the previous model might retrieve a shoe image, but it actually belongs to the category of *shoelaces*. This reveals that the previous approach did not align semantic information within the images, often leading to the retrieval of visually similar but categorically different images. Similar issues occur with categories like *vehicles & toy cars* or *flashlights & chargers*, where the images look alike but differ semantically. Our model mitigates these errors by guiding the image's embedding space with linguistic information, which enhances semantic alignment.

Based on the t-SNE results shown in 4, our method yields more clustered embeddings, suggesting that the model effectively captures the inherent structures or classes within the data. Each cluster represents a group

of similar images, closely corresponding to predefined classes, indicating that the model has learned meaningful and discriminative features for each class. This clustering enables the model to effectively distinguish between different classes.

Conversely, the previous approach results in a t-SNE visualization with less clustered and more continuous embeddings, making it difficult to identify distinct clusters or their correlation to predefined classes. This indicates that the model's learned representations are less discriminative, potentially capturing more generalized features shared across multiple classes. Although some clustering is visible, it may not relate to actual classes but rather to visually similar images, blurring the boundaries between different classes.

5. User Study

To further ensure the practical value of our system, we conducted a user study following rigorous psychological procedures. As for participants, we recruited 15 patent agents (48% female, average age: 33.4 years) to perform tasks related to design patent image retrieval.

As for procedure, we employed a double-blind test, where participants were unaware of the underlying retrieval system during their tasks. They could encounter either our retrieval system or a system based on the previous approach. Each patent agent handled 30 retrieval tasks, with these tasks randomly assigned to one of the two systems—15 tasks with our system and 15 with the previous approach. After each task, participants rated their satisfaction with the retrieval results on a scale from 1 to 5 and recorded the time taken to complete the task (in hours). To minimize randomness, we averaged the scores across systems for each participant, resulting in four scores per person (two systems × two scores).

The results of the paired t-test revealed significant differences in satisfaction levels, with patent agents showing a higher satisfaction with our system compared to the previous approach, $t(14) = 3.30, p < 0.01$. Regarding task completion time, agents completed tasks faster using our system, $t(14) = -4.30, p < 0.001$. These results indicate that our system is more efficient and better meets the practical needs of professionals in the field.

6. Conclusion

Our method has achieved new state-of-the-art results in the quantitative evaluation of mAP, Recall@K, and MRR@K, as well as in high-quality image retrieval during qualitative evaluation. For many years, current commercial design patent retrieval systems have had significant shortcomings. For example, traditional text-based

searches can be limiting due to the subjective interpretation of design features and the difficulty in describing visual details with text. Although learning-based image retrieval systems have started to emerge in the last two years, their practical value remains limited.

To address these issues, our proposal can effectively solve this problem. Firstly, we proposed a new learning-based architecture capable of learning image features with practical value. These representations not only contain visual information but are also aligned with corresponding (augmented) semantic text and classification data. This has substantial practical value because specific graphic semantic features such as curvature, edges, and geometric details are considered. Focusing solely on the image itself might overlook these critical visual elements. Secondly, we utilize a larger and more well-defined dataset, which encompasses a broader collection span, image-related metadata, and segmented sub-images. This makes the model more robust and enhances its accuracy. Thirdly, addressing the long-tail distribution in classification, our study is the first to propose distribution-aware losses, which have proven to be effective. Lastly, we conducted a user study to demonstrate the practical value of our system in the field, showing that it is more efficient, accurate, and time-saving.

In the future, we have several directions for further expansion: (*i*) Identifying similarities between this invention and prior arts, which involves not only using existing models for visualizations through explainable AI [49] but also leveraging data from examiner's reports (Office Actions) to guide further explorations. (*ii*) While our research currently focuses on prior art searches, future work could also explore other temporal dimensions, such as infringement searches. Additionally, image domain adaptation [50] could be used to enhance the effectiveness of searches across different domains, such as retrieving E-commerce images using patent drawings.

References

[1] K. Ajayi, X. Wei, M. Gryder, W. Shields, J. Wu, S. M. Jones, M. Kucer, D. Oyen, Deeppatent2: A large-scale benchmarking corpus for technical drawing understanding, Scientific Data 10 (2023) 772. doi:10.1038/s41597-023-02653-7.

[2] W. Shalaby, W. Zadrozny, Patent retrieval: a literature review, Knowledge and Information Systems 61 (2019) 631–660. doi:10.1007/s10115-018-1322-7.

[3] R. Krestel, R. Chikkamath, C. Hewel, J. Risch, A survey on deep learning for patent analysis, World Patent Information 65 (2021) 102035. doi:10.1016/j.wpi.2021.102035.

[4] S. Jiang, J. Luo, G. Ruiz-Pava, J. Hu, C. L. Magee, Deriving design feature vectors for patent images using convolutional neural networks, Journal of Mechanical Design 143 (2021) 061405. doi:10.1115/1.4049214.

[5] M. Kucer, D. Oyen, J. Castorena, J. Wu, Deeppatent: Large scale patent drawing recognition and retrieval, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 2309–2318. doi:10.1109/WACV51458.2022.00063.

[6] K. Higuchi, K. Yanai, Patent image retrieval using transformer-based deep metric learning, World Patent Information 74 (2023) 102217. doi:10.1016/j.wpi.2023.102217.

[7] H. Wang, Y. Zhang, Learning efficient representations for image-based patent retrieval, in: Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Springer, 2023, pp. 15–26. doi:10.1007/978-981-99-8540-1_2.

[8] R. N. Carney, J. R. Levin, Pictorial illustrations still improve students' learning from text, Educational psychology review 14 (2002) 5–26. doi:10.1023/A:1013176309260.

[9] S. Vrochidis, A. Moumtzidou, I. Kompatsiaris, Concept-based patent image retrieval, World Patent Information 34 (2012) 292–303. doi:10.1016/j.wpi.2012.07.002.

[10] A. Tiwari, V. Bansal, Patseek: content based image retrieval system for patent database (2004).

[11] Z. Zhiyuan, Z. Juan, X. Bin, An outward-appearance patent-image retrieval approach based on the contour-description matrix, in: 2007 Japan-China Joint Workshop on Frontier of Computer Science and Technology (FCST 2007), IEEE, 2007, pp. 86–89. doi:10.1109/FCST.2007.14.

[12] B. Huet, N. J. Kern, G. Guarascio, B. Merialdo, Relational skeletons for retrieval in patent drawings, in: Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205), volume 2, IEEE, 2001, pp. 737–740. doi:10.1109/ICIP.2001.958599.

[13] S. Vrochidis, S. Papadopoulos, A. Moumtzidou, P. Sidiropoulos, E. Pianta, I. Kompatsiaris, Towards content-based patent image retrieval: A framework perspective, World Patent Information 32 (2010) 94–106. doi:10.1016/j.wpi.2009.05.010.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763. doi:10.48550/arXiv.2103.00020.

[15] Z. Guo, Y. Tang, R. Zhang, D. Wang, Z. Wang, B. Zhao, X. Li, Viewrefer: Grasp the multi-

view knowledge for 3d visual grounding with gpt and prototype guidance, arXiv preprint arXiv:2303.16894 (2023). doi:`10.48550/arXiv.2303.16894`.

[16] R. Zhang, X. Hu, B. Li, S. Huang, H. Deng, Y. Qiao, P. Gao, H. Li, Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15211–15222. doi:`10.48550/arXiv.2303.02151`.

[17] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901. doi:`10.48550/arXiv.2005.14165`.

[18] C. Tian, W. Wang, X. Zhu, J. Dai, Y. Qiao, Vl-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition, in: European Conference on Computer Vision, Springer, 2022, pp. 73–91. doi:`10.48550/arXiv.2111.13579`.

[19] L. B. Christensen, R. B. Johnson, L. A. Turner, Research Methods, Design, and Analysis, 12 ed., Global Edition, 2014. Page count: 542; Dimensions: 23 x 18.6 cm; Book number: 00106962.

[20] M. Bhattarai, D. Oyen, J. Castorena, L. Yang, B. Wohlberg, Diagram image retrieval using sketch-based deep learning and transfer learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 174–175. doi:`10.48550/arXiv.2004.10780`.

[21] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S. C. Hoi, Deep learning for person re-identification: A survey and outlook, IEEE transactions on pattern analysis and machine intelligence 44 (2021) 2872–2893. doi:`10.1109/TPAMI.2021.3054775`.

[22] X.-Q. Ma, C.-C. Yu, X.-X. Chen, L. Zhou, Large-scale person re-identification based on deep hash learning, Entropy 21 (2019) 449. doi:`10.1109/TIP.2017.2695101`.

[23] Z. Liu, C. Rodriguez-Opazo, D. Teney, S. Gould, Image retrieval on real-life images with pre-trained vision-and-language models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2125–2134. doi:`10.48550/arXiv.2108.04024`.

[24] S. Karthik, K. Roth, M. Mancini, Z. Akata, Vision-by-language for training-free compositional image retrieval, 2024. doi:`10.48550/arXiv.2310.09291`. arXiv:`2310.09291`.

[25] A. Baldrati, L. Agnolucci, M. Bertini, A. Del Bimbo, Zero-shot composed image retrieval with textual inversion, arXiv preprint arXiv:2303.15247 (2023). doi:`10.48550/arXiv.2303.15247`.

[26] N. Cohen, R. Gal, E. A. Meirom, G. Chechik, Y. Atzmon, "this is my unicorn, fluffy": Personalizing frozen vision-language representations, in: European Conference on Computer Vision, Springer, 2022, pp. 558–577. doi:`10.48550/arXiv.2204.01694`.

[27] C. Deng, X. Xu, H. Wang, M. Yang, D. Tao, Progressive cross-modal semantic network for zero-shot sketch-based image retrieval, IEEE Transactions on Image Processing 29 (2020) 8892–8902. doi:`10.1109/TIP.2020.3020383`.

[28] P. Sangkloy, N. Burnell, C. Ham, J. Hays, The sketchy database: learning to retrieve badly drawn bunnies, ACM Transactions on Graphics (TOG) 35 (2016) 1–12. doi:`10.1145/2897824.2925954`.

[29] P. Xu, T. M. Hospedales, Q. Yin, Y.-Z. Song, T. Xiang, L. Wang, Deep learning for free-hand sketch: A survey, IEEE transactions on pattern analysis and machine intelligence 45 (2022) 285–312. doi:`10.48550/arXiv.2001.02600`.

[30] H. Wang, C. Deng, T. Liu, D. Tao, Transferable coupled network for zero-shot sketch-based image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (2021) 9181–9194. doi:`10.1109/TPAMI.2021.3123315`.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020). doi:`10.48550/arXiv.2010.11929`.

[32] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 13001–13008. doi:`10.48550/arXiv.1708.04896`.

[33] P. Chen, S. Liu, H. Zhao, J. Jia, Gridmask data augmentation, arXiv preprint arXiv:2001.04086 (2020). doi:`10.48550/arXiv.2001.04086`.

[34] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, Advances in neural information processing systems 36 (2024). doi:`10.48550/arXiv.2304.08485`.

[35] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, arXiv preprint arXiv:2301.12597 (2023). doi:`10.48550/arXiv.2301.12597`.

[36] H. C. Lo, C.-S. Fuh, Enhancing long-tailed 3d semantic segmentation with category-wise linguistic-visual representation, in: The 36th IPPR Conference on Computer Vision, Graphics, and Image Processing (CVGIP), Kinmen, Taiwan, 2023.

[37] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, Ad-

vances in neural information processing systems 30 (2017). doi:10.48550/arXiv.1703.04977.

[38] A. Kendall, R. Cipolla, Geometric loss functions for camera pose regression with deep learning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5974–5983. doi:10.48550/arXiv.1704.00390.

[39] D. Rozenberszki, O. Litany, A. Dai, Language-grounded indoor 3d semantic segmentation in the wild, in: European Conference on Computer Vision, Springer, 2022, pp. 125–141. doi:10.48550/arXiv.2204.07761.

[40] R. Huang, X. Pan, H. Zheng, H. Jiang, Z. Xie, C. Wu, S. Song, G. Huang, Joint representation learning for text and 3d point cloud, Pattern Recognition 147 (2024) 110086. doi:10.48550/arXiv.2301.07584.

[41] A. Baldrati, M. Bertini, T. Uricchio, A. Del Bimbo, Effective conditioned and composed image retrieval combining clip-based features, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 21466–21474.

[42] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Advances in neural information processing systems 32 (2019). doi:10.48550/arXiv.1912.01703.

[43] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778. doi:10.48550/arXiv.1512.03385.

[44] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114. doi:10.48550/arXiv.1905.11946.

[45] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022. doi:10.48550/arXiv.2103.14030.

[46] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al., Swin transformer v2: Scaling up capacity and resolution, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 12009–12019. doi:10.48550/arXiv.2111.09883.

[47] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023). doi:10.48550/arXiv.2303.08774.

[48] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023). doi:10.48550/arXiv.2307.09288.

[49] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, J. Zhu, Explainable ai: A brief survey on history, research areas, approaches and challenges, in: Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8, Springer, 2019, pp. 563–574. doi:10.1007/978-3-030-32236-6_51.

[50] M. Wang, W. Deng, Deep visual domain adaptation: A survey, Neurocomputing 312 (2018) 135–153. doi:10.48550/arXiv.1802.03601.