EPO Academic Research Programme 2021

Visual and Multimodal Patent Search at Scale (ViP@Scale)

Project Report

SUSHIL AWALE, ERIC MÜLLER-BUDACK,
AND RALPH EWERTH

March 31, 2025

TIB – Leibniz Information Centre for Science and Technology
Hannover, Germany

Table of Contents

1 Introduction

Patents are inherently multimodal documents, incorporating both textual descriptions and visual representations to protect inventions. This multimodal nature necessitates a patent retrieval system capable of efficiently processing both text and images. However, current research in patent retrieval has primarily concentrated on text-based approaches [12], leaving a significant gap in the field. Implementing a multimodal paradigm in patent retrieval could offer several benefits [29]. A robust visual patent retrieval system can, for example, allow for quicker, more focused searches, reducing the time required to find relevant patents for patent examiners [23]. This efficiency gain is particularly valuable given the ever-increasing volume of patent applications[1]. Additionally, by considering both textual and visual elements, the system can potentially identify a broader range of relevant patents, enhancing the overall recall of the search process. As a result, the risk of infringement-related litigation post-publication could be decreased [23].

1.1 Objectives

The main goal of the project *Visual and Multimodal Patent Search at Scale* (ViP@Scale) was to research and develop novel approaches and models for visual and multimodal patent retrieval. Patent images, which are often black-and-white images, are starkly different to natural images. Content in natural images have colour and texture, whereas in patent images objects are characterized by geometric features, i.e., geometric shapes and spatial relations between them [21]. In case of natural images, tremendous progress has been made to improve the retrieval methods due to powerful deep learning approaches, including (large) vision-language models (VLMs) [30]. However, these approaches are not explicitly optimized for patent images. Therefore, adapting VLMs pre-trained for the utility patent domains can drastically improve results for image patent search. With this motivation, the first objective of the project was adapting VLMs pre-trained on natural images to binary patent images by fine-tuning on a large patent image corpus. With the development of multimodal patent representation models, the second objective was the application of the representation models for downstream tasks such as retrieval, clustering, and classification. The final objective included the evaluation of the representation models on patent retrieval datasets. The main objectives of this project can be summarized as follows:

- Large-scale multimodal representation models for patent images,
- Application of multimodal models to downstream tasks such as patent retrieval and classification regarding core concepts (e.g., objects, patent domain, image type), and
- Evaluation of multimodal models on patent retrieval benchmark datasets

1.2 Related Work

Numerous research works have been conducted in the field of patent retrieval, with a predominant focus on text-based retrieval approaches [12]. This emphasis can be attributed to the fact that textual data serve as the primary medium for describing inventions in patents, and recent advancements in machine learning have significantly facilitated the analysis of textual information. In addition, most open-source patent retrieval datasets [1, 13, 18] have either provided only textual content for retrieval, or the relevance judgement is primarily based on textual or categorical metadata such as claims, abstract, and *International Patent Classification* (IPC) as well as *Cooperative Patent Classification* (CPC) codes. However, visual elements in patents provide crucial supplementary information about the inventions through innovative illustrations, in particular drawings. These visual components often convey details that the text alone cannot

[1] `https://report-archive.epo.org/about-us/annual-reports-statistics/statistics/2022/statistics/patent-applications.html`

adequately express and offer information that transcends linguistic and domain barriers inherent in patent documents [23, 29]. To this extent, studies on image-based patent retrieval [9, 14] and multimodal patent retrieval [19, 21, 22] have demonstrated that the incorporation of patent images can offer substantial benefits to the retrieval process.

In the early works on patent image retrieval, low-level vision-based methods extracted features like shape contours, relational skeletons, and *Adaptive Hierarchical Density Histograms* (AHDH) [23]. For example, PATSEEK [25] used *Edge Orientation Autocorrelograms* (EOAC) for shape-based retrieval, while Csurka et al. [3] use a *Fisher Vector* representation for images. These works incorporated additional metadata and figure descriptions for hybrid queries to address the limitations of purely visual methods [23], but these systems struggled with scalability and semantic understanding, often failing to generalize across diverse patent classes [16, 26].

The advent of deep learning introduced a new generation of approaches. The *DeepPatent* dataset [13] enabled end-to-end trainable models such as *PatentNet* [9] and others [14, 27]. In recent works on design patents [16, 22], multimodal architectures integrated visual and textual data to train VLMs such as CLIP [20] enhanced by generated descriptions from *large language models* (LLMs) and *large vision-language models* (LVLMs). While deep learning methods have significantly improved visual representations in patents, a major focus has been on design patent images. Although design and utility patents images are both black-and-white and abstract, design patent images generally show details on design aspects, while images in utility patents focus more on depicting the applicability of the inventions. Therefore, challenges still remain in developing a novel multimodal representation model for utility patent images.

2 Work Package Results

This section presents a comprehensive overview of the project's outcomes, organized according to the work packages (WPs) outlined in Table 1. Each work package represents a distinct phase or aspect of the project, collectively contributing to the project's overall objectives (Section 1.1). In the following subsections, we will summarize the main results, key findings, and outcomes of each work package.

Table 1: List of work packages for the project ViP@Scale

Work Package	Title
1	Requirements and interfaces for the integration in EPO services
2	Instance segmentation in patent images
3	Multimodal representation learning for patent images
4	Multimodal pre-search and classification of patents
5	Efficient models and evaluation at scale
6	Project Management, Coordination and Dissemination

In general, the proposed work packages were addressed as planned. However, during the course of the project, TIB and EPO jointly decided that approaches for instance segmentation in patent images (WP 2) are less crucial for patent retrieval than some other directions. Thus, WP 2 (Section 2.2) was replaced as follows. (1) Due to the significance and high impact of generative AI models in other domains [30], we explored how to use generative AI models for multimodal representation learning, pre-search, and classification of patents as explained in WP 3 (Section 2.3) and WP 4 (Section 2.4). (2) During the project, we noticed that benchmarks datasets for patent image retrieval, including the one provided by EPO, lack completeness, i.e., many relevant images are not labeled as such. This problem originates from the strategies that are used to automatically create these benchmarks [18]. In many cases, weak indicators such as citations are used to judge the relevance of patent images regarding a query patent. But citations are typically incomplete as typically only the most relevant works are cited. Alternatively, the task of patent image retrieval is reformulated, e.g., to find images from the same patent [13], which does not match typical use cases for patent image search. As a result, we focused on creating a high-quality benchmark dataset to reliably evaluate patent image retrieval as explained in WP 5 (Section 2.5).

Deliverables

Based on the WPs, the deliverables listed in Table 2 were agreed upon for the project. All the deliverables were submitted to EPO, and are referenced with the corresponding WPs. Due to the deviations described above, some adjustments were made for deliverable D2.2, as further explained in WP 2 (Section 2.2).

Table 2: List of deliverables for the project ViP@Scale

Deliverables	Title	Work Packages
D1	Requirements for multimodal patent search at EPO	1
D2.1	Dataset of annotated patent images	5
D2.2	Novel instance segmentation methods for patent images	2
D3.1	Dataset of multimodal patent information (images with related text)	3
D3.2	Multimodal deep learning model for patent data representation	3
D4.1	Models for downstream tasks	3, 4
D4.2	Optimized models for downstream tasks	3, 4
D5.1	Efficient models	5
D5.2	Results for large-scale evaluations	5
D6.1	Regular progress reports	6
D6.2	Publications in reputable venues	3, 4, 5
D6.3	Workshop organisation	6

2.1 WP 1: Requirements and Interfaces for the Integration in EPO Services

Work Package 1 (WP 1) focused on defining the system requirements and interfaces for integration into EPO services. This process involved comprehensive discussions of EPO's existing technology stack and systems, leading to the identification of key integration points. The initial requirements and interfaces were established during the bi-weekly meetings between EPO and TIB. These discussions resulted in the development of technical specifications, data exchange protocols, and the adoption of a common technology stack. The team adopted an agile approach, allowing for modifications as the project progressed.

Key Outcomes

- To simplify integration and for quick prototyping, the visual patent embeddings from the multimodal models were shared using *Google Cloud Platform.*
- For rapid prototyping, *Streamlit* was selected as the primary tool, which is also widely used at EPO.

2.2 WP 2: Instance Segmentation in Patent Images

Work Package 2 (WP 2) aimed to research and develop novel instance segmentation methods for patent images. To achieve this, the initial task was to focus on constructing a multimodal patent data corpus suitable for training and evaluating representation models. Subsequent efforts were to target the development of metadata extraction methods for patent images and segmentation techniques to isolate depicted objects. However, much of this work proved redundant when EPO provided TIB a pre-existing corpus, which already contained comprehensive metadata and segmentation boundaries for patent images (described in detail in WP 3 (Section 2.3). Moreover, TIB and EPO discussed on state-of-the-art approaches for instance segmentation including Meta's *Segment Anything* [11] that could be used to further refine existing results. Thus, as mentioned above, EPO and TIB jointly decided to instead extend other work packages.

2.3 WP 3: Multimodal Representation Learning for Patent Images

The goal of WP 3 was to adapt existing VLMs such as the CLIP model (Contrastive Language-Image Pretraining; [20]) to the patent domain. The CLIP model provides a multimodal vision-language representation by learning visual concepts from natural language supervision. Using a contrastive learning approach on large-scale datasets containing image-text pairs, it projects text and image inputs into a shared embedding space (shown in Figure 1). CLIP has demonstrated robust performance across diverse visual tasks, benchmarks, and domains including in patents [16, 19, 22]. However, these works have used CLIP either without any fine-tuning on patents [19] or by optimizing it based on design patent figures [16, 22], which are different to utility patent figures.

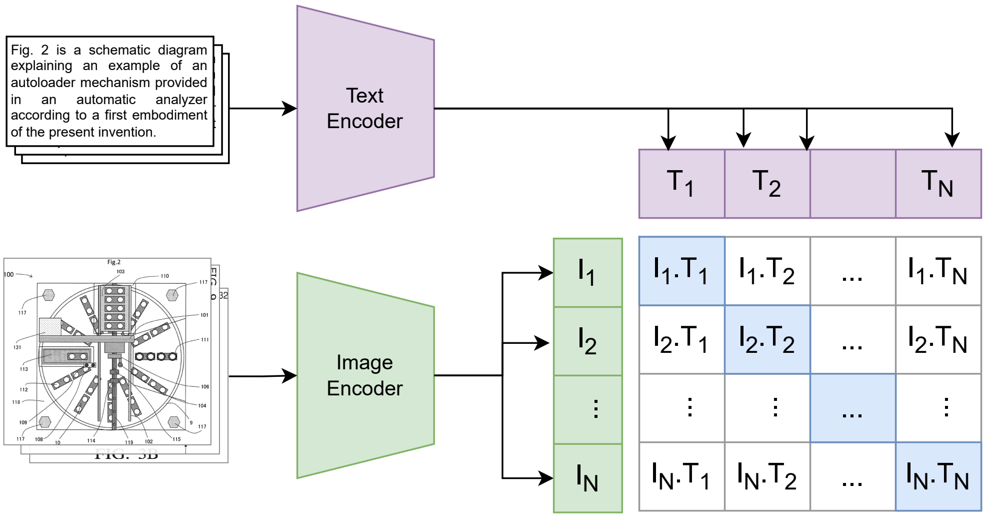


Fig. 2 is a schematic diagram explaining an example of an autoloader mechanism provided in an automatic analyzer according to a first embodiment of the present invention.
Text Encoder
T_1
T_2
T_N
Image Encoder
I_1
I_2
⋮
I_N
$I_1.T_1$
$I_1.T_2$
...
$I_1.T_N$
$I_2.T_1$
$I_2.T_2$
...
$I_2.T_N$
⋮
⋮
...
⋮
$I_N.T_1$
$I_N.T_2$
...
$I_N.T_N$

Figure 1: Contrastive language-supervised image pre-training of the CLIP model for utility patent domain using extracted figure descriptions. The CLIP model uses a text and image encoder to project image-text pairs into a shared embedding space. For this purpose, it maximizes the similarity of associated pairs of image I_n and text T_n.

To address the research gap described above, we focused on adapting the CLIP model [20] to the utility patent domain. This includes a variety of steps ranging from collecting and pre-processing a multimodal patent corpus that allows for language-supervised training over generative AI models to create or further refine descriptions of patent images to the fine-tuning appropriate VLMs for patent retrieval. In the following, details on the individual steps are provided.

2.3.1 EPO Utility Patent Corpus

A prerequiste to adapt the CLIP model for the utility patent domain, is to gather a suitable datasets comprised of image-text pairs. For this purpose, we leveraged a comprehensive multimodal corpus comprising images and related text of utility patents provided by the EPO which is referred to as *EPO Utility Patent Corpus* in the remainder of this report. The *EPO Utility Patent Corpus* contains patents across nine CPC sections, encompasses over one million patents, and contains approximately ten million patent images. Beyond the standard text descriptions and visual representations, the corpus is enriched with valuable metadata such as figure categories, figure labels, etc. associated with each patent image. Further, to facilitate the training and evaluation of patent retrieval models, the corpus is divided into three distinct groups: TRAIN, INDEX, and QUERY. A detailed breakdown of this dataset, including relevant statistics, is presented in Table 3 and 4. The image-text pairs required for training the CLIP model are taken from the patents in the TRAIN split, and the patents in the INDEX and QUERY splits are used for evaluation (discussed further in Section 2.5.1).

Dataset Group	# Patents	# Figures
TRAIN	66,290	595,629
INDEX	1,002,374	9,302,246
QUERY	1,756	20,699

Table 3: Number of patents and figures within the *EPO Utility Patent Corpus* for different dataset splits

Figure Type	TRAIN	INDEX	QUERY
Chemical	7,278	153,209	146
Code	1,462	71,454	53
Circuit	30,387	725,836	1,143
Diagram	72,590	1,149,266	2,326
Drawing	523,282	8,772,323	14,053
Flowchart	46,993	776,285	1,459
Geneseq	4,284	195,419	677
Graph	167,126	2,146,207	4,598
Math	623	25,363	36
Photo	39,238	420,071	789
Table	12,790	270,271	283
Text	927	21,553	33

Table 4: Number of images per figure type for each dataset split

Metadata

The *EPO Utility Patent Corpus* consists of metadata both at patent-level such as patent title, CPC codes, and full text description, and at sub-figure-level. A few selected sub-figure-level metadata are listed in Table 5. An example of a figure from the *EPO Utility Patent Corpus* with selected metadata is shown in Figure 2.

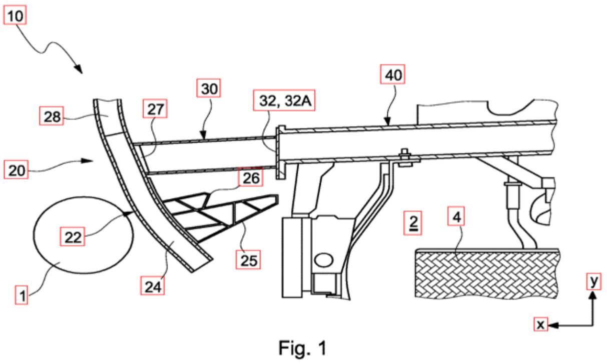


Figure Label: Fig. 1
Context Surrounding Figure Label:
1. Here, the drawings show: Fig. 1 a diagrammatic representation of a
2. to the invention from Fig. 1 at the time of the

Component Terms:
20 bumper crossmember
27 connection arrangement
30 crash box
22 end portions
25 first supporting element
40 longitudinal member
...

Figure 2: Example of a figure (left) from the patent US2019256021A1 randomly sampled from the *EPO Utility Patent Corpus* with select extracted metadata (right).

This carefully curated and organized corpus forms the foundation for our efforts to adapt

Table 5: List of selected metadata associated with each sub-figure in the patent image available as part of the *EPO Utility Patent Corpus*.

ID	Metadata	Description
F1	Type	Type of the sub-figure
F2	Type confidence score	Confidence score of type classifier model
F3	Label text	Text of the sub-figure label e.g. Fig. 1, Fig. 3A, etc.
F4	Bounding box coordinates for label	Coordinates of bounding box encapsulating a sub-figure label
F5	Bounding box coordinates	Coordinates of bounding box encapsulating a sub-figure
F6	Mentions	Snippets of text extracted from full text description containing the figure label text
C1	Component references	Alphanumeric references of all components in a sub-figure e.g. 32A
C2	Bounding box coordinates for component references	Coordinates of bounding box encapsulating component references
C3	Component term	Text labels associated with each component reference
C4	Component term mentions	Snippets of text extracted from full text description containing the component terms

CLIP to the specialized domain of utility patents, enabling more effective and accurate patent retrieval systems. The corpus's comprehensive coverage of various patent categories and rich metadata enhances its utility for training and ensures that the model can generalize well. However, to use the dataset for training the CLIP model, further filtering and preprocessing steps are beneficial. In Section 2.3.2, we discuss the filtering and preprocessing steps applied for this purpose. These steps are essential for removing noise and irrelevant data, thereby enhancing the model's performance in patent retrieval tasks.

2.3.2 Preprocessing Steps

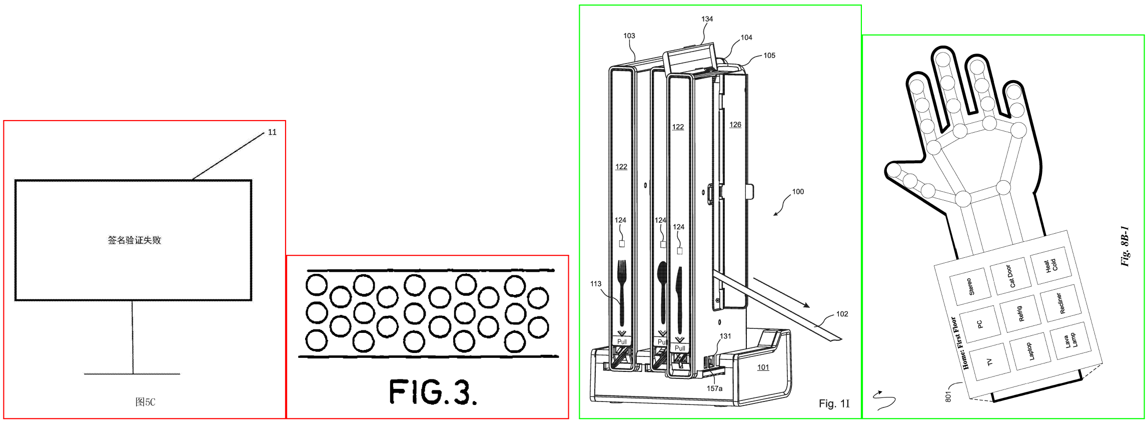
We performed the following preprocessing steps to use the *EPO Utility Patent Corpus* for fine-tuning CLIP.

Filtering by Figure Type

First, we jointly decided with EPO to only consider figures of type DRAWING (metadata F1; Table 5) since they best describe the invention in the patent. In addition, prior works [3, 1] have also shown that selecting only DRAWINGS results in the best performance in retrieval task. Further, we filter out figures of type DRAWING that were classified by an in-house model with a confidence score (metadata F2; Table 5) lower than 0.85 based on discussion with EPO.

Filtering Non-informative Images

In addition to filtering only DRAWINGS, we also filtered out "non-informative" images, i.e., patent images that are abstract and do not show any distinguishing or informative content. Figure 3 shows examples of both "informative" and "non-informative" patent images. For this filtering process, a binary classifier based on a CNN was provided by EPO.



Non-informative Non-informative Informative Informative

Figure 3: Examples of "non-informative", and "informative" patent images. The filtering process filters out the "non-informative" images

Segmentation and Padding

Utilizing the bounding box coordinates (metadata F5; Table 5), we cropped the patent figures to include only the figure content. The segmentation procedure eliminates information such as patent number, patent issuing office, and others that are irrelevant (or which can even add bias) for a language-supervised training. If available, this process also outputs the individual sub-figures with a drawing. For the adaptation of the CLIP model, we set the granularity of the patent figure at the sub-figure level. Additionally, we padded the images to ensure they had equal height and width, preventing any distortion or squeezing of their content.

Extraction of Descriptions and Component Terms

Next, for each sub-figure, we extracted **figure descriptions** (DESC) from the full text of the patent. The patent sub-figures and the extracted figure descriptions form the image-text pairs required for adapting the CLIP model (discussed in detail in Section 2.3.4) to the patent domain. To perform the extraction, we designed a rule-based algorithm that leverages regular expressions. This algorithm processes figure labels (metadata F3; Table 5) and figure mentions (metadata F6; Table 5) as input and generates the corresponding figure description from the full text. The extracted descriptions provide a concise summary of the sub-figure content. An example of such an extracted description is illustrated in Figure 1. However, extracting figure descriptions is not always feasible due to certain limitations. One key challenge arises from the optical character recognition (OCR) system used to extract figure labels, which occasionally fails to retrieve labels for some sub-figures. Consequently, we exclude these images from further processing since their corresponding figure descriptions cannot be automatically obtained. The final count of figures after the filtering process is shown in Table 6.

As an alternative source of text associated with the patent figures, we also utilized the available list of **component terms** (TERMS). Examples of component terms are shown in Figure 2. For fine-tuning CLIP using component terms, we arranged the terms in a comma-separated string prefixed with "an image composed of".

Formatting and Masking

For the next preprocessing step, we formatted the extracted figure descriptions by replacing figure labels such as "Figure 1" with a special token $< image >$, and masked the figure labels and component references with white patches. The main purpose for this is to prevent CLIP from establishing image-text associations based on these figure labels and component references, which may result in over-fitting to the training data.

We applied the preprocessing and filtering steps to all three splits of the *EPO Utility Patent Corpus* (Section 2.3.1) resulting in the following statistics shown in Table 6. We used this resulting dataset for training and evaluation purposes.

Table 6: Number of patents and figures within the *EPO Utility Patent Corpus* for different dataset splits after preprocessing

Dataset Group	# Patents	# Drawings
TRAIN	120,982	333,667
INDEX	660,011	4,825,159
QUERY	1,153	7,612

Beyond the above preprocessing steps, we also explored use of synthetic figure captions generated using LVLMs that address some short comings of using present existing patent figure descriptions and component terms. Section 2.3.3 elaborates further on this process.

2.3.3 Generating Figure Captions

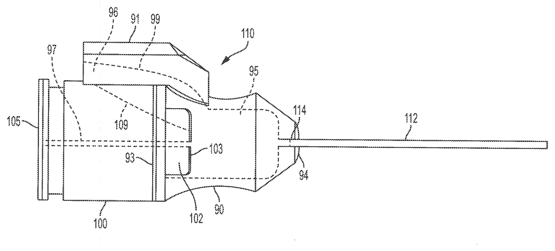


FIG. 11

Patent Title: Laser Video Endoscope

DESC: Fig. 11 is a sectional view of an assembled hand piece including distal and proximal portions and a probe of an endoscope according to an exemplary embodiment of the present invention

TERMS: An image composed of face, proximal end, probe, trifurcation zone, second portion, proximal portion, channels, fibers, opening, surface, channel, portion, hand piece, distal end

GenDESC: An image of an assembled hand piece including distal and proximal portions and a probe of an endoscope.

GenTERMs: An image of a hand piece with a probe and channels.

Figure 4: Fig 3: Example of a figure (left) from the patent WO2017100651A1 with different text captions for training CLIP (right). DESC and TERMS are the extracted figure descriptions and component terms respectively and GenDESC and GenTERMS are synthetic figure captions generated with LVLM using DESC and TERMS as context respectively.

The extracted figure descriptions may sometimes lack depth or specificity required, for example, "FIG. 1 and 2 are views for showing the prior art." or "Fig. 1 is schematic structural view of the invention.". Such descriptions, while technically accurate, offer minimal informative value about the specific contents or unique features of the figures they reference. Similarly, descriptions like "Figure 3 is a perspective view of the invention shown in Figure 1 in use." provide little additional context beyond indicating a change in viewpoint. These vague and non-descriptive captions pose significant challenges for language-supervised training of a vision-language model, as they lack the specificity and depth necessary for meaningful image-text associations. Similarly, component terms, while useful for identifying individual components, may not adequately represent the holistic nature of the invention or its unique features. For instance, a list of terms like "gear, shaft, lever" fails to capture the innovative mechanism or system they collectively form.

To address the aforementioned problems, we explored how to use large vision-language models (LVLMs) to generate more comprehensive and contextually relevant captions for patent images. LVLMs can understand the complex visual and textual relationships in patent figures, potentially producing more informative and accurate captions that better represent the invention's key features and overall concept. We used LLaVA [15], a state-of-the-art LVLM, to generate figure captions for patent figures using the following in-context prompting strategy. In-context prompting strategy is a technique used to enhance the performance of LLMs and LVLMs by providing relevant examples or instructions within the input prompt. This approach leverages the model's ability to adapt to new tasks without requiring additional training or fine-tuning. For example,

1. GenDESC: Use the patent image and the extracted figure description (DESC) as input context.

2. GenTERMS: Use the patent image along with the list of component terms (TERMS) as input context.

For both the variations of figure captions, we used the following LVLM prompt.

> $< image >$ For the patent image, please generate a brief CLIP-like image caption that starts with "an image of ..." and includes no mention of other figures. Use $< metadata = \{\text{DESC}, \text{TERMS}\} >$ as a reference.

Here, in this prompt template, $< image >$ and $< metadata >$ are replaced with the patent image and respective metadata information (DESC or TERMS). Figure 4 shows examples of generated figure caption.

2.3.4 Contrastive Language-image Pre-training for Utility Patent Domain

The extracted figure descriptions (DESC) and component terms (TERMS) as well as generated captions (GenDESC, GenTERMS), allowed us to formulate four distinct sets of image-text pairs that can be used to adapt CLIP to the patent domain resulting in four adapted CLIP models CLIP-DESC, CLIP-TERMS, CLIP-GenDESC, and CLIP-GenTERMS. A comparison of models using these different types of image-text pairs is provided in WP 5. We follow Goyal et al. [8] to pre-train CLIP using contrastive language-supervised image pre-training. This approach, called FLYP (Finetune Like You Pre-train), uses the same contrastive loss for adapting the model as was used during pre-training. The work demonstrated that mimicking the pre-training process lead to significant performance improvements.

Implementation Details

For the training of the CLIP model, we used the open source CLIP model training code [10]. For the training architecture, we use the vision transformer variant ViT-B-16 as vision encoder and a vanilla transformer model as text encoder. We initialized the weights using a pre-trained model named laion400m_e32, which was pre-trained on 400 million natural image-text pairs. We optimized the model for $33,000$ iterations with early stopping and using AdamW optimizer with a batch size of 256 and a learning rate of $3.54e - 6$. The hyperparameters and model architecture were selected using hyperparameter tuning.

We conducted comprehensive hyperparameter optimization across multiple dimensions:

1. **Model architecture**: Vision Transformer variants (ViT-B-16, ViT-B-32)

2. **Optimization parameters**: Learning rates (10^{-3} to 10^{-6}), weight decay values (0.1, 0.2) for AdamW

Further the preprocessing pipeline was critical due to the domain gap between natural images and patent images. To address the unique characteristics of patent images (discussed in Section 1.1)., we systematically evaluated different normalization parameters with mean and standard deviations from ImageNet [5], Inception [24], and our own TRAIN split. We achieved the best results with ImageNet normalization parameters.

Visualizing cross-modal alignment of CLIP

After adapting CLIP to the patent domain, we encoded patent images and their corresponding figure descriptions from QUERY split using the adapted CLIP model and performed dimensionality reduction of the embeddings using Uniform Manifold Approximation and Projection (UMAP) [17] to visualize them in a 2D-plane. Figure 5 shows the visualization, where the green circles represent the images, and the blue triangles represent the corresponding figure descriptions in the same projection space. We compare the image-text alignment before (left) and after adaption of CLIP (right).

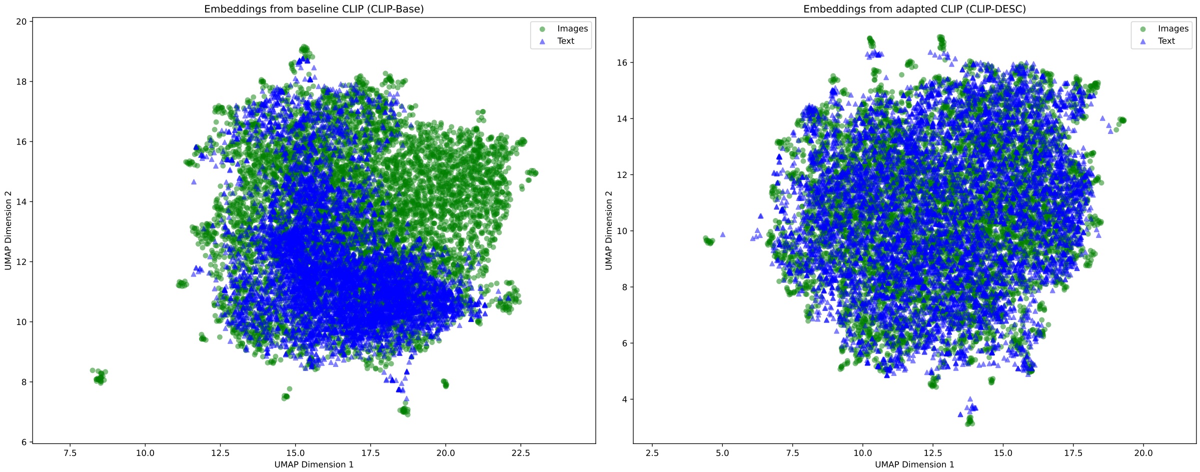


Figure 5: UMAP visualization of CLIP embeddings before and after adaptation to patent domain using figure descriptions for QUERY patents. **Left:** Embeddings from baseline CLIP model. **Right:** Embeddings from adapted CLIP-DESC model. Green circles represent image embeddings, blue triangles represent text embeddings.

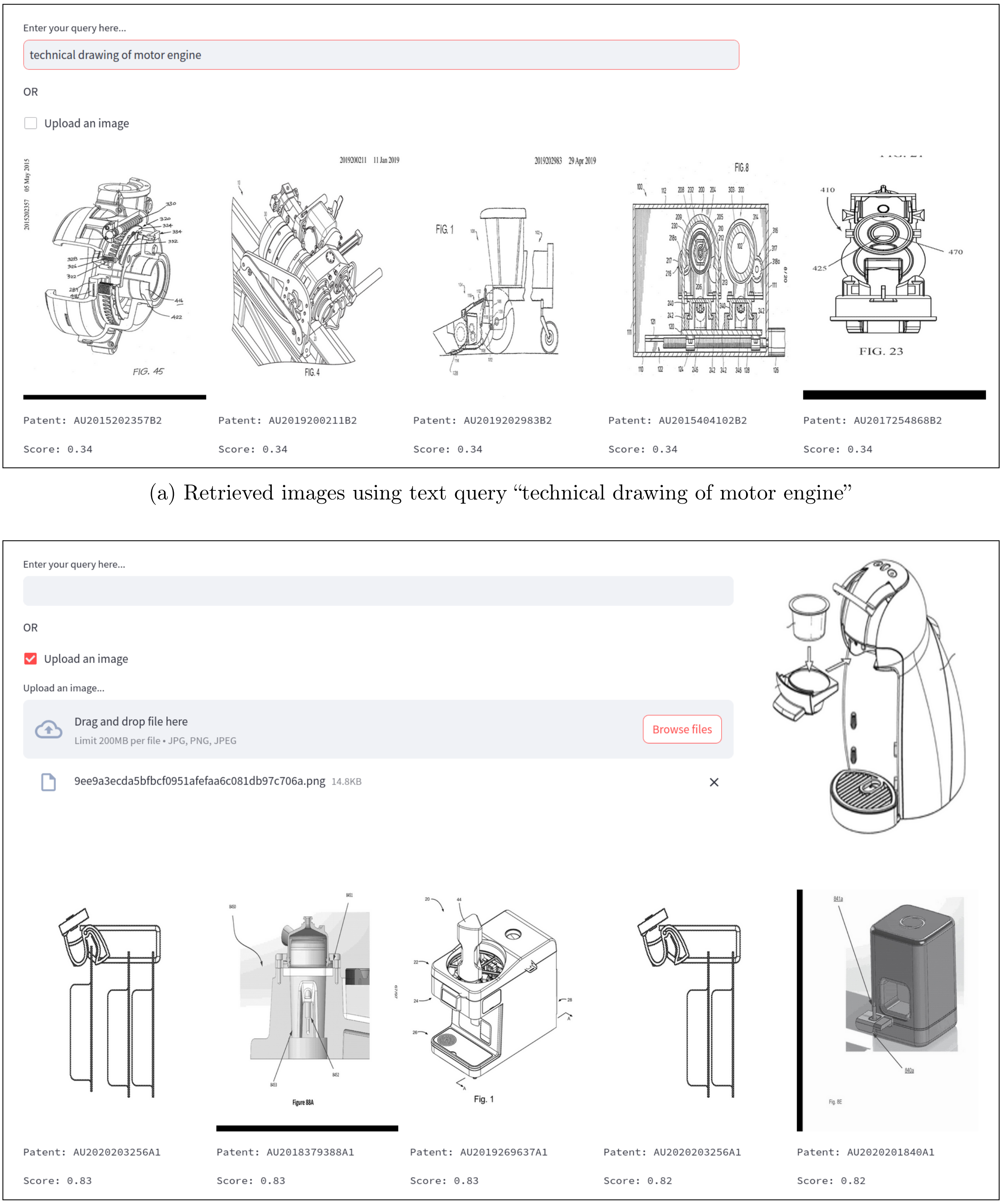
2.4 WP 4: Multimodal Pre-search and Classification of Patents

The main objective of WP 4 was to perform downstream tasks such as retrieval and classification using multimodal patent representation models. For the retrieval task (Section 2.4.1), the focus was on developing methods to effectively search and retrieve relevant patent images based on textual information (i.e., cross-modal text-to-image retrieval), visual information (i.e., image-to-image retrieval), or both. In terms of classification (Section 2.4.2), the work package explored the development of patent figure classification techniques, which could be leveraged to facilitate faceted search in large patent databases. The figure-based classification allows users to filter and navigate large patent collections using figures, complementing traditional text-based search methods. For example, users could narrow down their search results by selecting specific figures based on figure type, patent classification code, object-depicted or figure perspective viewpoint, greatly enhancing the efficiency of patent search systems.

2.4.1 Visual and Multimodal Patent Search

The main task in this work involved building visual and multimodal patent search systems that leverages the multimodal patent representations models developed in WP 3 (see Section 2.3). These models allow for various image search scenarios including cross-modal retrieval based on a textual description, image-to-image retrieval based on a query image, and the combination of both. For this purpose, we first created the embeddings of the patent figures in the INDEX split of the *EPO Utility Patent Corpus* using the CLIP model from WP 3. These embeddings were stored in a FAISS (Facebook Artificial Intelligence Similarity Search) [6] index, which is an open-source vector database developed by *Meta*, for efficient retrieval. Given a query image, description, we computed the embeddings using the same model and compared them to the embeddings of the images in the INDEX. For this purpose, we used the cosine similarity as the distance metric based on the *FlatIndex*, which is a plain vector store provided by FAISS, with vector dimension of 512 that corresponds to CLIP's output dimension. Finally, we sorted the images based on their similarity to create a final ranking.

To showcase the retrieval system, we developed a *Streamlit* application. First, we created a backend that implements the retrieval approach described above. Furthermore, we provided a simple interface (shown in Figure 6) that allows users to input a text or image for text-to-image and image-to-image retrieval, respectively. The image is passed to the backend which returns the final ranking to the user. The code and models for the demo was shared with EPO and served as basis for EPO's prototype. We plan to extend this demo with additional features such as user-driven re-ranking and cluster-based visual analytics as further discussed in Section 4 and Section 5.



Enter your query here...

technical drawing of motor engine

OR

□ Upload an image

Patent: AU2015202357B2
Score: 0.34

Patent: AU2019200211B2
Score: 0.34

Patent: AU2019202983B2
Score: 0.34

Patent: AU2015404102B2
Score: 0.34

Patent: AU2017254868B2
Score: 0.34

(a) Retrieved images using text query "technical drawing of motor engine"

Enter your query here...

OR

☑ Upload an image

Upload an image...

Drag and drop file here
Limit 200MB per file • JPG, PNG, JPEG

9ee9a3ecda5bfbcf0951afefaa6c081db97c706a.png 14.8KB

Patent: AU2020203256A1
Score: 0.83

Patent: AU2018379388A1
Score: 0.83

Patent: AU2019269637A1
Score: 0.83

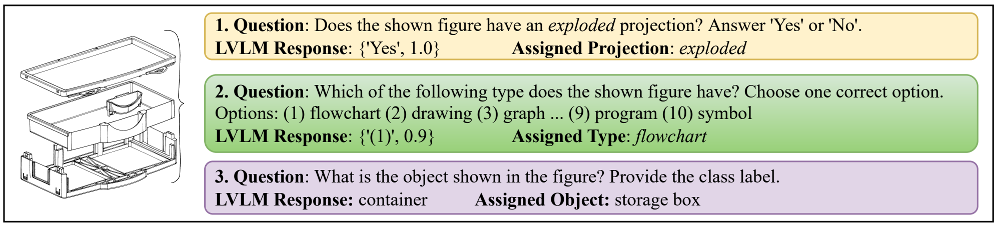
Patent: AU2020203256A1
Score: 0.82

Patent: AU2020201840A1
Score: 0.82

(b) Retrieved images using image query of coffee machine

Figure 6: Screenshot of the Streamlit application demo for (a) text-to-image retrieval and (b) image-to-image retrieval. The results show the retrieved images along with patent id (Patent) and cosine similarity score (Score).

2.4.2 Patent Figure Classification using Large Vision-language Model



1. Question: Does the shown figure have an *exploded* projection? Answer 'Yes' or 'No'.
LVLM Response: {'Yes', 1.0} **Assigned Projection**: *exploded*

2. Question: Which of the following type does the shown figure have? Choose one correct option.
Options: (1) flowchart (2) drawing (3) graph ... (9) program (10) symbol
LVLM Response: {'(1)', 0.9} **Assigned Type**: *flowchart*

3. Question: What is the object shown in the figure? Provide the class label.
LVLM Response: container **Assigned Object:** storage box

Figure 7: A figure from patent USD534354S1 showing a *drawing* of a *modular tool storage drawer* in *cross-sectional projection*, and three different questions asking about various aspects of the figure: 1) Binary question asking about projection 2) Multiple-choice question asking about figure type, and 3) Open-ended question asking about object depicted. The figure also shows the response (and token probability) generated from an LVLM and the corresponding concept assigned to the figure.

In this task, we explored the use of LVLMs for automating patent figure classification (shown in Figure 7), a task that is critical for improving efficiency in patent retrieval systems. We focused on adapting LVLMs to the unique characteristics of patent figures, which differ significantly from natural images, and demonstrate their effectiveness in handling complex classification tasks within the patent domain.

We explored figure classification of both utility and design patents, whereby we classify figures by TYPE (e.g., flowchart or technical drawing), PROJECTION (e.g., exploded or cross-sectional views), OBJECT (e.g., toothbrush or suitcase), and USPC (United States Patent Classification Scheme).

For this purpose, we utilized two publicly available patent figure classification datasets:

1. **Extended CLEF-IP 2011** [7], which contains 35,926 utility patent figures classified into 10 different figure TYPES, which include *block_or_circuit*, *chemical*, *drawing*, *flowchart*, *genesequence*, *graph*, *maths*, *program*, *symbol*, and *table* and fall into three IPC subclasses: A43B, A61B, and H10L.

2. **DeepPatent2** [1], which comprises of 2,785,762 segmented industrial design patent figures covering 22,394 unique PROJECTIONS, 132,890 unique OBJECTS, and 33 USPC classes.

We combined and adapted the above two datasets to produce two novel datasets that are suitable for training and evaluation of LVLM for patent figure classification. The first dataset, PATFIGVQA, is designed for visual question answering and enables finetuning LVLMs in few-shot learning scenarios. The second dataset, PATFIGCLS, supports classification tasks across multiple aspects: TYPE, PROJECTION, OBJECT, and USPC class.

To address the challenge of classifying figures across large label sets, such as the 1,400+ object classes, we developed a tournament-style classification strategy (MC-TS). This novel approach uses iterative multiple-choice questioning to narrow down labels efficiently, significantly reducing computational costs compared to binary classification method. Here, we experimented with 5: MC-TS (5), 10: MC-TS (10) and 20: MC-TS (20) options per question. Our methodology (shown in Figure 8) involved adapting INSTRUCTBLIP [4] to patent figures through finetuning, which allowed us to bridge the domain gap between natural images and technical patent figures. We tested various classification approaches, including binary classification (BC), open-ended classification (OC), and our tournament-style method. A BC approach uses multiple binary choice questions to narrow down to a single classification label, and an OC approach uses

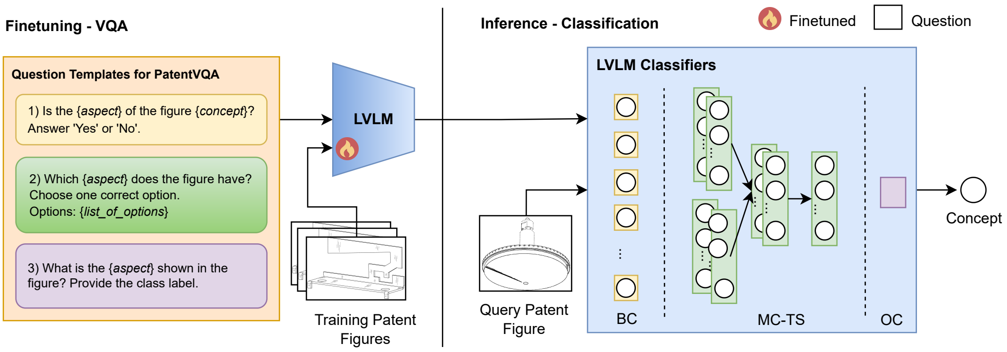

Finetuning - VQA
Question Templates for PatentVQA
1) Is the {aspect} of the figure {concept}? Answer 'Yes' or 'No'.
2) Which {aspect} does the figure have? Choose one correct option. Options: {list_of_options}
3) What is the {aspect} shown in the figure? Provide the class label.
LVLM
Training Patent Figures
Inference - Classification
Finetuned
Question
LVLM Classifiers
Query Patent Figure
BC
MC-TS
OC
Concept

Figure 8: Workflow of patent figure classification using different LVLM-based classification approaches. On the left, question templates for different question types used to create PatFigVQA dataset are shown. On the right, three different approaches to figure classification using a fine-tuned LVLM is shown, which include Binary Classification [BC], Multiple-choice Classification - Tournament-style Strategy [MC-TS], and Open-ended Classification [OC].

open-ended questions together with a mapping function that maps the open-ended answer to a classification label.

The evaluation framework included both zero-shot and few-shot scenarios, demonstrating the adaptability of LVLMs even with limited training data. We used Top-1 accuracy and SemEq metrics for evaluation. The SemEq metric uses LVLM as a judge to evaluate if two classification labels are semantically equivalent or not. The results of the experiments are shown in Table 7, which highlight the effectiveness of LVLMs in understanding and categorizing patent figures. We found that LVLM-based methods outperformed supervised traditional CNN-based classifiers for figure-type and USPC classification tasks while achieving comparable results for projection and object classification. This work was accepted as a full conference paper in the 47^{th} European Conference on Information Retrieval (ECIR) and will be presented on April 7, 2025 in Lucca, Italy (see Section 5). More details can be found in our publication (Awale et al. [2]).

2.5 WP 5: Efficient Models and Evaluation at Scale

In this work package, we considered two main tasks. The first task involved developing efficient and lean algorithms for large-scale multimodal patent retrieval, including collaboration with the EPO for “stress tests” to ensure system robustness and scalability. The second task focused on evaluating the multimodal patent representation models developed in WP 3 on a patent retrieval task using prevalent relevance data (Section 2.5.1). Our efforts primarily centered on the second task, where we evaluated the fine-tuned CLIP models on an image-based patent retrieval task using citation-based relevance judgments. This evaluation provided valuable insights into the limitations of using citation-based relevance judgement and highlighted the need for a new image-to-image retrieval benchmark. Hence, we prioritized creating this benchmark over the first task to provide a more accurate and relevant means of assessing multimodal patent retrieval system (Section 2.5.2).

2.5.1 Patent Retrieval Evaluation

In this task, we evaluated the multimodal patent representation model developed in WP 3 (Section 2.3) on multiple patent retrieval benchmarks. We formulated the evaluation task as an image-to-image retrieval task i.e. for a given query patent image, retrieve the most similar images from a database of patent images. The task involved encoding images into feature vectors, measuring similarity using distance metrics, and efficiently retrieving matches with indexing

Table 7: Top-1 accuracy and SemEq of supervised and few-shot models and different LVLM-based figure classification approaches (BC, MC_TS, OC) using InstructBLIP [4] in zero-shot and few-shot settings (n denotes the maximum number of training samples per class) for Type, Projection (Proj.), USPC, Objects.

Approach	**n**	**Type (10)**	**Proj. (7)**	**USPC (32)**		**Objects (1,447)**	
		Top-1	Top-1	Top-1	SemEq	Top-1	SemEq
Supervised (All results from [7])							
ResNet50	all	81.60	-				
ResNext101	all	85.01	-				
CLIP+MLP	all	82.44	-				
Zero-shot							
BC	-	46.44	14.50	7.80	10.60	1.38	24.40
MC-TS (5)	-	73.27	18.00	18.10	25.90	6.77	29.85
MC-TS (10)	-	57.69	14.90	17.80	25.70	6.36	30.82
MC-TS (20)	-	-	-	12.70	19.50	2.83	22.11
OC	-	30.96	11.60	5.10	13.20	5.18	25.57
Few-shot							
ResNet50	150	77.40	32.80	13.50	18.30	29.58	37.80
ResNext101	150	83.07	**38.80**	16.80	21.30	**47.96**	**56.25**
BC	150	67.31	16.40	15.50	17.80	6.70	18.80
MC-TS (5)	150	**87.98**	24.30	25.20	29.40	17.62	33.31
MC-TS (10)	150	87.12	23.90	**26.60**	**30.70**	17.00	33.10
MC-TS (20)	150	-	-	25.60	29.40	15.83	33.59
OC	150	87.31	34.60	18.90	21.90	18.24	42.23

tools such as FAISS [6]. We encoded all the patent images using each of our adapted CLIP models: CLIP-DESC, CLIP-TERMS, CLIP-GenDESC, and CLIP-GenTERMS. Finally, to determine which of the retrieved images are relevant to the query image, we utilized the corresponding ground truth datasets, which allowed us to calculate standard information retrieval metrics such as *mean Average Precision* (mAP).

Evaluation Benchmarks

We evaluated the adapted CLIP models on two patent retrieval benchmarks: (1) *EPO Utility Patent Corpus* and (2) *DeepPatent* [13], which use different forms of ground truth data to determine relevance between image pairs.

EPO Utility Patent Corpus The *EPO Utility Patent Corpus* (Section 2.3.1) uses citation-based relevance ground truth. In the patent domain, a citation refers to a documented reference in a patent search report where a citing patent A identifies a cited patent B as prior art. This relationship implies that the claims of patent A may be invalidated by patent B if the former demonstrates a lack of novelty or inventive step (obviousness of the invention). For example, if cited patent B predates patent A and discloses similar technical elements, it could negate patent A's novelty. A single citing patent can have invalidation risks from multiple cited patents, as each claim is independently assessed against prior art. Figure 9 illustrates the number of cited patents in the INDEX split for each citing patent in the QUERY split.

For *EPO Utility Patent Corpus*, we used patent images from QUERY split as the query image and the patent images from INDEX as the database of patent images to retrieve results

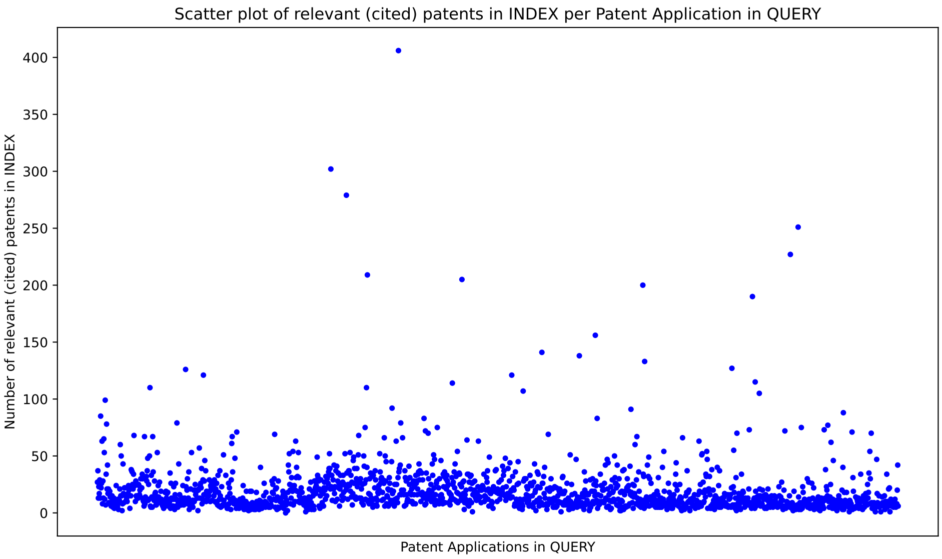


Figure 9: Scatter plot showing number of invalidating patents in INDEX of corpus per patent application in QUERY

from. While this process produces a ranking of patent images, the citation-based relevance ground-truth data require a ranking of patents. To bridge this gap, we implemented a mapping function to convert the image ranking into patent ranking. We employed the MAX aggregation method, which selects the highest similarity score between patent image pairs from two patents. To illustrate this aggregation method, consider a scenario where a query patent Q contains $n_Q = 3$ figures and a retrieved patent P has $n_P = 4$ figures. We compute $n_Q \times n_P = 12$ similarity scores for all possible figure combinations and select the highest score as the overall similarity measure between patents Q and P. This approach allowed us to transform the image-based ranking into a patent-based ranking suitable for our evaluation purposes.

DeepPatent The *DeepPatent* [13] dataset uses Re-ID (Re-identification) as the relevance ground truth, i.e., two patent images belonging to the same patent are considered to be relevant. The dataset comprises patent images from design patents, containing over 350,000 public domain patent DRAWINGS sourced from the United States Patent and Trademark Office (USPTO). Patent drawings within a patent are highly likely to be visually similar, which is more appropriate to evaluate image-to-image patent retrieval benchmark.

The results of our evaluation experiments are presented below.

Results

For evaluation, we computed the *mean Average Precision* (mAP@k) on the top-k results of the respective evaluation datasets, choosing the following values for k: 1, 10, 80, 500, 1000, and 2000. Table 8 presents the performance of baseline CLIP (CLIP-BASE) in comparison with the adapted CLIP models (CLIP-DESC, CLIP-TERMS, CLIP-GENDESC, CLIP-GENTERMS) on *EPO Utility Patent Corpus* and *DeepPatent* benchmarks.

Table 8 indicates that adapting the CLIP model using language-supervision for image-based patent retrieval tasks yielded modest performance improvements for *EPO Utility Patent Corpus*. However, the adapted models on the *DeepPatent* dataset demonstrated substantial performance gains over the baseline CLIP modelCLIP-BASE, despite not being specifically trained on the DeepPatent dataset. These results suggest that our pre-training approach effectively enhances the model's ability to handle patent images.

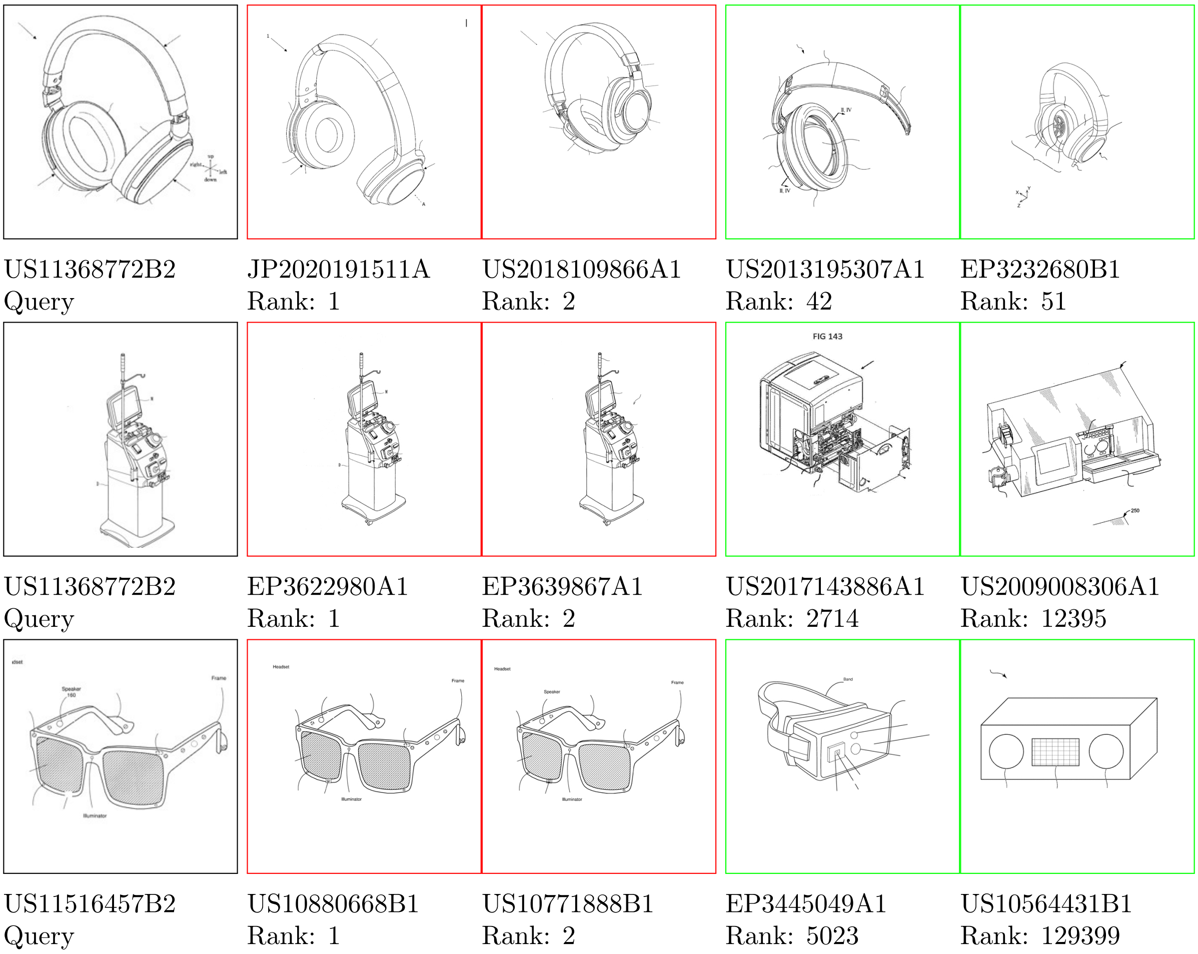
Table 8: Results for image-based patent retrieval using mAP@k for $k = 1, 10, 80, 500, 1000, 2000$ of the baseline CLIP (CLIP-base) in comparison with the adapted CLIP models (CLIP-DESC, CLIP-TERMS, CLIP-GenDESC, CLIP-GenTERMS) on *EPO Utility Patent Corpus* and *DeepPatent* benchmarks.

Text	mAP@1	mAP@10	mAP@80	mAP@500	mAP@1000	mAP@2000
Results for EPO Utility Patent Corpus						
CLIP-base	0.20468	0.22078	0.20764	0.18075	0.16821	0.15383
CLIP-DESC	**0.21336**	**0.23251**	**0.21202**	**0.18730**	**0.17404**	**0.15470**
CLIP-TERMS	**0.21336**	**0.23251**	**0.21202**	**0.18730**	**0.17404**	**0.15470**
CLIP-GenDESC	**0.21336**	0.23188	0.21464	0.18566	0.17098	0.15179
CLIP-GenTERMS	**0.21336**	0.23188	0.21464	0.18566	0.17098	0.15179
Results for DeepPatent dataset						
CLIP-base	0.56979	0.58448	0.49603	0.41091	0.38045	0.35062
CLIP-DESC	**0.67151**	**0.66748**	**0.56736**	**0.47628**	**0.44384**	**0.41286**
CLIP-TERMS	0.66352	0.65888	0.5592	0.46724	0.43632	0.40632
CLIP-GenDESC	0.58395	0.59388	0.50134	0.41100	0.38212	0.35377
CLIP-GenTERMS	0.59141	0.60147	0.50462	0.41668	0.38603	0.35862

Interestingly, the mAP values on the *DeepPatent* dataset are significantly higher than that for the *EPO Utility Patent Corpus*. We believe that this discrepancy is due to the misalignment between image-based patent retrieval task and the citation-based relevance judgement, i.e., patents judged to be relevant based on citations do not necessarily account for relevancy between images. In the patent domain, the citation-based relevance judgement is primarily based on text-based patent retrieval systems. Hence, the retrieved patents have a bias towards those with similar text. Further, the patent examiners are not necessarily required to consider the patent images during relevance judgement. As a result, a citing patent and a cited patent may be relevant, but the images between them may not necessarily be similar. Further to exacerbate the problem, patent examination is a time-sensitive process, and examiners are not incentivized to find multiple relevant patents. They typically stop searching once they find sufficient prior art to make a decision on patentability. This leaves a large number of patents judged to be not relevant (relevance judgement hole). In information retrieval, **relevance judgement holes** is a well-known concept, which refers to a gap or missing relevance judgement data between query and retrieved pairs.

As a result, while our developed multimodal patent retrieval model successfully retrieved visually similar patent images, the quantitative evaluation metrics based on citations did not accurately reflect the true performance of these models. On the other hand, on *DeepPatent* dataset, which accounts for image-to-image similarity, the models perform relatively well. Although *DeepPatent* dataset is suitable for our evaluation purposes, the retrieval task does not reflect the need of patent examiners, i.e., retrieving images from the same patent. These misalignments in the evaluation process highlight the need for more appropriate evaluation benchmark for image-based patent retrieval tasks.

Figure 10 presents qualitative results that support our hypothesis, demonstrating instances where visually similar patent images were retrieved but not necessarily reflected in the ground-truth labels derived from citation-based relevance judgments. This visual evidence underscores the limitations of relying solely on citation-based metrics for evaluating image-based patent retrieval systems.



US11368772B2 Query
JP2020191511A Rank: 1
US2018109866A1 Rank: 2
US2013195307A1 Rank: 42
EP3232680B1 Rank: 51

US11368772B2 Query
EP3622980A1 Rank: 1
EP3639867A1 Rank: 2
US2017143886A1 Rank: 2714
US2009008306A1 Rank: 12395

US11516457B2 Query
US10880668B1 Rank: 1
US10771888B1 Rank: 2
EP3445049A1 Rank: 5023
US10564431B1 Rank: 129399

Figure 10: Retrieved images and their rank that are labeled as relevant (green border) and irrelevant (red border) according to the *EPO Utility Patent Corpus* for two query images (black border). The qualitative results demonstrate that automatically curated citation-based relevance judgements can be (1) incomplete as images that are considered relevant by humans (retrieved within the top-2 in both examples) but are not labeled as such; and (2) inaccurate as the images itself can be irrelevant and are ranked lower according to human judgments while the patent text is related. The retrieved images are using CLIP-DESC model.

2.5.2 Patent Image-to-Image Retrieval Benchmark

The findings presented in the previous section underscored a critical gap and a need for a new patent retrieval benchmark that incorporates image-based relevance judgments. To address this gap, our goal was to build a high-quality patent image retrieval benchmark based on human annotations. To gain insights into the task's requirements and potential challenges, we initiated our efforts with a small-scale pilot study. This pilot study served as a foundation for understanding the requirements of image-based patent relevance assessment and developing a sophisticated strategies to create a challenging and comprehensive benchmark for patent image search.

Pilot Study

Sample Selection In our pilot study, we manually curated 45 query images, selecting one representative image from five randomly selected patents evenly divided across nine CPC sections from the QUERY split. The selection ensured coverage of diverse technical domains. For each query image, we applied the vanilla CLIP model (CLIP-BASE), pre-trained on natural images, to identify two images from the INDEX split. The selection process focused on finding the two most visually similar images to the query with an important distinction: one image was sourced from a *cited patent* to the query patent, while the other came from a *non-cited patent.* As a result, we retrieved 90 image pairs for annotation.

Annotation To facilitate the annotation process, we developed a custom tool using Streamlit, enabling efficient evaluation across three critical dimensions: visual relevance, semantic relevance, and subpart relevance. One of the project members manually assessed each image pair by answering three specific questions, each with three possible responses (Yes, No, Unsure):

1. Are the images the same or visually similar?
2. Do the images depict the same or very similar inventions/concepts?
3. Do the images illustrate related components or subsystems of a larger invention?

First, these questions were answered only based on *visual* information from the image pair. However, since reliably answering these questions might require further context, we also provided additional, *multimodal* metadata such as patent title, figure description, etc. The statistics can be found in the Table 9. The statistics show that the more image pairs between query and non-cited patents were deemed to be visually and semantically similar than the image pairs between query and cited patents.

Table 9: Distribution of annotation responses for different relevancy questions between *cited patents* and *non-cited patents.* The responses were provided using only the visual information of the image pair as well as with additional textual information from the patent (denoted as multimodal).

Mode	Question	Response	Cited Patents	Non-cited Patents	Total
Visual	Q1	Yes	10	30	40
		No	34	15	49
		Unsure	1	0	1
	Q2	Yes	4	14	18
		No	27	5	32
		Unsure	14	26	40
	Q3	Yes	4	14	18
		No	26	5	31
		Unsure	15	26	41
Multimodal	Q1	Yes	11	30	41
		No	34	15	49
		Unsure	0	0	0
	Q2	Yes	18	20	38
		No	18	24	42
		Unsure	9	1	10
	Q3	Yes	19	20	29
		No	18	24	42
		Unsure	8	1	9

Results of the Pilot Study Finally, we built a novel image-to-image patent retrieval benchmark using the human-annotated patent image pairs. We consider image pairs with positive annotations (response: yes) as relevant. Similar to prior experiments, we indexed all the candidate patent images, and evaluated the baseline and adapted CLIP models on this new dataset using mAP@1 metric. We evaluate at only k=1 as we only have one relevant image in the index. Table 10 shows the results on the new benchmark dataset. The results show high mAP scores compared to the *EPO Utility Patent Corpus* and *DeepPatent* dataset benchmarks demonstrating that the baseline and our adapted CLIP models are well-suited to retrieving visually and semantically similar images. In particular, the CLIP models adapted using synthetic text captions (CLIP-GenDESC, CLIP-GenTERMS) also show strong performance in retrieving semantically similar images and images with subpart relationship.

Table 10: mAP@1 using baseline CLIP and pre-trained CLIP models on our patent image-to-image retrieval benchmark

Model	Visual Stage Annotations			Multimodal Stage Annotations		
	Q1	Q2	Q3	Q1	Q2	Q3
CLIP-base	0.85	0.86	0.86	0.82	0.5	0.48
CLIP-DESC	**0.88**	0.86	0.86	**0.85**	0.54	0.52
CLIP-TERMS	**0.88**	0.86	0.86	**0.85**	0.54	0.52
CLIP-GenDESC	0.85	**0.93**	**0.93**	0.82	**0.58**	**0.55**
CLIP-GenTERMS	0.85	**0.93**	**0.93**	0.82	**0.58**	**0.55**

Key Learnings The pilot study helped us understand some of the key requirements in the development of an image-based patent retrieval benchmark. We summarize our learnings from the pilot study as follows:

- **Circular Bias**: A circular bias was identified due to the use of the same image encoder for both sampling candidate images and evaluating their relevance. This overlap compromises the objectivity of the evaluation process, as the encoder inherently favors its own learned representations.
- **Relevance Judgement Hole**: The analysis of response statistics (Table 9) showed that positively annotated image pairs predominantly belonged to non-cited patents, indicating the presence of a relevance judgment hole problem (discussed in Section 2.5.1)
- **Annotation Guideline**: Requirement for a comprehensive and detailed annotation guideline was identified, where each of the relevance dimensions need to be defined specifically

Final Benchmark Annotation

With these insights from the pilot study, we revised our methodology to address the shortcomings and to create a more challenging dataset which requires avoiding the circular bias and identifying not only easy but also hard positives and negatives samples, which involved making the following changes:

Sample Mining We first made revisions to the process of sample mining. First, in order to avoid the circular bias that was prevalent in the dataset developed in the pilot study, we modified the sampling technique to use multiple image retrieval models instead of using only the CLIP model. For example, we used CLIP together with PatentNet [9], ResNext101 [28], which have different model architectures and number of model parameters, and were trained on different domains including the patent domain. The image encoders only capture visual relevance, so additionally, we also used text retrieval models to incorporate semantic relevance. The text retrieval models are based on retrieving patent images using different text metadata such as figure descriptions, component terms and patent titles. Using text retrieval and image retrieval models, we formulated four distinct groups of samples:

- **Group 1: Easy Positives** - This group includes samples that were ranked high by both text and image retrieval models.
- **Group 2: Hard Positives** - This group includes samples that were ranked high by text retrieval models but lower by image retrieval models.
- **Group 3: Hard Negatives** - This group includes samples that were ranked high by image retrieval models but lower by text retrieval models.
- **Group 2: Hard Positives** - This group includes samples that were ranked low by both text and image retrieval models.

Here, the high and low ranks were determined based on percentile of the retrieval scores. An advantage of having easy sample groups is that they do not require human annotations to annotate all the sample pairs. Moreover, formulating such positive and negative sample groups allows us to apply contrastive representation learning, which is in our future plans.

Annotation Guideline Realizing the need for a comprehensive annotation guidelines, we developed one with clear instructions and guidelines for annotation including example images. A draft of the new annotation guideline can be found in the Appendix A.

With these revisions, we then plan to create patent image retrieval benchmark annotated by multiple humans across multiple key dimensions to facilitate a comprehensive evaluation of patent representation models. The modifications are planned to be completed by end of May 2025 and we intend to share the annotated dataset with EPO.

2.6 WP 6: Project Management, Coordination and Dissemination

Project Management and Coordination

Management and coordination of the project was realized mainly through the regular bi-weekly online meetings between TIB and EPO (supervisors: Rahim Delaviz and Alexander Klenner-Bajaja). In these bi-weekly meetings, TIB reported on the progress status regarding accomplishments, challenges and other obstacles to synchronize the work with the project objectives. To keep track of the project's progress, a mid-term interim report was prepared and shared with all stakeholders. This report served as a comprehensive overview of the project's status, highlighting key milestones achieved, challenges encountered, and plans for the remaining project duration.

Presentations

We presented ViP@Scale at multiple workshops.

1. A presentation on the project was given on July 12, 2023 at the Annual European Patent Office Academic Research Programme Workshop 2023. The workshop proved to be highly productive, with valuable feedback and suggestions received during the question and answer session. Our discussant, Anna Chatzimichali from the University of the West of England, provided particularly insightful feedback. Her recommendation to develop the patent retrieval system with the end user in mind was especially valuable and has influenced our approach to the project.

2. We were invited by Dr. Hidir Aras (FIZ Karlsruhe) to present the project at the 1st workshop of the DFG-funded research project *Patents4Science* (`https://www.patents4science.org/`) in Berlin on October 5, 2023. This workshop was organized by project members of the FIZ Karlsruhe (Leibniz Institute for Information Infrastructure), INP Greifswald (Leibniz Institute for Plasma Science and Technology), IWT Bremen (Leibniz Institute for Materials-oriented Technologies), and INM Saarbrücken (Leibniz Institute for New Materials) and provided a good opportunity to promote the project, exchange ideas and use cases, as well as discuss future collaborations.

3. Ralph Ewerth is invited to give a keynote talk at the SIGIR Workshop PatentSemTech 2025.

4. We will present our work on *Patent Figure Classification using Large Vision-language Models* [2] presented in Section 2.4.2 at the European Conference on Information Retrieval in Lucca, Italy on April 7, 2025.

Dataset, Models, and Code

All resources including datasets, models, and source code developed within the project were shared with EPO directly. For this purpose, *Google Cloud Platform*[2] was used to facilitate the exchange among team members. The shared models were deployed by EPO during the project to build a demo of the image search system using *Streamlit*. Furthermore, datasets, source code, and models used in our published and planned research papers (see Section 5) will be made publicly available.

3 Use of funds

We confirm that the funds for ViP@Scale were used as requested. A detailed financial report is attached to this project report, providing a breakdown of all expenditures.

4 Summary & Future Work

In this project, we focused on researching and developing advanced multimodal representation models specifically tailored to utility patents. A key aspect of our work involved exploring practical applications of these multimodal models in patent-related downstream tasks: patent retrieval and figure classification. To validate the effectiveness of our multimodal approaches, we conducted comprehensive evaluations using a large-scale patent retrieval benchmark, which provided valuable insights into the performance of our models and helped identify areas for further improvement.

For the development of our multimodal patent representation model, we performed domain adaptation of the CLIP vision-language model—originally pretrained on natural images—to utility patents by leveraging both extracted patent texts and synthetic descriptions generated by LVLMs for language-supervised alignment. This approach combined patent-specific textual data with patent figures to enhance cross-modal alignment and create more robust multimodal patent representations. Utilizing this developed model, we implemented a patent retrieval system capable of facilitating both text-to-image and image-to-text retrieval. Additionally, we explored the classification of patent figures using LVLM-based classification methods. This classification approach enables faceted patent search, significantly aiding in large-scale patent search tasks by providing more nuanced and targeted search options.

[2] `https://cloud.google.com/`

Finally, to assess the efficacy of our multimodal approaches, we conducted thorough evaluations using a citation-based patent retrieval benchmark. This evaluation process yielded valuable insights, particularly regarding the limitations of image-based patent retrieval methods. Our findings highlighted a critical gap in the field: the absence of a suitable benchmark for evaluating image-to-image retrieval in the patent domain. This realization underscored the need for a human-annotated image-to-image retrieval benchmark, which would provide a more accurate framework for evaluating multimodal patent retrieval systems.

Looking ahead, our future work will focus on advancing the multimodal patent representation model through contrastive representation learning to better align visual and semantic similarities in the projection space. This task will entail (1) developing sample mining techniques to formulate image pairs with positive and negative pairings suitable for contrastive representation learning, and (2) creating a high-quality image-to-image patent retrieval benchmark based on human annotations based on several relevance judgements that are specifically designed for typical use cases in patent retrieval task. We also plan to augment the patent retrieval system by incorporating interactive components such as user-driven re-ranking mechanisms and cluster-based visual analytics, while integrating LVLM-powered classification techniques directly into the retrieval pipeline. These ongoing developments—already in active progress—aim to establish new standards for technical patent analysis, with planned dissemination through major conferences and workshops in multimodal and information retrieval.

5 Publications

We have worked on three papers during the course of the project. One paper has been accepted so far, two further papers (one research and one demo paper) are currently being prepared and we plan to submit them for review in spring 2025.

Published Articles

[2] Sushil Awale, Eric Müller-Budack, and Ralph Ewerth. “Patent Figure Classification using Large Vision-language Models”. In: *European Conference on Information Retrieval, ECIR 2025, Luuca, Italy, April 05-11, 2025*. Lecture Notes in Computer Science. Springer, 2025. DOI: 10.48550/ARXIV.2501.12751.

This paper present a novel approach using large vision-language models along with two datasets for patent visual questions answering classification (see WP 4, Section 2.4.2).

Abstract: Patent figure classification facilitates faceted search in patent retrieval systems, enabling efficient prior art search. Existing approaches have explored patent figure classification for only a single aspect and for aspects with a limited number of concepts. In recent years, large vision-language models (LVLMs) have shown tremendous performance across numerous computer vision downstream tasks, however, they remain unexplored for patent figure classification. Our work explores the efficacy of LVLMs in patent figure visual question answering (VQA) and classification, focusing on zero-shot and few-shot learning scenarios. For this purpose, we introduce new datasets, PatFigVQA and PatFigCLS, for fine-tuning and evaluation regarding multiple aspects of patent figures (i.e., type, projection, patent class, and objects). For a computational-effective handling of a large number of classes using LVLM, we propose a novel tournament-style classification strategy that leverages a series of multiple-choice questions. Experimental results and comparisons of multiple classification approaches based on LVLMs and Convolutional Neural Networks (CNNs) in few-shot settings show the feasibility of the proposed approaches.

Source Code: https://github.com/TIBHannover/patent-figure-classification

Planned Articles

Based on the work and future work presented in this report, we are planning to submit two papers at renowned conferences (or workshops), with authorship shared between our team and the EPO.

Preliminary title: "iPatent - Interactive Patent Retrieval and Clustering" We plan to submit a demo paper (Section 2.4.1) including the extensions described in the future work (Section 4) for interactive patent retrieval. We aim to submit this work to the demo track of ACM Multimedia or the Patent Text Mining and Semantic Technologies (PatentSemTech) workshop co-located with ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR).

Preliminary title: "Multimodal Patent Retrieval Benchmark Dataset" In this paper, we will present our approaches for multimodal representation learning including a novel high-quality benchmark to reliably evaluate models for for image retrieval in patents. For this purpose, we will finalize the corpus and train several models as described in the future work (Section 4). Depending on the results and contributions, we aim to submit this work either at the research track of International Conference on Information and Knowledge Management (CIKM 2025) or as a resource paper at the ACM Multimedia Conference (ACMMM 2025).

References

[1] Kehinde Ajayi, Xin Wei, Martin Gryder, Winston Shields, Jian Wu, Shawn M. Jones, Michal Kucer, and Diane Oyen. "DeepPatent2: A Large-Scale Benchmarking Corpus for Technical Drawing Understanding". In: *Scientific Data* 10.1 (Nov. 2023), p. 772. ISSN: 2052-4463. DOI: 10.1038/s41597-023-02653-7.

[2] Sushil Awale, Eric Müller-Budack, and Ralph Ewerth. "Patent Figure Classification using Large Vision-language Models". In: *European Conference on Information Retrieval, ECIR 2025, Luuca, Italy, April 05-11, 2025.* Lecture Notes in Computer Science. Springer, 2025. DOI: 10.48550/ARXIV.2501.12751.

[3] Gabriela Csurka, Jean-Michel Renders, and Guillaume Jacquet. "XRCE's Participation at Patent Image Classification and Image-based Patent Retrieval Tasks of the Clef-IP 2011". In: *CLEF 2011 Labs and Workshop, 2011, Amsterdam, The Netherlands, Septemeber 19-22, 2011.* CEUR-WS.org, 2011. URL: https://ceur-ws.org/Vol-1177/CLEF2011wn-CLEF-IP-CsurkaEt2011.pdf.

[4] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. "InstructBLIP: Toward and relative positios General-purpose Vision-Language Models with Instruction Tuning". In: *Advances in Neural Information Processing Systems, NeurIPS 2023, New Orleans, LA, USA, December 10-16, 2023.* 2023. URL: http://papers.nips.cc/paper%5C_files/paper/2023/hash/9a6a435e75419a836fe47ab6793623e6-Abstract-Conference.html.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "ImageNet: A large-scale hierarchical image database". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2009, Miami, Florida, USA, June 20-25, 2009.* IEEE Computer Society, 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[6] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. "The Faiss library". In: (2024). arXiv: 2401.08281 [cs.LG].

[7] Junaid Ahmed Ghauri, Eric Müller-Budack, and Ralph Ewerth. "Classification of Visualization Types and Perspectives in Patents". In: *International Conference on Theory and Practice of Digital Libraries, TPDL 2023, Zadar, Croatia, September 26-29, 2023.* Springer, 2023, pp. 182–191. DOI: 10.1007/978-3-031-43849-3_16.

[8] Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. "Finetune like you pretrain: Improved finetuning of zero-shot vision models". In: *Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023.* IEEE, 2023, pp. 19338–19347. DOI: 10.1109/CVPR52729.2023.01853.

[9] Kotaro Higuchi and Keiji Yanai. "Patent image retrieval using transformer-based deep metric learning". In: *World Patent Information* 74 (2023), p. 102217. ISSN: 0172-2190. DOI: https://doi.org/10.1016/j.wpi.2023.102217.

[10] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. *OpenCLIP.* Version 0.1. If you use this software, please cite it as below. July 2021. DOI: 10.5281/zenodo.5143773.

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. "Segment Anything". In: *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023.* IEEE, 2023, pp. 3992–4003. DOI: 10.1109/ICCV51070.2023.00371.

[12] Ralf Krestel, Renukswamy Chikkamath, Christoph Hewel, and Julian Risch. "A survey on deep learning for patent analysis". In: *World Patent Information* 65 (2021), p. 102035. ISSN: 0172-2190. DOI: `https://doi.org/10.1016/j.wpi.2021.102035`.

[13] Michal Kucer, Diane Oyen, Juan Castorena, and Jian Wu. "DeepPatent: Large scale patent drawing recognition and retrieval". In: *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*. IEEE, 2022, pp. 557–566. DOI: `10.1109/WACV51458.2022.00063`.

[14] Yu-Hsun Lin, Min-Chian Hung, and Chen-Fan Lee. "Density-Refine: Patent Image Retrieval by Density-Based Region Extraction and Feature Fusion". In: *Journal of Mechanical Design* 147.8 (Feb. 2025), p. 081703. ISSN: 1050-0472. DOI: `10.1115/1.4067749`. eprint: `https://asmedigitalcollection.asme.org/mechanicaldesign/article-pdf/147/8/081703/7429671/md-24-1532.pdf`.

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. "Visual Instruction Tuning". In: *Advances in Neural Information Processing Systems, NeurIPS 2023, New Orleans, LA, USA, December 10-16, 2023*. 2023. URL: `http://papers.nips.cc/paper%5C_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html`.

[16] Hao-Cheng Lo, Jung-Mei Chu, Jieh Hsiang, and Chun-Chieh Cho. "Large Language Model Informed Patent Image Retrieval". In: *Proceedings of the 5th Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2024) co-located with the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024), Washington D.C., USA, July 28th, 2024*. Ed. by Hidir Aras, Ralf Krestel, Linda Andersson, Florina Piroi, Allan Hanbury, and Dean Alderucci. Vol. 3775. CEUR Workshop Proceedings. CEUR-WS.org, 2024, pp. 51–60.

[17] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. "UMAP: Uniform Manifold Approximation and Projection". In: *The Journal of Open Source Software* 3.29 (2018), p. 861.

[18] Florina Piroi, Mihai Lupu, Allan Hanbury, and Veronika Zenz. "CLEF-IP 2011: Retrieval in the Intellectual Property Domain". In: *CLEF 2011 Labs and Workshop, Notebook Papers, Amsterdam, The Netherlands, September 19-22, 2011*. Vol. 1177. CEUR Workshop Proceedings. CEUR-WS.org, 2011. URL: `https://ceur-ws.org/Vol-1177/CLEF2011wn-CLEF-IP-PiroiEt2011.pdf`.

[19] Kader Pustu-Iren, Gerrit Bruns, and Ralph Ewerth. "A multimodal approach for semantic patent image retrieval". In: *Patent Text Mining and Semantic Technologies co-located with ACM SIGIR Conference on Research and Development in Information Retrieval, PatentSemTech@SIGIR 2021, Aachen, Germany, July 15, 2021*. Aachen, Germany: RWTH Aachen, 2021. URL: `http://ifs.tuwien.ac.at/patentsemtech/2021/fls/2021/7%5C_Pustu-Iren.pdf`.

[20] 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: *International Conference on Machine Learning, ICML 2021, July 18-24, 2021, Virtual Event*. PMLR, 2021, pp. 8748–8763. URL: `http://proceedings.mlr.press/v139/radford21a.html`.

[21] Daniel Seebacher, Manuel Stein, Halldór Janetzko, and Daniel A. Keim. "Patent retrieval: a multi-modal visual analytics approach". In: *EuroVis Workshop on Visual Analytics*. SERIES16416. Groningen, The Netherlands: Eurographics Association, 2016, pp. 13–17. ISBN: 9783038680161.

[22] Homaira Huda Shomee, Zhu Wang, Sathya N. Ravi, and Sourav Medya. "IMPACT: A Large-scale Integrated Multimodal Patent Analysis and Creation Dataset for Design Patents". In: *Annual Conference on Neural Information Processing Systems, NeurIPS 2024, Vancouver, BC, Canada, December 10-15, 2024*. 2024. URL: `http://papers.nips.cc/paper%5C_files/paper/2024/hash/e3301977b92f28e32639ec99eb08f4a1-Abstract-Datasets%5C_and%5C_Benchmarks%5C_Track.html`.

[23] Panagiotis Sidiropoulos, Stefanos Vrochidis, and Ioannis Kompatsiaris. "Content-based binary image retrieval using the adaptive hierarchical density histogram". In: *Pattern Recognition* 44.4 (2011), pp. 739–750. DOI: `10.1016/J.PATCOG.2010.09.014`.

[24] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning". In: *Conference on Artificial Intelligence, AAAI 2017, San Francisco, California, USA, February 4-9, 2017*. Ed. by Satinder Singh and Shaul Markovitch. AAAI Press, 2017, pp. 4278–4284. DOI: `10.1609/AAAI.V31I1.11231`. URL: `https://doi.org/10.1609/aaai.v31i1.11231`.

[25] Avinash Tiwari and Veena Bansal. "PATSEEK: Content Based Image Retrieval System for Patent Database". In: *The Fourth International Conference on Electronic Business - Shaping Business Strategy in a Networked World*. Ed. by Jian Chen. Academic Publishers/World Publishing Corporation, 2004, pp. 1167–1171.

[26] Stefanos Vrochidis, Anastasia Moumtzidou, and Ioannis Kompatsiaris. "Concept-based patent image retrieval". In: *World Patent Information* 34.4 (2012), pp. 292–303. ISSN: 0172-2190. DOI: `https://doi.org/10.1016/j.wpi.2012.07.002`.

[27] Hongsong Wang and Yuqi Zhang. "Learning Efficient Representations for Image-Based Patent Retrieval". In: *Pattern Recognition and Computer Vision, PRCV 2023, Xiamen, China, October 13-15, 2023*. Ed. by Qingshan Liu, Hanzi Wang, Zhanyu Ma, Weishi Zheng, Hongbin Zha, Xilin Chen, Liang Wang, and Rongrong Ji. Vol. 14431. Lecture Notes in Computer Science. Springer, 2023, pp. 15–26. DOI: `10.1007/978-981-99-8540-1\_2`.

[28] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. "Aggregated Residual Transformations for Deep Neural Networks". In: *Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 5987–5995. DOI: `10.1109/CVPR.2017.634`.

[29] Johanna Zellmer and Christa Womser-Hacker. "Elicitation of requirements for innovative visual patent retrieval based on interviews with experts". In: *Information Research: an international electronic journal* 27 (Jan. 2022). DOI: `10.47989/irisic2234`.

[30] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. "Vision-Language Models for Vision Tasks: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.8 (2024), pp. 5625–5644. DOI: `10.1109/TPAMI.2024.3369699`.

A Patent Figure Similarity Annotation Guideline

Task Description

The objective of this annotation task is to evaluate the relevance between pairs of patent images across three key dimensions. Relevance, in the context of this annotation task, refers to the degree to which the characteristics of one patent image visually align with the characteristics of another patent image, relative to specified dimensions. Annotators should assess similarity across the following three dimensions: Visual, Semantic, and Subpart Relevance. During annotation, each dimension should be considered independently.

1. Visual Relevance

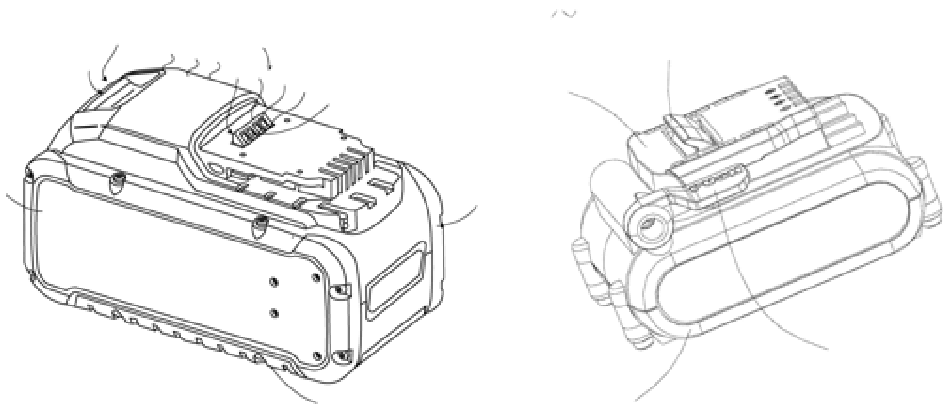
For visual relevance, we ask the question "How similar are the two shown images visually?" We assess the visual relevance of the two images based on the following points:

- Overall shape and structural match
- Layout and arrangement of components
- Level of detail and complexity
- Ignore leading lines and pointed arrows

Score	Condition
2 - Relevant	Nearly identical or very similar visual appearance
1 - Moderate Relevance	Moderately similar overall look
0 - No Relevance	Completely different visual appearance

Table 11: Score card for annotating visual relevance between shown patent image pairs

Guide: Would these drawings look similar if traced?
Example Here, the following two images are visually similar (2 - Relevant) as the overall shape and structure match, although are shown in different perspectives and are different inventions.



(a) Patent image showing *battery pack* (b) Patent image showing *power brick*

Figure 11: Example of an image pair for visual relevance annotation

2. Semantic Relevance

For semantic relevance, we ask the question “How semantically close are the concepts depicted in the shown two images?” We assess the semantic relevance of the two images based on the following points:

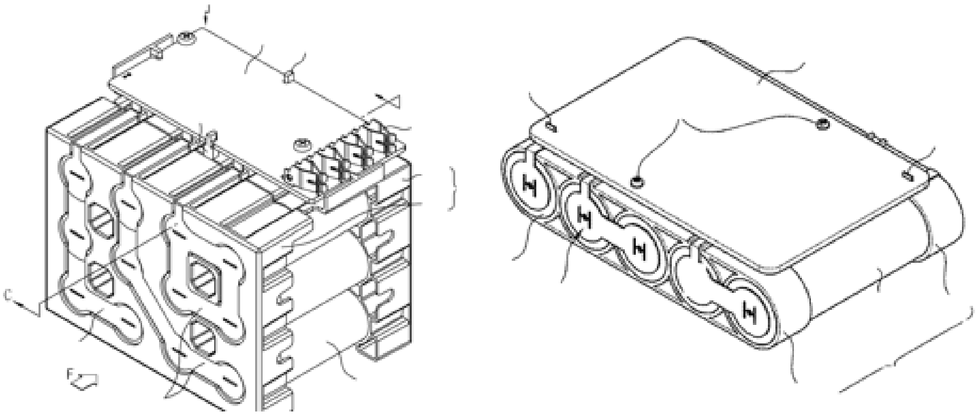
- Functional purpose of the invention
- Core operating principles
- Problem addressed by the invention
- Application of the invention

Score	**Condition**
2 - Relevant	Nearly identical or very similar function and approach
1 - Moderate Relevance	Related function or similar approach
0 - No Relevance	Completely different function/purpose

Table 12: Score card for annotating semantic relevance between shown patent image pairs

Guide: Do these inventions solve the same problem?

Example For example, the following two images are showing two different battery modules. Here, the relevance is high (2 - Relevant) as both inventions have the same function, same working principle and same application, although the images look different.



(a) Patent image showing *battery pack* (b) Patent image showing *battery pack*

Figure 12: Example of an image pair for semantic relevance annotation

3. Subpart Relevance

For subpart relevance, we ask the question "Is one image showing one or more subpart of the same invention depicted in the other image?" We assess the subpart relevance of the two images based on the following points:

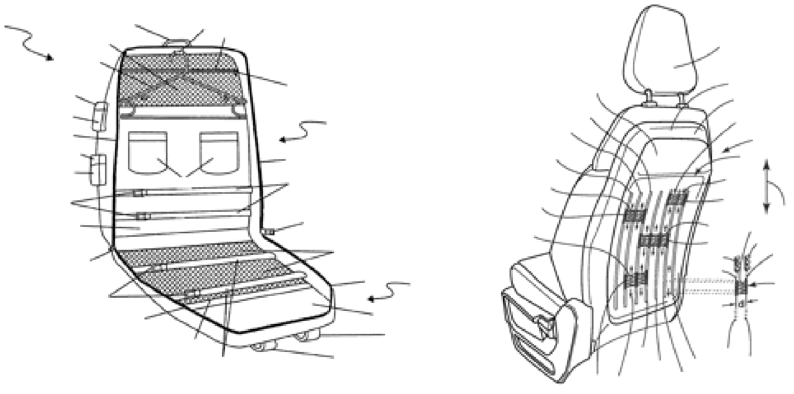
- One image showing key or specific component/s of the invention shown in other image
- Hierarchical relationship exists between the concepts shown in the images

Score	**Condition**
1 - Relevant	One image showing one or more subpart of the other invention
0 - No Relevance	No shared components

Table 13: Score card for annotating subpart relevance between shown patent image pairs

Question to Ask Yourself - Is this image a detailed view or component of the other?

Example For example, the following two images are showing two different subparts of the same concept. Here, the images are relevant (1 - Relevant) as one is a subpart of the other.



(a) Patent image showing *internal structure of a chair*

(b) Patent image showing *back of a chair*

Figure 13: Example of an image pair for semantic relevance annotation