Hierarchical Multi-Positive Contrastive Learning for Patent Image Retrieval

Kshitij Kavimandan[1,*], Angelos Nalmpantis[2], Emma Beauxis-Aussalet[1] and Robert-Jan Sips[2]

[1]Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[2]TKH AI, Amsterdam, The Netherlands

Abstract

Patent images are technical drawings that convey information about a patent's innovation. Patent image retrieval systems aim to search in vast collections and retrieve the most relevant images. Despite recent advances in information retrieval, patent images still pose significant challenges due to their technical intricacies and complex semantic information, requiring efficient fine-tuning for domain adaptation. Current methods neglect patents' hierarchical relationships, such as those defined by the Locarno International Classification (LIC) system, which groups broad categories (e.g., "furnishing") into subclasses (e.g., "seats" and "beds") and further into specific patent designs. In this work, we introduce a hierarchical multi-positive contrastive loss that leverages the LIC's taxonomy to induce such relations in the retrieval process. Our approach assigns multiple positive pairs to each patent image within a batch, with varying similarity scores based on the hierarchical taxonomy. Our experimental analysis with various vision and multimodal models on the DeepPatent2 dataset shows that the proposed method enhances the retrieval results. Notably, our method is effective with low-parameter models, which require fewer computational resources and can be deployed on environments with limited hardware.

Keywords

Information Retrieval, Patent Image Retrieval, Hierarchical Multi-Positive Contrastive Learning

1. Introduction

Patent images are technical drawings that illustrate the novelty of a patent, often conveying their details more effectively than natural language written in text [1]. Thereby, technical patent reports are typically accompanied by multiple images capturing different aspects of the invention. With the rapidly growing volume of patents, efficient patent image retrieval systems are becoming an essential component for searching these vast collections.

Many advances in information retrieval have been largely driven by the power of attention based models [2, 3] and the knowledge acquired during extensive pretraining phases, mainly focused on the language domain. While similar models, such as Vision Transformer (ViT) [4] and ResNet [5], have provided remarkable results on a plethora of vision tasks, they still fall short when processing technical drawings since their pretraining mainly involves natural images [6]. In response, to address this domain shift, researchers have released specialized sketch datasets [7, 8] that facilitate model fine-tuning on such images. Similarly, large scale datasets containing patent images have emerged to address their unique intricacies and enable the development of efficient patent image retrieval methods.

DeepPatent [1] was the first large scale dataset designed for training and evaluating patent image retrieval systems, comprising over $350,000$ images across $45,000$ patents, enabling the development of PatentNet, which exhibited significant improvements in patent image retrieval. Additionally, several studies investigated the generation of synthetic text descriptions by leveraging the zero-shot capabilities of (vision) language models [9, 10], allowing the application of multimodal models, such as CLIP [11], on patent image retrieval. Inspired by DeepPatent, DeepPatent2 [12] provided an extension of

6th Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech) 2025
*Corresponding author.
kshitijk3188@gmail.com (K. Kavimandan); a.nalmpantis@tkh.ai (A. Nalmpantis); e.m.a.l.beauxisaussalet@vu.nl (E. Beauxis-Aussalet); r.sips@tkh.ai (R. Sips)
0009-0002-3760-5882 (K. Kavimandan); 0000-0002-1505-4656 (A. Nalmpantis); 0000-0002-4657-892X (E. Beauxis-Aussalet); 0000-0002-2316-7183 (R. Sips)

the dataset, scaling to more than 2.7 million images with patents spanning from 2007 to 2020 while also incorporating additional metadata like the object's name. Despite the advances in patent image retrieval [13, 14], many methodologies determine the relevance of images based on their association with the same patent. This criterion neglects the rich hierarchical taxonomies of patents that are defined by standardized classification systems. Such hierarchical similarities could potentially enhance the effectiveness of patent image retrieval systems.

In this paper, we aim to address this limitation by leveraging the hierarchical taxonomy of patents as defined by the Locarno International Classification (LIC) system [15], which organizes industrial designs into a structured taxonomy consisting of 32 main classes, each further divided into various subclasses. Figure 1 provides an example of how patents are organized within this hierarchical taxonomy. For brevity, we omit illustrating all classes entailed in the LIC taxonomy. While many studies aim to capture the inherent hierarchical information of data, ranging from representation learning methods [16, 17] to specialized architectures [18], it remains unclear how to properly leverage patents' hierarchical relations for improving patent image retrieval.

To this end, we propose a hierarchical multi-positive contrastive learning method that explicitly integrates these hierarchical relations of patents into the training process. Our method extends upon previous works on patent image retrieval [1] and contrastive learning approaches [11, 19] by treating patent images of the same hierarchical main class, subclass and patent ID as positive with varying degrees of similarity. Figure 1 compares conventional contrastive learning methods with the proposed approach. With the conventional method shown in Figure 1(b), each image is associated only with one positive pair that belongs to the same patent ID. In contrast, the proposed approach in Figure 1(c) respects the hierarchical taxonomy, assigning higher positive scores to images with finer taxonomic relationships. For example, two images from the same patent receive the highest positive score, reflecting their direct relationship. Images that belong only to the same Locarno subclass are assigned a slightly lower positive score, while those that share only the same Locarno main class receive an even lower score.

In our experimental analysis with various architectures, we demonstrate that our approach enhances the retrieval performance. Notably, the proposed method shows great effectiveness with low parameter models which can be deployed in resource constrained environments where computational efficiency is also crucial.

The rest of the paper is structured as follows. First, in Section 2, we formulate the proposed hierarchical multi-positive contrastive learning method for patents. Then, in Section 3, we provide the details of the experimental setup, facilitating the reproducibility of our results. In Section 4, we report our findings and demonstrate the effectiveness of our approach. Finally, in Section 5, we draw the conclusions of this study and discuss future directions.

2. Methodology

To induce hierarchical relations among patents, we propose a Hierarchical Multi-Positive Contrastive Learning approach that leverages the hierarchical taxonomy provided by the LIC system. Our approach enables the model to align patent images of the same main class, subclass and patent ID incrementally closer in the embedding space.

Let X be a collection of patent images, $x \in X$ a sample image from the dataset and $z \in \mathbb{R}^d$ the corresponding image embedding provided by an image encoder. Considering a batch of $2K$ images that form K positive pairs $(x_i, \tilde{x}_i)$ and the embeddings of the anchor image z_i and its positive pair $\tilde{z}_i$, the contrastive loss [11, 20] is defined as:

$$L_i = -\log \frac{\exp(\text{sim}(z_i, \tilde{z}_i))}{\sum_{j=1}^{K} \exp(\text{sim}(z_i, \tilde{z}_j))} \tag{1}$$

where $\text{sim}(z_i, \tilde{z}_j)$ indicates the cosine similarity between the two vector embeddings z_i and $\tilde{z}_j$.

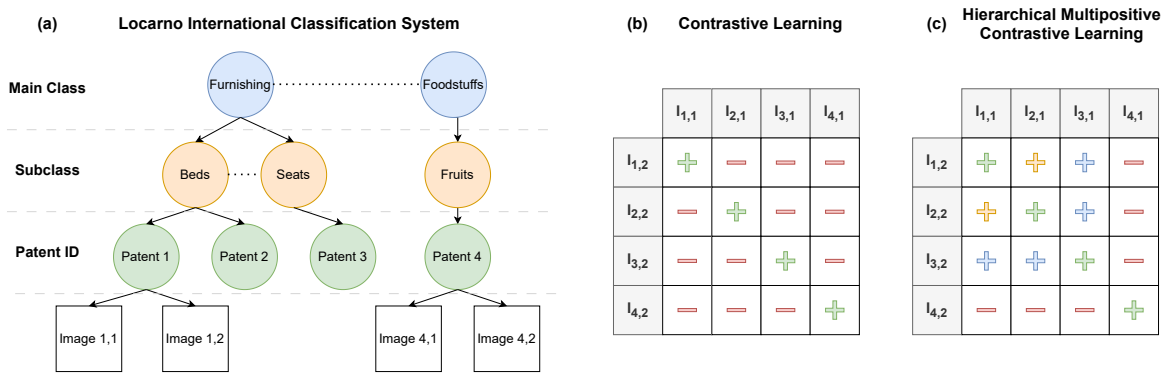

Figure 1: Overview of the Hierarchical Multi-Positive Contrastive Learning approach using the Locarno International Classification system. Figure (a) is an example of the hierarchical taxonomy of patents provided by the Locarno International Classification system. The images from Patent 2 and Patent 3 are omitted. Figure (b) presents how positive pairs within a batch are defined in conventional contrastive learning approaches. Figure (c) summarizes the proposed Hierarchical Multi-Positive Contrastive Learning approach, where the anchor images have multiple positive pairs within the batch, each assigned a different similarity score based on the hierarchical relationship from (a). $\mathrm{I}_{i,j}$ denotes the j-th image that belongs to patent i.

The loss defined in Equation 1, as well as similar losses employed in prior work, such as in PatentNet, are unable to properly capture the hierarchical relations of patents within the batch. In contrast, L_i should accommodate multiple positive pairs for the anchor image x_i and assign a different relevance score to each pair depending on their hierarchical relations within the LIC taxonomy.

Let h_{ij} define the relevance score between two images x_i and $\tilde{x}_j$:

$$h_{ij} = \begin{cases} s_p & \text{if } x_i \text{ and } \tilde{x}_j \text{ belong to the same patent ID} \\ s_s & \text{if } x_i \text{ and } \tilde{x}_j \text{ belong to the same subclass} \\ s_m & \text{if } x_i \text{ and } \tilde{x}_j \text{ belong to the same main class} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $s_p > s_s > s_m$ are positive scalar values that reflect the importance of matching at different hierarchical levels. The function h_{ij} assigns the highest relevance score to the most specific case (same patent ID) with progressively lower scores for broader relationships. Additionally, let H_i be the normalization factor for the patent image x_i:

$$H_i = \sum_{j=1}^{K} h_{ij} \tag{3}$$

Then, the hierarchical multi-positive contrastive loss is defined as:

$$L_i = -\sum_{j=1}^{K} \frac{h_{ij}}{H_i} \log \frac{\exp(\mathrm{sim}(z_i, \tilde{z}_j))}{\sum_{l=1}^{K} \exp(\mathrm{sim}(z_i, \tilde{z}_l))} \tag{4}$$

This formulation enables the model to learn representations that align each image x_i with multiple other images from the batch based on their hierarchical proximity within the LIC taxonomy.

In the case where the text description t_i of the patent image x_i is available, we can incorporate language supervision by adding an additional term to L_i:

$$-\lambda \sum_{j=1}^{K} \frac{h_{ij}}{H_i} \log \frac{\exp(\mathrm{sim}(z_i, y_j))}{\sum_{l=1}^{K} \exp(\mathrm{sim}(z_i, y_l))} \tag{5}$$

where y_j denotes the embedding of the text description t_j provided by a language encoder. The hyperparameter λ is a weighting factor controlling the language supervision.

Note that Equation 1 is a special case of Equation 4. The two equations are equivalent when only a single positive pair exists with a score of 1, and all other pairs are assigned a score of 0.

While our implementation leverages the LIC system, this approach generalizes to other hierarchical classification systems, such as the Cooperative Patent Classification system. Alternative taxonomies can be seamlessly integrated by appropriately defining the scoring function h_{ij} to reflect their specific hierarchical structures.

3. Experimental Setup

For conducting the experiments, we use the DeepPatent2 dataset [12] for the year 2007, which contains multiple images per patent along with the patent's code from the LIC system and a short textual description of the depicted object. The experimental setup is similar to Kucer et al. [1]. We split the data using a $72.25/12.75/15$ ratio for training, validation and testing, respectively. In training, we sample 64 patents, where for each we randomly pick 2 images forming a positive pair based on the patent ID. For testing, we sample 2 images from each patent, with each image being used individually to form a query. The rest of the patent images from the test set form the database used for searching. All images are reshaped with a resolution of 224×224. During training, we use the following augmentation techniques to avoid overfitting: horizontal flip with an applying probability of 0.3, rotation by a maximum of 10 degrees with a probability of 0.5, and Gaussian noise with a probability of 0.2. For testing, no augmentation methods are deployed.

We conduct experiments with the ViT [4], CLIP [11], and ResNet [5] architecture of different sizes. The vision models, ViT and ResNet, are initialized from a pretrained version on ImageNet, while CLIP models are pretrained using the dataset from [11].

We use the AdamW optimizer [21] with a learning rate of 0.0001 and weight decay of 0.01. All models are trained for 20 epochs until convergence with early stopping based on the validation set. Each experiment is repeated for multiple random seeds. For the ViT and ResNet models, we repeat the experiments for 5 different seeds, while for the CLIP models, which require more computational resources, we use 3 different seeds. The temperature τ and the hyperparameter λ are set to 0.1 and 0.2, respectively. For the scoring function h_{ij}, we set $s_p = 1$, $s_s = 0.35$ and $s_m = 0.2$, emphasizing the patent ID level while still incorporating information from higher levels. These values offer a balanced performance across all levels and a fair comparison with the baselines that mainly focus on the patent ID level. Note that a different scoring function could be used depending on the significance of each hierarchical level for the use case at hand.

The models are evaluated using the mean Average Precision (mAP), the normalized Discounted Cumulative Gain (nDCG), the Top-K Mean Reciprocal Rank (MRR@K) and the Top-K Accuracy (Acc@K).

The experiments are conducted using PyTorch [22], PyTorch Lightning [23], and the transformers library from Hugging Face [24]. The training process of a model takes approximately 2.5 hours on a single NVIDIA A100 GPU.

4. Experimental Results

Table 1 presents the results obtained with the ResNet and ViT models. Baseline denotes the pretrained models without any additional finetuning on the patent domain. CL denotes the conventional contrastive learning approach defined in Equation 1 while HMCL indicates the proposed hierarchical multi-positive contrastive learning defined in Equation 4.

The proposed method is evaluated at all hierarchical levels (Patent ID, Subclass and Main Class), with each successive level containing incrementally more relevant images. The set of relevant items at the Patent ID level is a subset of those at the Subclass level, which also form a subset of the relevant items at the Main Class level. This explains why mAP may decrease as hierarchical levels increase (thus, results may be compared row by row, and not column by column).

Table 1
Performance of the ResNet and ViT models on patent image retrieval. All results are averaged across 5 runs with different seeds.

Model	Method	Patent ID		Subclass		Main Class	
		mAP	nDCG	mAP	nDCG	mAP	nDCG
ResNet Models							
ResNet-18	Baseline	0.153	0.366	0.053	0.484	0.076	0.610
	CL	0.208	0.437	0.079	0.514	0.095	0.626
	HMCL	**0.221**	**0.453**	**0.085**	**0.521**	**0.101**	**0.631**
ResNet-34	Baseline	0.152	0.366	0.058	0.494	0.080	0.613
	CL	0.212	0.447	0.086	0.523	0.100	0.629
	HMCL	**0.223**	**0.457**	**0.091**	**0.528**	**0.102**	**0.631**
ResNet-50	Baseline	0.168	0.386	0.059	0.494	0.080	0.616
	CL	0.248	0.482	0.089	0.528	0.102	0.633
	HMCL	**0.251**	**0.484**	**0.093**	**0.530**	**0.106**	**0.636**
ViT Models							
ViT Tiny	Baseline	0.150	0.365	0.054	0.487	0.077	0.612
	CL	0.304	0.536	0.099	0.541	0.102	0.637
	HMCL	**0.310**	**0.542**	**0.105**	**0.546**	**0.110**	**0.642**
ViT Small	Baseline	0.179	0.401	0.065	0.503	0.085	0.620
	CL	**0.349**	**0.577**	0.115	0.557	0.112	0.646
	HMCL	0.347	0.575	**0.119**	**0.560**	**0.119**	**0.650**
ViT Base	Baseline	0.171	0.392	0.064	0.499	0.084	0.619
	CL	**0.333**	**0.564**	0.115	**0.556**	0.113	0.646
	HMCL	0.324	0.555	**0.116**	**0.556**	**0.118**	**0.648**
ViT Large	Baseline	0.166	0.386	0.069	0.506	0.088	0.621
	CL	**0.348**	**0.576**	0.118	0.562	0.112	0.645
	HMCL	0.345	0.573	**0.121**	**0.563**	**0.115**	**0.647**

Overall, the hierarchical multi-positive contrastive loss enhances retrieval performance across all hierarchical levels. Notably, the proposed approach provides significant improvements with the ResNet architecture and lower parameter models such as ViT Tiny. While with larger ViT models, we notice improved performance at the Subclass and Main Class levels, we observe a slight deterioration at the Patent ID level. This trade-off is expected, as images from higher hierarchical levels have a higher similarity score and get higher in the ranking list. Also, we calculate the standard deviation between the runs, but we do not observe any significant difference between the methods. For the Patent ID level, the standard deviation is approximately ± 0.005, for the Subclass level, it is ± 0.002, and for the Main Class level, it is ± 0.001 for both methods and metrics.

Table 2 reports the results with the CLIP model. First, we evaluate only the ViT component from a pretrained CLIP, in isolation from the language encoder. Additionally, we experiment in a multimodal setting with minimal language supervision where the textual descriptions are defined using the following format:

"This is a patent image of a [OBJECT_NAME]."

where [OBJECT_NAME] represents the object's description provided by DeepPatent2. These models provide significant improvements compared to the ViT and ResNet models from Table 1. This can potentially being attributed to the extensive and contextualized pretrained phases of CLIP. Additionally, language supervision further improves performance. Finally, we observe a similar performance trade-off between Patent ID and the higher hierarchical levels, as previously shown in Table 1 for ViT Base and ViT Large. In the case of the CLIP models, the deterioration in performance at the Patent ID level is

Table 2
Performance of the CLIP models on patent image retrieval. All results are averaged across 3 runs with different seeds. The top section reports the results when only the vision component from CLIP is used for training with no language supervision.

Model	Method	Patent ID		Subclass		Main Class	
		mAP	nDCG	mAP	nDCG	mAP	nDCG
ViT component from CLIP							
CLIP-B/16*	Baseline	0.179	0.400	0.077	0.565	0.077	0.563
	CL	**0.373**	**0.596**	0.120	0.609	0.121	0.609
	HMCL	0.356	0.582	**0.128**	**0.613**	**0.139**	**0.618**
CLIP-L/14*	Baseline	0.213	0.437	0.088	0.577	0.087	0.574
	CL	**0.454**	**0.663**	0.136	0.626	0.135	0.624
	HMCL	0.452	0.661	**0.146**	**0.633**	**0.155**	**0.634**
CLIP Models							
CLIP-B/16	Baseline	0.179	0.400	0.077	0.565	0.077	0.563
	CL	**0.401**	**0.619**	0.128	0.617	0.125	0.613
	HMCL	0.386	0.608	**0.137**	**0.623**	**0.148**	**0.625**
CLIP-L/14	Baseline	0.213	0.437	0.088	0.577	0.087	0.574
	CL	**0.458**	**0.665**	0.149	0.634	0.148	0.632
	HMCL	0.439	0.651	**0.156**	**0.639**	**0.171**	**0.642**

more pronounced, resulting from greater improvements in Subclass and Main Class levels.

Figure 2 reports the results with ResNet-18 and ResNet-50 using the metrics MRR@K and Acc@K for $K \in \{1, 5, 10, 20\}$, providing a more comprehensive overview of the retrieved list. For all levels (Patent ID, Subclass, and Main Class), the proposed approach outperforms the conventional contrastive learning method, with more relevant items being found at higher ranks in the retrieved list.

Finally, we project the embeddings of ViT Base into 2 dimensions using PCA. Figure 3 illustrates the samples from 5 subclasses (where 2 subclasses belong to the same main class). We notice that without any hierarchical information induced during training, the classes have a higher overlap and are less distinctly separated. In contrast, the proposed approach leads to more coherent clustering, with samples from the same subclass positioned closer together and subclasses of the same main class being closer in the embedding space.

5. Conclusion

In this paper, we presented a hierarchical multipositive contrastive learning approach to improve patent image retrieval. We integrated the hierarchical relationships of patents defined by the LIC system into the training process, allowing the models to capture this rich information in the embedding space. Our approach considers multiple positive pairs within a batch for an anchor image, with each pair being assigned a different relevance score, which reflects how closely their patents are classified within the chosen hierarchical taxonomy (e.g., LIC). Experimental results demonstrated that our approach enhanced performance at all hierarchical levels, exhibiting notable improvements with low parameter models.

Our findings suggest that incorporating the hierarchical information of patents can improve patent image retrieval, opening several promising avenues for future research. One direction could be to explore hyperbolic embeddings, which are inherently more suitable for capturing hierarchical structures [16]. Finally, our study was specifically focused on the LIC taxonomy. Future directions could investigate alternative taxonomies, for example the Cooperative Patent Classification system, which provides a more granular hierarchical structure with additional levels.

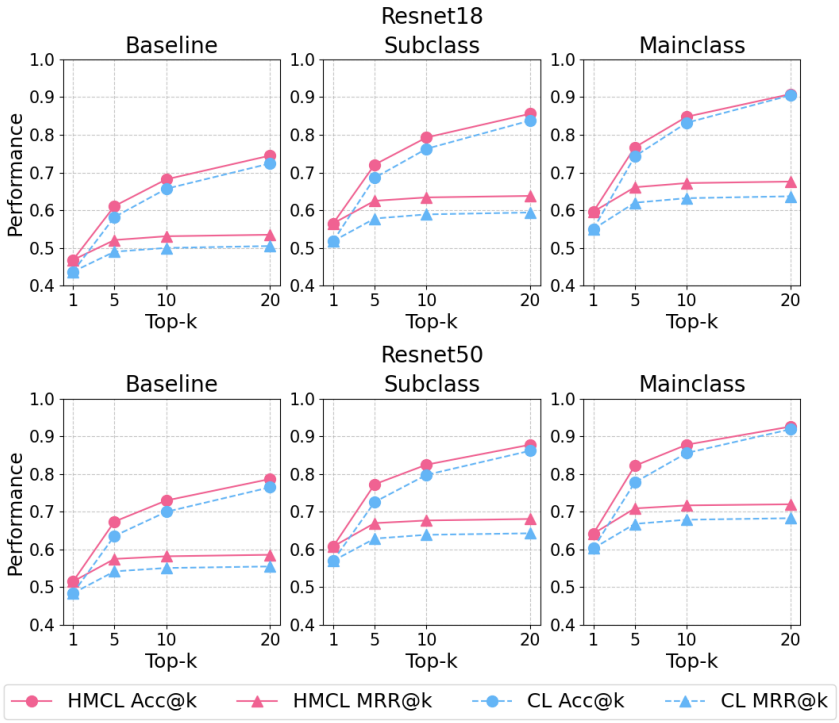

Figure 2: MRR@k and Acc@k for the Baseline, the conventional Contrastive Learning (CL) and the Hierarchical Multipositive Contrastive Learning (HMCL) method (for k = 1, 5, 10, and 20).

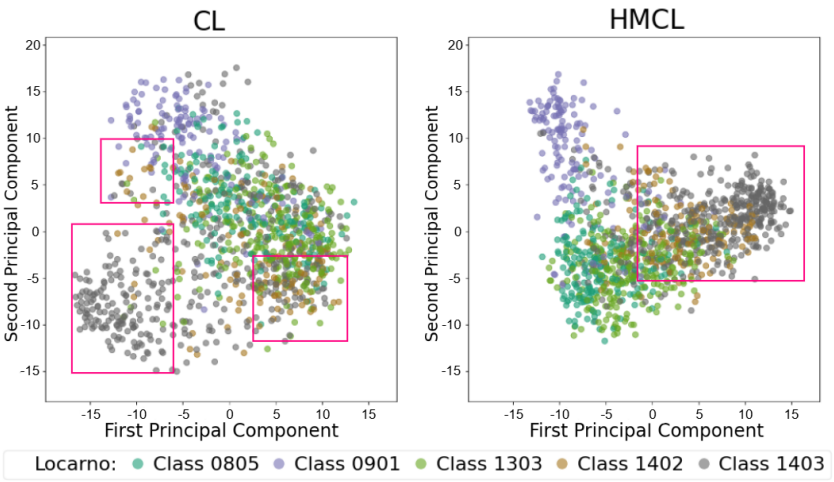

Figure 3: First and second principal component of the embeddings of 5 different subclasses from the LIC system for the conventional Contrastive Learning (CL) and the Hierarchical Multi-Positive Contrastive Learning (HMCL) method. The Subclass 1402 and 1403 from LIC belong to the same Main Class, while the rest to different ones. The red boxes enclose the main regions of Subclass 1402 and 1403.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] 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.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.
[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.
[5] 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.
[6] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, W. Brendel, Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, in: International conference on learning representations, 2018.
[7] H. Wang, S. Ge, Z. Lipton, E. P. Xing, Learning robust global representations by penalizing local predictive power, Advances in neural information processing systems 32 (2019).
[8] 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.
[9] D. Aubakirova, K. Gerdes, L. Liu, Patfig: Generating short and long captions for patent figures, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2843–2849.
[10] H.-C. Lo, J.-M. Chu, J. Hsiang, C.-C. Cho, Large language model informed patent image retrieval, 2024. URL: https://arxiv.org/abs/2404.19360. arXiv:2404.19360.
[11] 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.
[12] 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.
[13] 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.
[14] K. Higuchi, K. Yanai, Patent image retrieval using transformer-based deep metric learning, World Patent Information 74 (2023) 102217.
[15] World Intellectual Property Office, Locarno classification, https://www.wipo.int/classifications/locarno/, 2025. Accessed: April 2025.
[16] P. Mettes, M. Ghadimi Atigh, M. Keller-Ressel, J. Gu, S. Yeung, Hyperbolic deep learning in computer vision: A survey, International Journal of Computer Vision 132 (2024) 3484–3508.
[17] A. Nalmpantis, P. Lippe, S. Magliacane, Hierarchical causal representation learning, in: Causal Representation Learning Workshop at NeurIPS 2023, 2023.
[18] 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.
[19] Y. Tian, L. Fan, P. Isola, H. Chang, D. Krishnan, Stablerep: Synthetic images from text-to-image models make strong visual representation learners, Advances in Neural Information Processing Systems 36 (2023) 48382–48402.
[20] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, 2019. URL: https://arxiv.org/abs/1807.03748. arXiv:1807.03748.
[21] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.
[22] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep

learning library, 2019. URL: https://arxiv.org/abs/1912.01703. arXiv:1912.01703.

[23] W. Falcon, The PyTorch Lightning team, PyTorch Lightning, 2019. URL: https://github.com/Lightning-AI/lightning. doi:10.5281/zenodo.3828935.

[24] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Q. Liu, D. Schlangen (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6/. doi:10.18653/v1/2020.emnlp-demos.6.