DeepPatent: Large scale patent drawing recognition and retrieval

Michal Kucer, Diane Oyen, Juan Castorena
Los Alamos National Laboratory
Los Alamos, NM
{michal, doyen, jcastorena}@lanl.gov

Jian Wu
Old Dominion University
Norfolk, VA
jwu@cs.odu.edu

Abstract

We tackle the problem of analyzing and retrieving technical drawings. First, we introduce DeepPatent, a new large-scale dataset for recognition and retrieval of design patent drawings. The dataset provides more than 350,000 design patent drawings for the purpose of image retrieval. Unlike existing datasets, DeepPatent provides fine-grained image retrieval associations within the collection of drawings and does not rely on cross-domain associations for supervision. We develop a baseline deep learning model, named PatentNet, based on best practices for training retrieval models for static images. We demonstrate the superior performance of PatentNet when trained on our fine-grained associations of DeepPatent against other deep learning approaches and classic computer vision descriptors. With the introduction of this new dataset, and benchmark algorithms, we demonstrate that the analysis and retrieval of technical drawings remains an open challenge in computer vision; and that patent drawing retrieval provides a real-world testbench to spur research.

1. Introduction

Drawings, illustrations, and free-hand-sketches are often used to convey important scientific or technical information more easily than can be described in text [19,38]. Research indicates that humans can learn faster and gain deeper understanding from carefully constructed illustration, as opposed to text alone [5, 28]. A technical drawing[1] is a visual description of an object or concept, conveying important information to a person who does not need to have specific expertise to understand the image [14]. Technical and scientific illustration remain a vital part of conveying information in science and technology [19]; especially in archaeology [29,41], medicine [16,21], design [11], and fashion [20]. Yet, information retrieval for scientific, technical, and scholarly information relies primarily on text-based retrieval — ignoring the drawings even though the drawings may convey more human-accessible information than the text [4, 27]. Our broad goal is to advance computer vision in this area of understanding visual information which is specifically created for human cognition of abstract, technical, and scientific concepts.

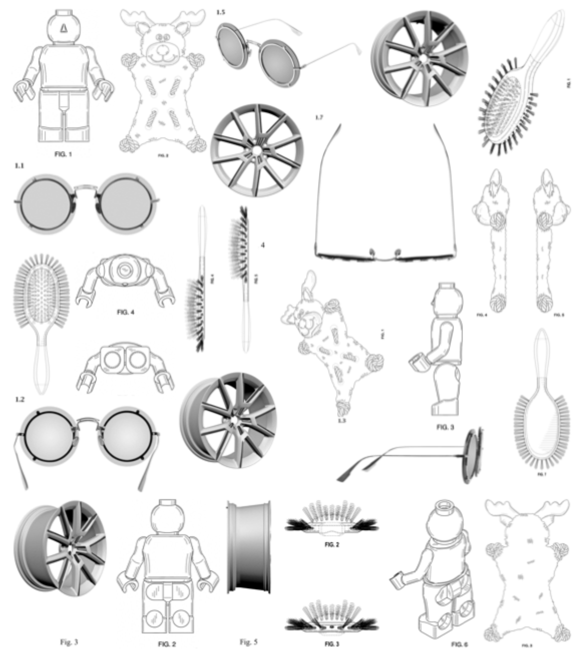


Figure 1: Mosaic of example drawings from five patents: *toy figure, moose-shaped animal toy, spectacles, toy wheel, hairbrush*. Each patent contains multiple drawings of a single object from various views. One need not be an expert to match which of these drawings belong to each of the five patents (answer given in Supplement).

To better understand technical drawings, we focus on patent drawings, as in Figure 1, which are a rich yet not very well explored domain [35, 44, 50]. Patent drawings

[1] We use "drawing" as the most appropriate term in computer vision, whereas "illustration" would be more appropriate in the art community.

are similar to free-hand sketches in that they are both drawings and thus share some properties: lack of background or any contextual information present in natural photos, abstractness, sparseness, etc [47]. In contrast, patents are often more detailed and of higher quality than free-hand sketches (existing sketch datasets self-describe as "badly drawn bunnies" [38]); giving a faithful — rather than exaggerated — representation of an object [14]. Furthermore, patents typically provide a drawing of an object from several different viewpoints including viewpoints that would rarely be featured in a free-hand-sketch (such as undersides and aerial views) [38]; much as a photo could be taken of an object from any viewpoint. We find that commercial image retrieval tools perform poorly on these patent drawings (see Supplement).

Despite the similarity of technical drawings to free-hand-sketch, our empirical results highlight critical gaps in the capability of computer vision approaches to retrieve semantically-similar technical drawings. The impressive advances in sketch-based recognition rely on stroke information and/or associated natural images [38, 47, 48]; neither of which are typically available for technical drawings. DeepPatent provides within-domain fine-grained associations by grouping drawings within a patent as positive examples of relevant drawings. Examples in Figure 1 show that some drawings may not be easily identifiable as a particular object out-of-context; such as the top-view of a brush or side-view of an animal toy; but these are still recognizable to a human as belonging to the same patent as the other views of the same object. These are reasons that we believe the best approach to content-based drawing recognition and retrieval is not classification, but image retrieval.

In addition to the dataset, we train baseline deep learning models for drawing retrieval: PatentNet is a deep network trained on the DeepPatent dataset. PatentNet is based on best practices from recent retrieval approaches that leverage non-associated data from natural images and sketches [15, 36]. We benchmark PatentNet, simpler deep learning approaches, and classic image descriptors on our DeepPatent dataset. We find that the classic image descriptors are brittle due to their reliance on similarity of visual features and lack of learning semantic similarity. Our evaluation of learning-based approaches demonstrates the improved ability to identify images with similar objects through training with fine-grained associations on patent drawings.

Key contributions of this paper are:

- DeepPatent dataset: We collect, process, and make available a large dataset of patent drawings aimed at understanding and retrieving technical illustrations.

- We benchmark several methods to evaluate their efficacy on the DeepPatent dataset including: traditional methods (e.g. fixed image descriptors) and deep learning methods; and provide a strong baseline deep learning approached named PatentNet which is trained on DeepPatent with classification loss, contrastive loss and triplet loss.

The rest of the paper is organized as follows. In Section 2, we discuss the work related to drawing retrieval. In Section 3 we discuss the dataset collected by mining the patent database. In Section 4 we discuss the models used for retrieving patent images. Section 5 discusses and illustrates the quantitative and qualitative results. Section 6 concludes.

2. Related work

Image-based patent retrieval datasets Patent image retrieval has not received as much attention as text-based patent retrieval [27, 34]. CLEF-IP 2011 [35] provides two image-based patent challenge datasets, but only 211 patents are included for the retrieval task; and the image classification task considers 9 classes of broad image types (such as flow chart and chemical structure) rather than fine-grained retrieval. The *concept* dataset is a collection of 1000 patent drawings with a classification challenge of labeling 8 different types of shoe (ski boot, high heel, etc); and another set of 2000 mechanical drawings with categories of relevance [44]. As modern deep learning approaches require more data for training and evaluation, we introduce a large-scale database with more than 350,000 images.

Image-based patent retrieval methods Current approaches use image descriptors to find visually-similar drawings [9, 31, 45] but these approaches perform poorly on DeepPatent. Evidence of the limited effectiveness of visual-similarity approaches are given in a recent survey [50]. Machine learning approaches focus on classification problems; to predict an international patent classification (IPC) label [22], or for classification of 8 types of shoes [1, 44].

Content-based drawing retrieval datasets ImageNet-Sketch provides 50 drawings per class for the 1000 classes of the ImageNet Challenge (ILSVRC) [37] for a total of 50k images [47]. Like our DeepPatent dataset, ImageNet-Sketch provides in-the-wild examples of drawings. However, the variety of drawing styles makes the breadth of the domain quite large with a limited number of examples per class and no fine-grained associations. Retrieval-by-sketch is an important computer vision research topic with its related datasets. The TU-Berlin drawing dataset consists of 20k human sketches covering 250 classes [12], while QuickDraw has 50M sketches covering 345 classes [24]; yet these datasets do not have fine-grained associations. The Sketchy dataset [38] does provide fine-grained associations for 75k free-hand-sketches but we demonstrate better retrieval by training on our DeepPatent dataset.

Content-based drawing retrieval methods There are only a few existing approaches for content-based drawing

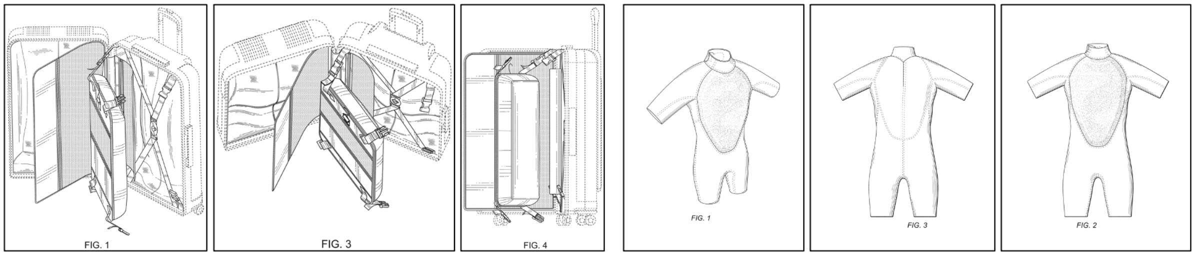


Figure 2: Qualitative examples from DeepPatent showing two objects with three views each.

or sketch retrieval [7], and they use weaker models than PatentNet for encoding the sketch information [46, 48, 49] - either AlexNet [25] or Sketch-A-Net [51]. Much more common are multi-modal problems such as sketch-based image retrieval (SBIR) where objects of higher complexity (natural images) are retrieved through queries of simpler representation (sketches) [38, 48]. Sketch-retrieval methods typically rely on information which we do not have in technical drawings: stroke information, associations with natural images, attributes, or well-defined classification labels [2, 47, 52, 53]. When such auxiliary information is not available, these methods reduce to the traditional approach of classification pre-training and ranking optimization [15, 36] which we implement as PatentNet.

3. The DeepPatent dataset

We introduce the DeepPatent dataset for large-scale drawing retrieval experiments, which we collected, evaluated, and will make easily accessible as a benchmark challenge. With over 350,000 public domain images it is the largest image collection focused on patent drawings. One of the main benefits of design patent drawings is the presence of multiple views of each object in the patent (on the order of 10 drawings per patent) as we can see in Figure 2.

Design patents (as opposed to the more numerous *utility* patents), according to USPTO, capture the visual characteristics or aspects of the object, and thus mostly drawings of the particular object are present (rather than the flowcharts, plots, mathematical expressions, and text-heavy mechanical diagrams in *utility* patents). These detailed, abstract drawings are intended to convey crucial information about the patented object that is better described in picture than in words. When searching patents, people often rely on visual comparison of images in patents to quickly identify relevant prior art [35]. Yet, searching patent drawings based on computer vision is an open challenge.

3.1. Data collection

We first mine the United States Patent and Trademark Office (USPTO) bulk downloads website to collect patent drawings [42]. To establish a computer vision benchmark for technical drawing retrieval, we select only the drawings from patents of the *design* category.

The dataset consists of a total of 45,000 unique design patents that span the year 2018 and the first half of the year 2019. We randomly sample 15% of given patents and reserve them as a test set. This results in 13,133 queries and more than 38,000 database drawings belonging to 6927 patents. To choose queries, we sample 1 or 2 drawings from each test set patent and withhold them. The rest of the drawings from the test-set of patents are set aside to serve as a database of drawings to search through. In the remaining set of patents, 15% of those are further sampled to serve as a validation set resulting in 254787 images in the training set (across 33364 patents), and 44815 images in validation (across 5888).

USPTO provides a weekly bulk download of patents including figures (drawings) in TIF format and an XML containing the text and metadata of all patents awarded in the week. From the bulk download, we extract just the drawings and metadata XML from the *design* category of patents, and convert the TIF images to PNG using ImageMagick. PNG is more widely accepted by computer vision software packages, so we provide this conversion for consistency when comparing methods. Our set of images with metadata is less than 10% of the size of the original bulk download.

3.2. Dataset availability and distribution

The DeepPatent drawing retrieval dataset will be available for download from Google Drive and will have an associated DOI. The dataset includes all image files in PNG format and patent ID labels. The images and labels are distributed as Public Domain CC0 license[2]. Works created for the purpose of USPTO patent application are generally not subject to copyright [42]. See Supplement for notes on the ethics of distributing this dataset.

3.3. Comparison with other datasets

Drawings in DeepPatent are much more detailed than simple sketches and provide more viewpoints for each object. We cannot count the strokes in static images, as is

[2] https://creativecommons.org/publicdomain/zero/1.0/

a standard metric in quantifying complexity of sketches (Sketchy and TU-Berlin report the median strokes per image as 14 and 13 respectively [12, 38]). Instead, we count the number of connected components in each image, noting that for a sketch, the number of strokes is an upper limit on the number of connected components in the corresponding rendered image. The median number of connected components in DeepPatent (705) is orders-of-magnitude larger than in Sketchy sketches (2). The median number of connected components in ImageNet-Sketch (244) is more similar to DeepPatent than to Sketchy — indicating that patent drawings may be similar in level of detail to the broader class of drawings found on the web.

We investigate the similarity of DeepPatent to photos of objects by generating edge maps for photos in ImageNet and Sketchy, using the Canny edge detector [3], and then counting the connected components of the edge map. The median number of connected components in DeepPatent (705) is similar to that of the edge maps generated from the ImageNet validation set (673), providing evidence that the complexity of shapes in the DeepPatent dataset is similar to that of photos of objects. See Supplement for implementation details and quantified results. Yet, edge maps generated from photos look noisier than the clean lines of technical illustrations; and furthermore, not all technical illustrations have a meaningful photo-like representation.

4. PatentNet model for drawing retrieval

This section describes details of the baseline model for patent drawing retrieval, which we denote PatentNet. Though much work has been devoted to creating sketch-specific architectures, many of the accuracy improvements over standard CNN models come from leveraging the stroke information that are available with free hand sketches, or using a hybrid CNN-RNN architecture to process the raster and stroke versions of the sketches, or using multi-domain information to share weights [2, 48]. This auxiliary information is not available for static drawings, therefore our baseline PatentNet model adopts the best practices found in literature for natural static images [15, 36].

Network structure: The base network for all of our models, is either the ResNet18 or ResNet50 [18], as both demonstrate strong performance in many tasks and provide a better baseline and are faster to train than the ever popular VGG-16 [39]. Differing from the original architectures, we replace average pooling with Generalized Mean (GeM) pooling [36]:

$$\mathbf{f}^{(g)} = [f_1^{(g)} \dots f_k^{(g)} \dots f_n^{(g)}], \tag{1}$$

and

$$f_k^{(g)} = \left(\frac{1}{|\mathcal{X}_k|} \sum_{x \in \mathcal{X}_k} x^{p_k} \right)^{\frac{1}{p_k}}, \tag{2}$$

where $\mathcal{X}_k$ is the k^{th} feature map, and p_k is the pooling parameter for that map. These parameters can be set manually, or finetuned separately [36]. In our implementation, we use the same value of $p = 3$ for all feature maps. After pooling we have a single n-dimensional feature vector. Furthermore, similarly to [15], the pooled feature vector $\mathbf{f}^{(g)}$ is l_2 normalized. This output is then passed through a fully connected layer and again l_2 normalized. This last step is equivalent to learning a whitening and dimensionality reduction end-to-end [36]. Also, the l_2 serves to normalize the vector as the feature vectors are going to be compared via inner product.

Weight initialization: First we explore whether self-supervised training of network weights using the patent data could provide a better baseline retrieval performance than using ImageNet trained weights. Singer et al. [40] suggest that networks trained on natural images achieve worse classification performance on drawings and even worse on sketches. To obtain the self-supervised baseline set of weights, we use a self-supervised method introduced by Gidaris et al. [13] which learns image features by learning to predict image rotations for the particular dataset, and we refer to this model as RotNet.

Patent based retrieval: Our patent retrieval models follow the training protocol outlined by Gordo et al. [15]. With weights initialized according to the previous section (step 1), the network weights are then finetuned by classification (step 2), and then further finetuned with retrieval loss (step 3). Step 2: Due to the availability of large datasets of sketch images, we compare whether fine-tuning the ImageNet weights on sketches could provide a better baseline performance as opposed to patents. For sketches, we finetune the network by predicting the 125 classes of sketches of the Sketchy dataset [38]. For finetuning on patents, each patent is given its own label, i.e. all images for a given patent have the patent ID as a class label. Such model is referred to as the *classification loss* (Cl) model.

Step 3: After finetuning the network weights, the classification layer is removed and the rest of the network serves as a base for the retrieval network. We finetune the networks using either the triplet (Tri) or contrastive (Ct) loss. The triplet loss acts on triplets of images:

$$L(I_q, I^+, I^-) = \frac{1}{2} \max\left(0, m + \|q - d^+\|^2 - \|q - d^-\|^2\right), \tag{3}$$

where the triplets (I_q, I^+, I^-) and (q, d^+, d^-) are the notations for images and their feature representations for the triplet (query, positive example, and negative example) respectively. Letter m is the margin parameter.

Contrastive loss [36] acts on matching and non-matching pairs of images:

$$\mathcal{L}(I_i, I_j) = \begin{cases} \frac{1}{2}\|q - d^+\|^2 & \text{if } Y(i,j) = 1 \\ \frac{1}{2}(\max(0, m - \|q - d^-\|))^2 & \text{if } Y(i,j) = 0 \end{cases} \tag{4}$$

5. Results and experiments

In this section we discuss further details of our implementation, comparison models, and experimental results on the DeepPatent dataset.

5.1. Implementation details

All networks are implemented using the PyTorch [32] deep learning framework. The training is performed on NVIDIA Quadro RTX 8000 (48 GB of VRAM) paired with an Intel Xeon CPU. During training, before an image is processed through the network it goes through a set of augmentations that include random flipping and rotations. Then, following [38], a mean and standard deviation of 0.5 is subtracted and divided for each color channel (this is different from traditional pre-processing in which means and standard deviations come from the ImageNet [37] dataset). Though the patent drawings are black and white images, they are treated as 3-channel color images as most deep architectures take color images.

5.2. Evaluation

Following recent retrieval papers [15, 36] to evaluate each of the methods, we calculate the mean Average Precision (mAP) [17] , and Top-K Accuracy $Acc@K$ [26].

Let $\mathcal{X}$ be the set of all images in the dataset, and let $S \subset \mathcal{X}$ be the database of images we search through. Given a query image q, let S_q^+ and S_q^- be the set of matching and non-matching images. Given a distance metric D, and a ranking $\{x_1, x_2, \ldots, x_n\}$ for images in S ($D(x_i, q) \leq D(x_j, q)$ if $i \leq j$). The the Average Precision (AP) for a given query q can be computed as:

$$Prec@K = \frac{1}{K}\sum_{i=1}^{K} 1\left[x_i \in S_q^+\right], \tag{5}$$

$$AP = \frac{1}{|S_q^+|}\sum_{K=1}^{N} 1\left[x_K \in S_q^+\right] \cdot Prec@K \tag{6}$$

where N is the number of images in the database. Lastly, we report the mean Average Precision (mAP) over all queries in the dataset. The Top-K Accuracy is computed as:

$$Acc@K = \frac{1}{Q}\sum_{i=1}^{Q} 1\left[\mathcal{S}_{q_i}^+ \cap \mathcal{S}_{q_i}^K\right], \tag{7}$$

where $1\left[\mathcal{S}_q^+ \cap \mathcal{S}_q^K\right]$ is an indicator function that indicates whether the Top-K retrieved set of images contains at least one image matching the query, and Q is the number of query images in the test set.

Method	mAP	$Acc@1$	$Acc@5$	$Acc@20$
RotNet RN50	0.169	0.416	0.510	0.584
ImageNet RN50	0.291	0.634	0.716	0.779
Sketchy RN50 Cl	0.229	0.532	0.631	0.703
Patent RN18 Cl	0.284	0.590	0.694	0.763
Patent RN50 Cl	0.366	0.682	0.783	0.844
Patent RN18 Ct	0.275	0.578	0.680	0.754
Patent RN18 Tri	0.278	0.586	0.689	0.756
Patent RN50 Ct	0.332	0.636	0.745	0.819
Patent RNet50 Tri	0.379	0.701	0.794	0.851

Table 1: Quantitative comparison of various design choices for the retrieval network on the validation set of the DeepPatent dataset. The first three models compare the retrieval performance of baseline network weights trained via: (RotNet ResNet50) self-supervised training on DeepPatent dataset; (ImageNet ResNet50) supervised training on the ImageNet dataset; and (Sketchy ResNet50) finetuning on the Sketchy dataset. All PatentNet models are pre-trained on ImageNet and finetuned on DeepPatent. Cl denotes the network after classification finetraining. Ct and Tri denote the networks after retrieval finetuning using the contrastive and triplet losses respectively.

5.3. Comparison models

In this section we provide a description of the comparison models. We first describe models that take inspiration from sketch-recognition and retrieval, which will either be trained on sketch recognition, sketch-based image retrieval (SBIR), or on DeepPatent. Then we describe traditional computer vision methods that are currently used to perform patent retrieval.

Sketch-a-Net is a seminal model, as it is the first deep network to beat human level performance on sketch recognition [51]. Sketch-a-Net serves as a network of choice for many works on SBIR, where it is used as the feature descriptor for static images of sketches [6, 23, 52]. The model contains specific architectural choices that are aimed at improving sketch understanding (e.g. larger filters and pooling regions).

Sketchy-Resnet is a sketch-based network trained for the purpose of sketch-to-image retrieval. Motivated by the work of Bhattarai et al. [1], we include this model to assess the domain generalization performance between sketches and patent drawings. The network is trained in a two-step process similar to [38]. First, two ImageNet pre-trained ResNet50 networks are re-trained on Sketchy photos and sketches to predict the 125 Sketchy categories. Next, the networks are optimized for retrieval using the triplet loss on fine-grained associated sketches. The triplet loss is in the end combined with the softmax classification loss for predicting the object categories.

Adaptive hierarchical density histogram (AHDH) creates adaptively-sized regions of the image by hierarchically calculating the centroid of the region and estimates the distribution of black points in these regions and is demonstrated to work well on patent drawings [45]. **Histogram of oriented gradients (HOG)** [10] counts the occurrence of discrete number of gradient orientations in an image patch. **VisHash** generates a signature based on relative brightness in regions of the image and is demonstrated to match visually similar images on a wide variety of image types including drawings [31]. **Local binary patterns (LBP)** [30] is a rotation-invariant texture descriptor that classifies each local region into one of 58 so-called *uniform* patterns. The normalized histogram of these patterns is used as the image descriptor. **Fisher vectors (FV)** [8, 33] are an extension of the popular bag-of-visual-words representations which generate a fixed-length image representation. Implementation details are given in the Supplement.

5.4. Results

5.4.1 Modular evaluation of the retrieval model

We first compare Step 1 weight-initialization strategies for the model. The standard in image retrieval (including sketch-based) is to pre-train using ImageNet [38], yet a recent study suggests that models trained on natural images may not be the most appropriate [40]. Therefore, we compare the retrieval performance of an ImageNet trained ResNet50 model with the RotNet ResNet50 model initialized by self-supervised feature learning on the patent data directly. As we can see from the first two rows of Table 1, the ImageNet pre-trained weights achieve better retrieval performance as opposed to self-supervised training on the target data, despite the network never having "seen" any patents. This suggests that networks trained on natural images might be a good starting point for developing models on drawings.

We explore the the following choices for Step 2 training the retrieval model: (a) the set of data used for classification fine-tuning akin to Gordo et al. [15], (b) the backbone architecture, and (c) the ranking loss. Due to the maturity of sketch-based datasets, e.g. Sketchy [38], and the similarity of sketches to drawings, we check if using Sketchy would would provide a better fine-tuning over patents. As we can see from Table 1, fine-tuning on patent drawings provides better performance, and sketches make the performance worse even as compared to the baseline ImageNet weights. For backbone networks, we choose to compare the ResNet18 and ResNet50 models. As we can see from rows 4 and 5 in Table 1, a deeper ResNet50 model achieves better performance on patent retrieval as compared to ResNet18. We can see that ResNet18 achieves slightly lower performance when trained with either retrieval loss as compared to classification model, though the difference between the two losses is negligible. In the case of ResNet50, the triplet loss learns a better retrieval model, significantly outperforming the contrastive loss in this case. As the best model, ResNet50 Tri is used in comparison to other deep features and traditional approaches.

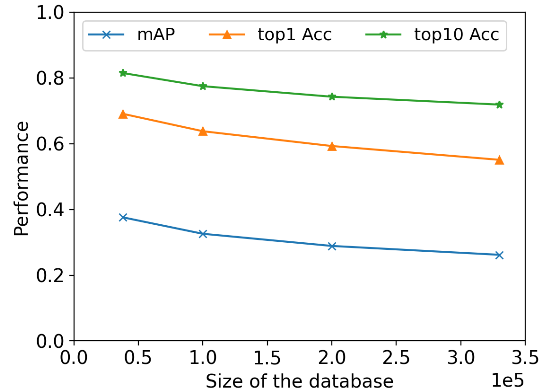
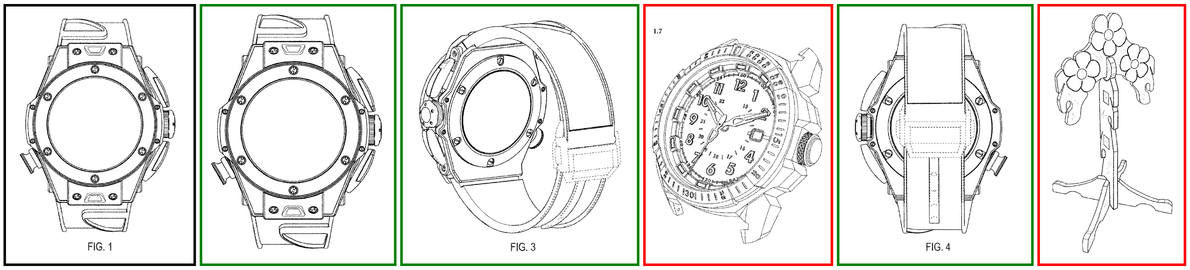


Figure 3: Plot of the model performance in terms of the mean average precision (mAP), top-1 accuracy, and top-10 accuracy as a function of the database size.

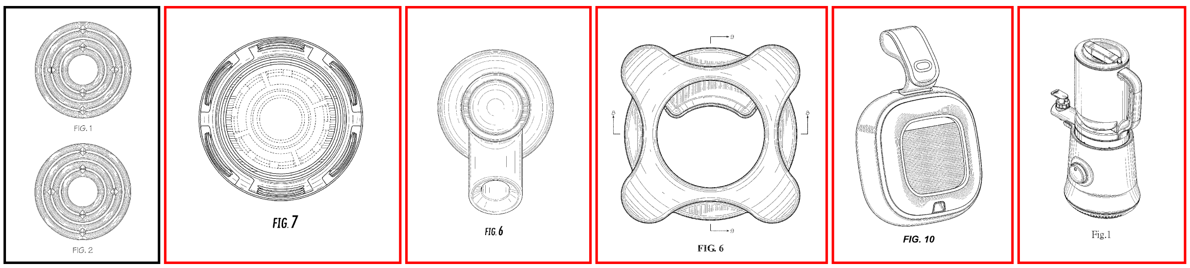
Scalability The test set itself consists of more than fifty thousand images spread across 6927 patents, split into 13,133 queries and 38,834 database images. To further demonstrate the challenge of patent drawing retrieval, we experiment with expanding the size of the database we search through by adding in the training and validation images. Note that this is not a problem, as we do not use training and validation images as queries. Figure 3 shows the plot of the model performance as a function of the database size. As we can see, adding additional images further increases the difficulty of the problem. As we can see mAP and top-1 accuracy go from 0.376 and 69.1% respectively when we search through roughly 38,000 images down to 0.262 and 55.1% when we search through roughly 350,000 images. This demonstrates the difficulty of the problem as we see a significant drop in the metrics by simply searching though one and a half years worth of images.

5.4.2 Comparison to classic computer vision

We perform quantitative comparison of PatentNet to non-learning computer vision methods and other deep features using the metrics defined in Section 5.2. Table 2 shows the improved performance of PatentNet, the best model from our deep architecture studies (ResNet50 Tri), against the previous top-performing image descriptors for drawing retrieval. Furthermore all learning-based models (even the self-supervised RotNet) outperform all of these classic ap-



(a) Retrieval example showing some patents contain very similar images



(b) Sample failure case

Figure 4: Qualitative examples of retrieval results for PatentNet

Method	mAP	$Acc@1$	$Acc@5$	$Acc@20$
AHDH	0.095	0.288	0.343	0.399
VisHash	0.093	0.274	0.340	0.402
SIFT FV	0.092	0.206	0.289	0.375
HOG	0.083	0.272	0.317	0.359
LBP	0.069	0.210	0.252	0.343
PatentNet	0.376	0.691	0.784	0.841
SANet Sketches	0.086	0.258	0.324	0.388
SANet Patent	0.135	0.361	0.451	0.536
SK Sketches Cl	0.229	0.532	0.631	0.703
SK Sketches RT	0.156	0.428	0.513	0.586
SK Photo RT	0.132	0.353	0.452	0.539

Table 2: Comparison of PatentNet, traditional computer vision approaches, and other deep representations in retrieval performance on the DeepPatent test set. For PatentNet, we use the best performing model on the validation set from the various design choices. SANet denotes the Sketch-a-Net network, SK denotes the Sketchy-ResNet SBIR model and the additional term denotes the domain the network was trained on. For SK models, Cl denotes the models after classification pretraining and RT indicates that the domain specific model was trained SBIR as described in Section 5.3

proaches. The superior performance of deep-learning approaches validates the creation of the large-scale DeepPatent dataset. Additionally, we compare our model to other learning-based approaches - Sketch-A-Net and Sketchy-ResNet, an SBIR model trained on the Sketchy dataset. As we can see, PatentNet and ResNet50 pretrained on either patent drawings or Sketches outperform the Sketch-a-Net model (trained on either domain). Though this is not surprising as Sketch-a-Net is based on the AlexNet architecture and achieves much weaker performance as compared to ResNet50 and newer models. Furthermore, we can see that PatentNet and ResNet50 outperform Sketchy-ResNet deep features.

It is surprising to find that the performance of the ImageNet-pretrained ResNet50 (ImageNet RN50) drops when finetuned on sketches (Sketchy RN50 Cl in Table 1 and SK Sketches Cl in Table 2), and further drops when the networks are trained for SBIR (SK Sketches RT and SK Photo RT in Table 2). To investigate how patent drawing data informs sketches, we train two SBIR ResNet 50 models with different initialisations for the network used to extract features for sketches. The sketches branch is either pretrained on sketches from the Sketchy database or patent drawings, and then the sketches branch and photo branch are fine-tuned for sketch-based image retrieval using the Sketchy database. However in this case, we find that the cross-domain retrieval performance for the sketch-trained network is significantly better than the network pre-trained on patents. This demonstrates that despite the seeming similarity between sketches and drawings, they comprise two different domains and further that methods must be developed to achieve cross-modal understanding.

5.4.3 Qualitative results

Figure 4 shows qualitative examples from the PatentNet model. For qualitative comparison of PatentNet with other

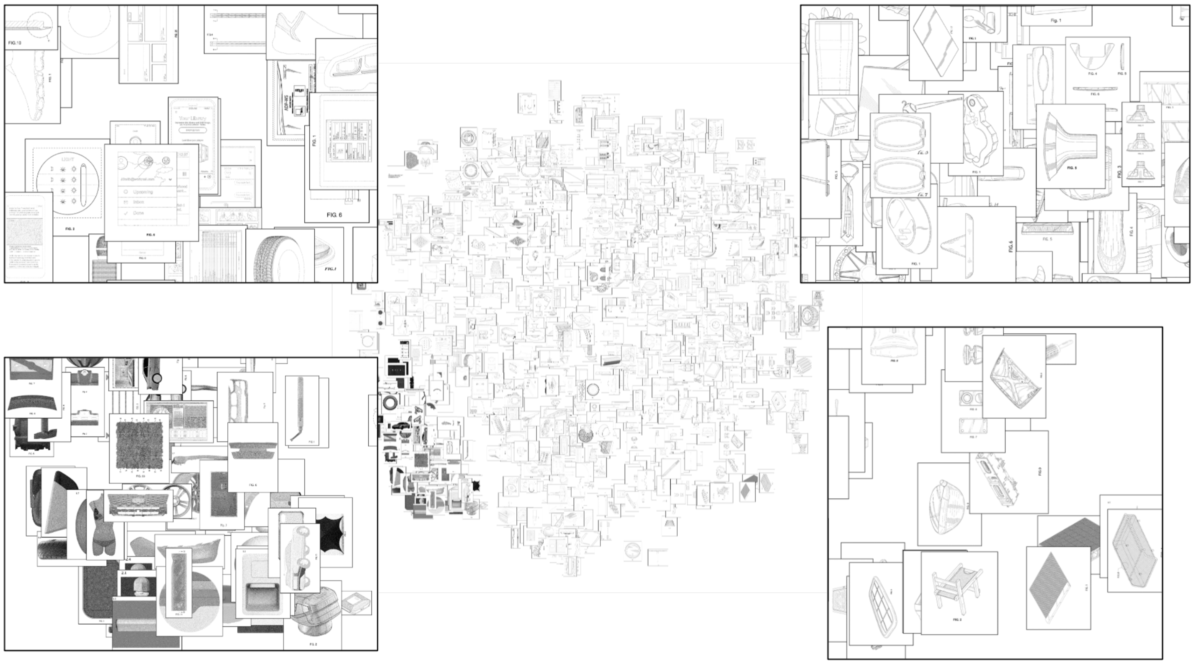


Figure 5: t-SNE visualization of a random subset of 1000 images from the test set.

models, please refer to the Supplement. Figure 4a shows an example of a successful retrieval in which all but one of the retrieved examples are arguably relevant to the query. We are interested in such cases, as this could help one search for prior art. Figure 4b, shows a failure case in which when searching for drawings of toy rings, the model fails to retrieve any correct examples. Though none of the retrieved examples are correct, we can see the top three retrieved examples are circular similar to the query.

Figure 5 visualizes a subset of the testing set by projecting the feature descriptors of the PatentNet model into a two dimensional space using t-SNE [43]. From further inspection, we notice that patent drawings that are photo-like renderings (bottom left) are all clustered together, most likely due to their similar texture. In top left of the figure, we highlight a cluster of images that each depict some sort of a display, table or a figure. Lastly, on bottom left we show a cluster of drawings that show objects from a very similar perspective. This indicates that the network has learned some higher level semantics about the drawings, despite only providing fine-grained associations with any further supervision through class labels or attributes.

5.5. Opportunities and future work

The DeepPatent dataset and findings from the develop of PatentNet open up new opportunities for future work.

Learning image representations at different levels of abstraction In our earlier discussion, we pointed out the unique nature of patent drawings as being in a level of abstraction between sketches and natural images. Given that previous datasets were either limited in size, or focused on particular patent types (i.e. shoes) [44], we believe a large collection of patents such as this would pave the way to building models that could understand objects at various levels of abstraction [40].

Learning robust image representations Recent work notes that Imagenet-trained models rely heavily on color, texture and background pixels rather than the foreground and shape features that are most prominent to people; and so DeepPatent could be used to develop models that are more sensitive to shape and robust to image type [47].

6. Conclusion

We introduce the DeepPatent dataset, a large-scale collection of patents for content-based drawing retrieval. The dataset contains over 350,000 design patent drawings split into train, validation and test sets. We find that although deep learning methods outperform hashing based methods, our retrieval networks achieve a much better performance both in terms of the mean Average Precision as well as Top-K retrieval accuracy. From our results, we see that patent drawing retrieval is a challenging problem and we hope this will spur further research into developing methods that effectively analyze abstract drawings that are prevalent in technical publications and on the web.

References

[1] M. Bhattarai, D. Oyen, J. Castorena, L. Yang, and B. Wohlberg. Diagram image retrieval using sketch-based deep learning and transfer learning. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 663–672, 2020.

[2] Tu Bui, Leonardo Ribeiro, Moacir Ponti, and John Collomosse. Sketching out the details: Sketch-based image retrieval using convolutional neural networks with multi-stage regression. *Computers & Graphics*, 71:77–87, 2018.

[3] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):679–698, 1986.

[4] Cornelia Caragea, Jian Wu, Alina Ciobanu, Kyle Williams, Juan Fernández-Ramírez, Hung-Hsuan Chen, Zhaohui Wu, and Lee Giles. Citeseer x: A scholarly big dataset. In *European Conference on Information Retrieval*, pages 311–322, 2014.

[5] Russell N Carney and Joel R Levin. Pictorial illustrations still improve students' learning from text. *Educational Psychology Review*, 14(1):5–26, 2002.

[6] John Collomosse, Tu Bui, Michael Wilber, Chen Fang, and Hailin Jin. Sketching with style: Visual search with sketches and aesthetic context. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2679–2687, 2017.

[7] Antonia Creswell and Anil Anthony Bharath. Adversarial training for sketch retrieval. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pages 798–809, Cham, 2016. Springer International Publishing.

[8] Gabriela Csurka. Document image classification, with a specific view on applications of patent images. In *Current Challenges in Patent Information Retrieval*, 2016.

[9] 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 *Conference and Labs of the Evaluation Forum*, volume 2, 2011.

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition*, 2005.

[11] Koos Eissen. *Sketching Drawing Techniques for Product Designers*. Taylor & Francis Group, 2006.

[12] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Transactions on Graphics*, 31(4), 2012.

[13] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

[14] Betina Giemsa. Technical illustration in the 21st century: A primer for today's professionals. Technical report, Parametric Technology Corporation, Needham, MA, USA, 2007.

[15] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision*, 2017.

[16] Rachel Hajar. Medical illustration: Art in medical education. *Heart Views*, 12(2):83, 2011.

[17] K. He, Y. Lu, and S. Sclaroff. Local descriptors optimized for average precision. In *Computer Vision and Pattern Recognition*, pages 596–605, 2018.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[19] Elaine RS Hodges. *The Guild Handbook of Scientific Illustration*. John Wiley & Sons, 2003.

[20] John Hopkins. *Fashion Design: The Complete Guide*. Bloomsbury Publishing, 2020.

[21] Peter S Houts, Cecilia C Doak, Leonard G Doak, and Matthew J Loscalzo. The role of pictures in improving health communication: a review of research on attention, comprehension, recall, and adherence. *Patient Education and Counseling*, 61(2):173–190, 2006.

[22] Shuo Jiang, Jianxi Luo, Guillermo Ruiz Pava, Jie Hu, and Christopher L. Magee. A convolutional neural network-based patent image retrieval method for design ideation. *Computers and Information in Engineering Conference (CIE)*, 2020.

[23] Tony Xiang Jifei Song, Yi-zhe Song and Timothy Hospedales. Fine-grained image retrieval: the text/sketch input dilemma. In Gabriel Brostow Tae-Kyun Kim, Stefanos Zafeiriou and Krystian Mikolajczyk, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 45.1–45.12. BMVA Press, September 2017.

[24] J. Jongejan, H. Rowley, T. Kawashima, J. Kim, and N. Fox-Gieg. The Quick, Draw! - A.I. Experiment. `https://quickdraw.withgoogle.com/`, 2016.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

[26] M. Kucer and N. Murray. A detect-then-retrieve model for multi-domain fashion item retrieval. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 344–353, 2019.

[27] Shaobo Li, Jie Hu, Yuxin Cui, and Jianjun Hu. DeepPatent: Patent classification with convolutional neural networks and word embedding. *Scientometrics*, 117(2):721–744, 2018.

[28] Richard E Mayer. Illustrations that instruct. *Advances in Instructional Psychology*, 10:253–284, 2019.

[29] Brian Leigh Molyneaux. *The Cultural Life of Images: Visual Representation in Archaeology*. Routledge, 2013.

[30] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[31] D. Oyen, M. Kucer, and B. Wohlberg. VisHash: Visual similarity preserving image hashing for diagram retrieval. In *Applications of Machine Learning*, volume 11843. International Society for Optics and Photonics, 2021.

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

[33] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the Fisher kernel for large-scale image classification. In *European Conference on Computer Vision*, page 143–156, 2010.

[34] Florina Piroi, Mihai Lupu, and Allan Hanbury. Overview of CLEF-IP 2013 lab. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 232–249, 2013.

[35] Florina Piroi, Mihai Lupu, Allan Hanbury, and Veronika Zenz. CLEF-IP 2011: Retrieval in the intellectual property domain. In *Conference and Labs of the Evaluation Forum*, 2011.

[36] F. Radenović, G. Tolias, and O. Chum. Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[38] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: Learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (proceedings of SIGGRAPH)*, 2016.

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015.

[40] Johannes Singer, Martin N Hebart, and Katja Seeliger. The representation of object drawings and sketches in deep convolutional neural networks. In *NeurIPS 2020 Workshop SVRHM*, 2020.

[41] Mélanie Steiner and Lindsay Allason-Jones. *Approaches to Archaeological Illustration: A Handbook*. 2005.

[42] United States Patent and Trademark Office. Terms of use for USPTO websites. https://www.uspto.gov/terms-use-uspto-websites, 2021.

[43] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[44] Stefanos Vrochidis, Anastasia Moumtzidou, and Ioannis Kompatsiaris. Concept-based patent image retrieval. *World Patent Information*, 34(4):292–303, 2012.

[45] Stefanos Vrochidis, Symeon Papadopoulos, Anastasia Moumtzidou, Panagiotis Sidiropoulos, Emanuelle Pianta, and Ioannis Kompatsiaris. Towards content-based patent image retrieval: A framework perspective. *World Patent Information*, 32(2):94 – 106, 2010.

[46] Fang Wang and Yi Li. Spatial matching of sketches without point correspondence. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 4828–4832, 2015.

[47] Haohan Wang, Songwei Ge, Eric P Xing, and Zachary C Lipton. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, 2019.

[48] Peng Xu, Timothy M. Hospedales, Qiyue Yin, Yi-Zhe Song, Tao Xiang, and Liang Wang. Deep learning for free-hand sketch: A survey and a toolbox. *arXiv preprint arXiv:2001.02600*, 2020.

[49] Peng Xu, Yongye Huang, Tongtong Yuan, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, Zhanyu Ma, and Jun Guo. Sketchmate: Deep hashing for million-scale human sketch retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[50] Liping Yang, Ming Gong, and Vijayan K. Asari. Diagram image retrieval and analysis: Challenges and opportunities. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

[51] Yongxin Yang and Timothy M. Hospedales. Deep neural networks for sketch recognition. *CoRR*, abs/1501.07873, 2015.

[52] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen Change Loy. Sketch me that shoe. In *Computer Vision and Pattern Recognition (CVPR)*, pages 799–807, 2016.

[53] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. Sketchnet: Sketch classification with web images. In *Computer Vision and Pattern Recognition*, pages 1105–1113, 2016.