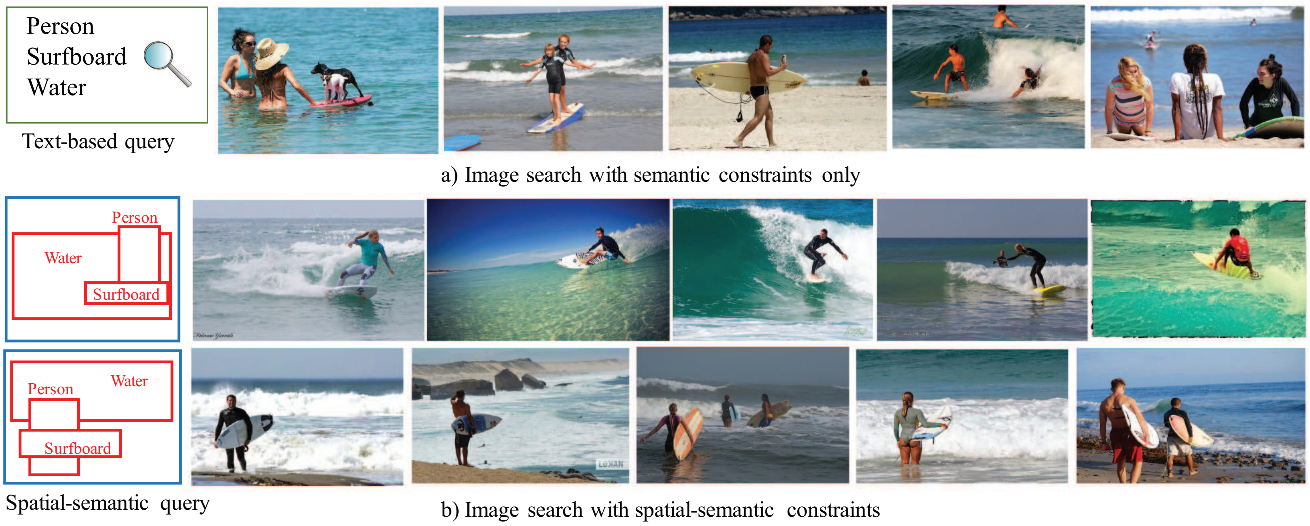
Spatial-Semantic Image Search by Visual Feature Synthesis

Long Mai[1], Hailin Jin[2], Zhe Lin[2], Chen Fang[2], Jonathan Brandt[2], and Feng Liu[1]
[1]Portland State University
[2]Adobe Research
[1]{mtlong,fliu}@cs.pdx.com, [2]{hljin,zlin,cfang,jbrandt}@adobe.com



a) Image search with semantic constraints only

b) Image search with spatial-semantic constraints

Figure 1: Spatial-semantic image search. (a) Searching with content-only queries such as text keywords, while effective in retrieving relevant content, is unable to incorporate detailed spatial intents. (b) Spatial-semantic image search allows users to interact with the 2-D canvas to express their search intent both spatially and semantically.

Abstract

The performance of image retrieval has been improved tremendously in recent years through the use of deep feature representations. Most existing methods, however, aim to retrieve images that are visually similar or semantically relevant to the query, irrespective of spatial configuration. In this paper, we develop a spatial-semantic image search technology that enables users to search for images with both semantic and spatial constraints by manipulating concept text-boxes on a 2D query canvas. We train a convolutional neural network to synthesize appropriate visual features that captures the spatial-semantic constraints from the user canvas query. We directly optimize the retrieval performance of the visual features when training our deep neural network. These visual features then are used to retrieve images that are both spatially and semantically relevant to the user query. The experiments on large-scale datasets such as MS-COCO and Visual Genome show that our method outperforms other baseline and state-of-the-art methods in spatial-semantic image search.

1. Introduction

Image retrieval is essential for various applications, such as browsing photo collections [6, 52], exploring large visual data archives [15, 16, 38, 43], and online shopping [26, 37]. It has long been an active research topic with a rich literature in computer vision and multimedia [8, 30, 55, 56, 57]. In recent years, advances in research on deep feature learning have led to effective image and query representations that are shown effective for retrieving images that are visually similar or semantically relevant to the query [12, 14, 25, 53].

However, in many search scenarios, such as recalling a specific scene from personal albums or finding appropriate stock photos for design projects, users want to indicate not only which visual concepts should appear in the image, but also how these concepts should be spatially arranged within the scene. This paper presents a *spatial-semantic image search* method, which allows users to interact with a 2D canvas to construct search queries. As shown in Figure 1 (b), by manipulating text boxes representing visual concepts, users can naturally express their search intent both spatially and semantically. In contrast, searching with content-only queries, such as text keywords, while effective in retrieving images with relevant content, is unable to represent such detailed spatial queries (Figure 1 (a)).

The main challenge in developing such a spatial-semantic image search technology is to design appropriate image and query representations [8, 57]. Traditional search paradigms often use text-based or image-based queries for which effective feature representations have been well studied [18, 31, 41, 46]. However, the representation for the query of spatial-semantic image search is not as well studied. Early research on spatial-semantic image search [4, 34, 59] mostly follows example-based approaches, extracting low-level visual features from the separately retrieved visual exemplars to represent the canvas queries. Despite promising performance, these methods rely on carefully designed algorithms for both feature extraction and feature matching, which often cannot be generalized well.

In this paper, we present a learning-based approach to visual feature synthesis to support spatial-semantic image search. Instead of manually constructing features from separate visual exemplars, our method learns to synthesize visual features directly from the user query on the 2D canvas. Specifically, we develop a convolutional neural network to synthesize visual features that simultaneously capture the spatial and semantic contraints from the user canvas query. We train this neural network by explicitly optimizing the retrieval performance of its synthesized visual features. This learning strategy allows our neural network to generate visual features that can be used to effectively search for spatially and semantically relevant images in the database.

Our experiments on large-scale datasets like MS-COCO and Visual Genome show that our method outperforms other baseline and existing methods in spatial-semantic image search. Our study demonstrates that our method can support users to retrieve images that match their search intent. Furthermore, our experiments indicate that our feature synthesis method can capture relationships among different visual concepts to predict useful visual information for concepts that were not included in training. This demonstrates its potential for generalization to novel concepts.

2. Related Work

Our work is related to multi-modal image retrieval research in which the database and the queries belong to different modalities. Advances in feature learning have recently provided effective feature representations for different modalities such as text [14, 31, 41, 42], images [1, 2, 17, 18, 19, 45, 46], and hand-drawn sketches [47, 53, 58, 61], which have been shown to greatly improve the retrieval performance. Most existing works target traditional image search paradigms which focus on retrieving images with relevant semantic content or visual similarity. The learned representations are thus designated to capture only semantic information or visual information. The spatial-semantic image search paradigm targeted in this paper, on the other hand, requires a special type of query that contains not only the semantic concepts but also their spatial information. In this paper, we provide a feature synthesis approach to learn effective visual representation for such spatial-semantic canvas queries.

A common approach in representation learning with multi-modal data is to learn a joint embedding to map all modalities into a common latent space [3, 14, 31, 32, 50, 53]. In this paper, we follow a different approach which fixes the image feature representation and learn to synthesize that visual feature from the user given queries. This approach can take advantage of the well established image features, such as the ones obtained by pre-trained deep neural networks, which faithfully preserve important visual and semantic information from images [10]. Another advantage is the flexibility provided by the fact that the image feature is not affected by the query representation learning, which helps avoids the cost of re-processing the database when the feature synthesis model is changed [5].

In the context of incorporating spatial information into image search systems, Zavesky *et al.* [62, 63] present visualization methods to display the search results by arranging the retrieved images onto the 2D layout of the search page according to their content similarity. However, these works focus on the problem of visualizing and browsing the retrieved images which are obtained by text-based queries without spatial information. Our work, on the other hand, addresses a different problem and focuses on retrieving relevant images with respect to the spatial and semantic constraints specified in the canvas queries.

Most relevant to our research are the existing works on spatial-semantic image search [4, 34, 59]. These research mostly follow exemplar-based approaches. A set of visual exemplars for each visual element in the query are pre-determined. Low-level features such as SIFT [36] and color histograms are then extracted from these exemplars to form the visual representation for the queries. Our spatial-semantic image search framework proposed in this paper is different from these methods in two important aspects. First, instead of relying on visual exemplars for feature extraction, our method provides a model-based framework which explicitly learns a synthesis model to synthesize the visual features directly from user queries. Second, instead of manually designing ad-hoc processes for feature computation and matching, we provide a data-driven approach which learns the feature synthesis model from training data so as to explicitly optimize the retrieval performance.

In recent years, convolutional neural network models have shown great successes in generating image data [9, 20, 48]. Recent research have been able to generate realistic images from different input information such as texts

[39, 51], attributes [11, 60, 64], and images from different views [13, 27, 44]. In a recent work, Reed *et al.* [49] present a method to synthesize an image given a scene canvas which is similar to our spatial-semantic query. Inspired by the success of these research on image generation, this paper leverages convolutional neural network models to synthesize a visual representation from input semantic information. However, different from image generation research, our goal is not to generate realistic image data. Instead, we aim to synthesize useful features for image search. The training framework for our model, as a result, needs to be tailored to optimize the retrieval performance.

3. Visual Feature Synthesis

We implement our visual feature synthesis model using a convolutional neural network architecture. Our framework first represents the 2-D input canvas query as a three dimensional grid Q whose depth dimension corresponds to the semantic vector such as *Word2Vec* [41] for the concept appearing at each spatial position in the query. We note that using semantic vector representation instead of the one-hot-encoding alternative can help exploit the relationship in the semantic space and generalize to larger classes of concepts. The grid entries corresponding to unspecified regions in the canvas are set as zeros. During synthesis, the query grid Q is then passed through the feature synthesis model to synthesize the visual feature f_Q for the query.

Figure 2 illustrates our visual feature synthesis framework. While the model is applicable to queries with any number of concepts, we found it best to train the network for only single-concept queries due to two important reasons. First, multi-concept queries often contain concept boxes overlapping one another, which causes ambiguity in terms of semantic representation at the overlapped regions. Moreover, as the number of concepts increases, the number of feasible images that match any specific spatial configuration of those concepts is often limited, which limits the amount of available data to train the synthesis model directly for multi-concept queries. In general, when an input query consists of multiple concepts, we first represent it as multiple single-concept sub-queries. We then synthesize visual features independently for each sub-query and combine them together with the max operator in the end to form the final feature for the whole query. We note that other methods for combining the input *Word2Vec* descriptors or the output features at the overlapping regions can also be used [23].

3.1. Model training

Let Q denote a spatial-semantic canvas query and f_Q denote the visual feature synthesized from Q using our feature synthesis network, we define the per-query loss function as

$$L(f_Q) = w_S L_S(f_Q) + w_D L_D(f_Q) + w_R L_R(f_Q) \quad (1)$$

where L_S, L_D, and L_R are the three individual loss terms (described below) modelling three objectives that guide the network learning. The relative loss weights w_S, w_D, and w_R are heuristically determined as 0.6, 0.3, 0.1, respectively to emphasize the importance of the feature similarity loss L_S as it is most related to the retrieval performance. These hyper-parameters are fixed in all of our experiments. During training, the model parameters are iteratively updated to minimize the stochastic loss function accumulated over the training data.

The ultimate goal of our feature synthesis model is to synthesize useful features for retrieving relevant images given canvas queries. We therefore explicitly design our loss function to encourage good retrieval performance for each training query.

3.1.1 Similarity Loss

For a given training query Q, let I_Q denote a training image that is relevant to Q (such a query-image pair can be readily obtained from image datasets with available bounding box annotations such as MS-COCO [33] and Visual Genome [29]). We design the similarity loss term L_S to encourage the synthesized feature f_Q to resemble the known visual feature f_{I_Q} extracted from I_Q. Formally, the similarity loss L_S is defined as

$$L_S(f_Q) = 1 - \cos(f_Q, f_{I_Q}) \quad (2)$$

Minimizing this loss function equivalently maximizes the cosine similarity between the feature synthesized from each query and those from its relevant images in the database. These relevant images, as a result, are likely to be highly ranked during retrieval.

Training the feature synthesis model using only the similarity loss, however, is not sufficient. While similarity loss training encourages the relevant features to be learned, it often cannot emphasize the discriminative features that helps distinguish between concepts. As a result, images of irrelevant concepts sharing some visual similarity with the relevant ones may also be ranked high, leading to noisy retrieval results. We propose to address this limitation by incorporating two additional loss functions, namely discriminative loss, and ranking loss.

3.1.2 Discriminative Loss

We incorporate the discriminative loss function to encourage the synthesized features not only to be relevant, but also discriminative with respect to the concepts in the query. A common approach to learn discriminative features is to train

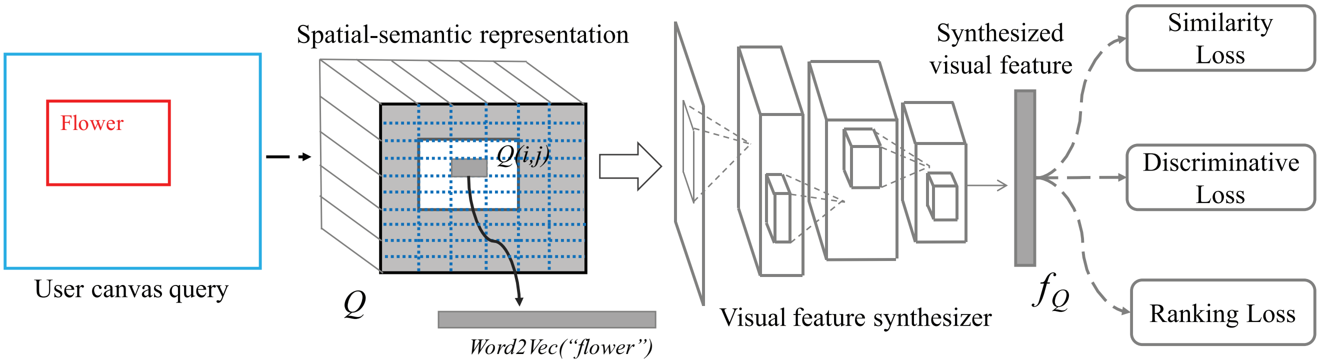


Figure 2: A canvas query is represented in a spatial-semantic representation consisting of a three-dimensional grid Q where $Q(i, j)$ contains the *Word2Vec* semantic vector of the concept appearing at the position (i, j). The grid-based representation of the canvas is then passed through the convolutional feature synthesis network to synthesize the visual feature f_Q for the query. The feature synthesis network is trained to jointly minimize three dedicated loss functions to encourage good retrieval performance at each training query.

a classification model F_D on top of the learned features f_Q. Learning the classification model jointly with the feature synthesis model, however, is problematic as this setup tends to force the synthesized feature to retain mostly semantic information while ignoring useful visual information.

Our idea is to train the classification model F_D with the actual image features f_{I_Q} instead of the synthesized features f_Q and then use it (with fixed weights) to compute the classification loss on the synthesized features during the synthesis model training. As F_D is trained on the real image features for classification, it can encode discriminative visual features for each concept. Using it to guide the training of the feature synthesizer therefore encourages the synthesized features to capture similar discriminative features. The discriminative loss function for a query Q is defined as

$$L_D(f_Q) = CrossEntropy(F_D(f_Q), c_Q) \tag{3}$$

where $F_D(f_Q)$ denotes the class prediction of the classification sub-network F with the synthesized feature f_Q as input. c_Q denotes the concept specified in the query Q. The classification loss is realized by the standard cross-entropy objective function widely used in classification network training.

In our implementation, we implement the classification sub-network F_D as a fully connected layer with 4,096 neurons, each followed by a ReLU activation unit.

3.1.3 Ranking Loss

To further encourage good retrieval performance, we define a ranking loss function which encourages proper ranking for the images given the synthesized features. Following previous works on ranking-based feature learning [21, 53, 54], we define the ranking loss L_R using the triplet loss ranking formulation

$$L_R(f_Q) = \max(0, \alpha - \cos(f_Q, f_{I_Q}) + \cos(f_Q, f_{I_{\bar{Q}}})) \tag{4}$$

where $F_{I_{\bar{Q}}}$ denotes the feature extracted from an image $I_{\bar{Q}}$ which is irrelevant to the query q. α denotes the margin for the ranking loss, which was empirically determined to be 0.35 in our framework. Minimizing this loss encourages the proper ranking of the images given the synthesized features of the queries.

3.2. Implementation Details

For all experiments reported here, we use the features extracted from the fourth inception module of the GoogLeNet network as our target image visual features. In particular, we use the pre-trained GoogLeNet model implemented in Torch[1]. The network was pre-trained on the ImageNet classification task with 1,000 object classes. Our preliminary study shows that this feature is particularly appropriate for our task as it can effectively capture high-level semantic information from the images and at the same time naturally retains most spatial information. We note that our framework is general and can adopt any type of image features. For networks with fully-connected layers (e.g. AlexNet or VGG), we can also use the fully convolutional structure (i.e. the convolutional and pooling layers below the fully connected layers in the network) to retain the spatial information.

We implement our feature synthesis model using a convolutional network architecture with three convolution layers with 3×3 filter size, interleaved by two stride-2 max-pooling layers, each followed by a ReLU activation function and batch normalization [22]. Our network takes as input the canvas cube representation of size $31 \times 31 \times 300$ and output the feature of size $7 \times 7 \times 832$ which is the size of the target GoogLeNet feature layer. The number of feature maps in the two intermediate convolutional blocks are 256 and 512, respectively. The network is trained using the ADAM algorithm [28] for 100,000 iterations with the mini-batch

[1] https://github.com/soumith/inception.torch

size of 17 and initial learning rate of 0.01. To encourage the network to capture the spatial information during training, we mask out the regions in the predicted feature outside of the object regions in the query canvas before computing the loss. After training, our network takes one *ms* to synthesize the feature for one query on an NVIDIA Geforce Titan X GPU. Once the feature is generated, our system can search over 5 millions images in less than a second.

4. Experiments

We evaluate our method on the combination two large-scale datasets MS-COCO [33] and Visual Genome [29]. The MS-COCO dataset contains 123,287 images in its training and validation set with bounding boxes provided for 81 object categories[2]. The Visual Genome dataset contains 108,077 images with provide bounding box annotations for a wide variety of visual concepts[3]. We combine MS-COCO and Visual Genome to obtain a combined dataset with 179,877 images consisting of the bounding box annotation for most image regions, including not only objects but also non-object and background concepts such as sky, grass, beach, and water. The images in the dataset are randomly partitioned into the database set with 105,000 images and the training set with 74,877 images. During training, the training single-concept queries Q are obtained by sampling the bounding boxes in each training image. The image itself serves as the corresponding relevant image I_Q. The irrelevant image $I_{\bar{Q}}$ is randomly picked among the images not containing the concept specified in Q. In principal, we can apply this procedure to train our model with all available concepts in the dataset. However, to ensure each concept has enough data to train, we only sample training queries for the concepts that is contained in at least 1,000 images. As a result, we use the concept list with 269 concepts covering a variety of categories.

Testing queries: Our evaluation requires a large and diverse set of testing queries, which covers a wide variety of concepts with different number of concepts per-query and concepts appear in different sizes. To avoid relying on human efforts in query set creation, which is costly and difficult to control for such a diverse query set, we automatically generate the testing queries from images with their annotated bounding boxes. We randomly select 5,000 images in the database set and create testing queries from them. For each image, we randomly sample its bounding boxes to obtain up to six queries, containing from one to six concept instances. To avoid including many small objects that are insignificant to the scene content, we sample the bounding boxes with the probabilities proportional to their sizes. This process gives us 28,699 testing queries in total.

Spatial-semantic relevance score: the relevance between an input query Q and a retrieved database image I is defined as

$$R(Q, I) = \frac{1}{|B_Q|} \sum_{b_i \in B_Q} \max_{b_j \in B_I} \mathbb{I}(c(b_i) = c(b_j)) \frac{b_i \cap b_j}{b_i \cup b_j} \quad (5)$$

where B_Q and B_I denote the set of annotated bounding boxes in the query Q and the image I, respectively. $\mathbb{I}$ represents the indicator function which takes the value 1 if its argument is true and zero otherwise. $c(b_i)$ and $c(b_j)$ denote the semantic class of the concept represented by the boxes b_i and b_j, respectively. This relevance score evaluates how the retrieved images are relevant to the input queries according to both semantic content and spatial configuration.

4.1. Spatial-Semantic Search Performance

We compare our methods to different approaches for spatial-semantic image search, including the text-based approach, the image-based approach with known image features, and the exemplar-based approach introduced in [59].

Text-based approach: We use the annotation provided with each database image to rank the images according to how many number of concepts specified in the queries contained in the image, regardless of their positions.

Image-based approach: This approach represents each query with a known image features extracted from an example image. We consider an oracle-like approach where the image used to represent each test query is selected to be the ground-truth one that was used to generate that query. This makes a strong baseline as the feature for each query is obtained from the image that is guaranteed to be highly relevant to the query. In particular, we consider two types of features: the GNet-Conv feature extracted at the GoogLeNet's fourth inception convolutional layer (similar to the one used in our model) and the GNet-1024 feature extracted at the layer prior to classification, which forms a feature vector with 1024 dimensions.

Exemplar-based approach [59]: We also experiment with Xu *et al.*'s approach [59] using our own implementation. Affinity Propagation (AP) clustering is first applied to select 6 exemplars for each concept in the query from which visual features are constructed and used to search the database images. The original framework in [59] employs low-level visual features such as SIFT and local color histogram for feature extraction. In this experiment, we also consider a variant where the low-level features are replaced by the local features obtained from the output of the GoogLeNet convolutional layer.

We compare the performance of all methods according to three standard metrics that are widely used in the context of learning-to-rank and information retrieval:

Normalized Discounted Cumulative Gain (NDCG): NDCG is among the most popular evaluation metrics used

[2] http://mscoco.org/

[3] https://visualgenome.org/api/v0/

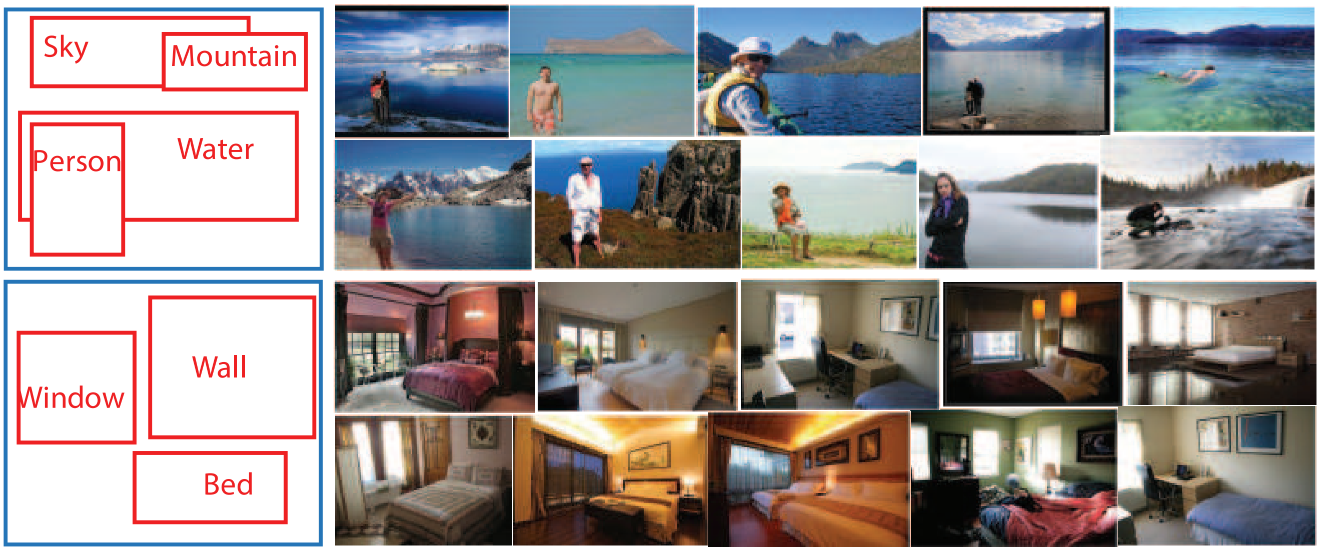
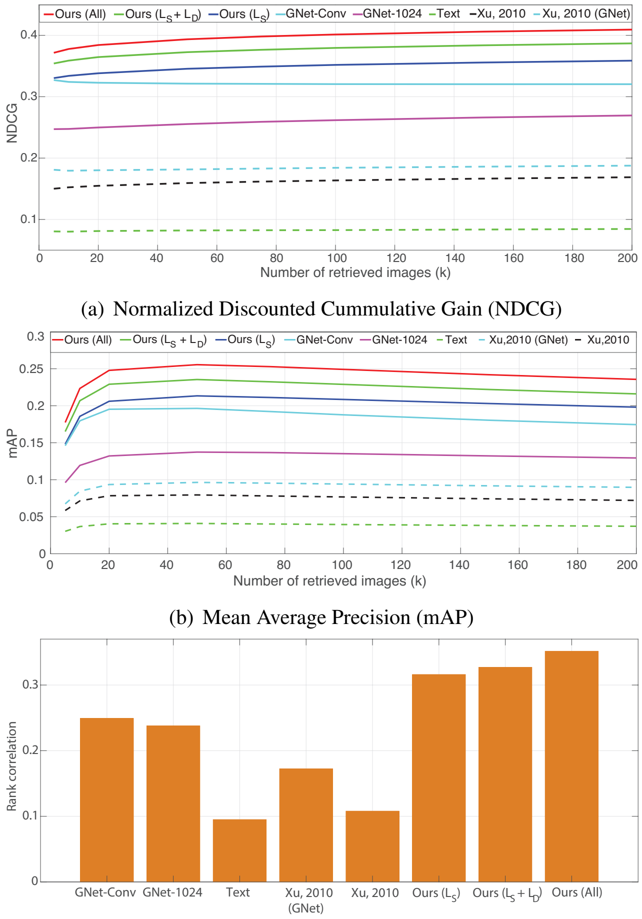


Figure 3: Visual feature synthesis for spatial-semantic image search. Given a user provided canvas query depicting spatial-semantic contraints, we use our feature synthesis model to synthesize the visual features from the canvas query. The synthesized features are used to search against the database image visual features to retrieve images relevant to the query both semantically and spatially.



(a) Normalized Discounted Cummulative Gain (NDCG)

(b) Mean Average Precision (mAP)

(c) Spearman Rank Correlation

Figure 4: Spatial-semantic image search performance.

to evaluate information retrieval systems [7, 24]. NDCG measures the accumulated relevance score obtained by the top-k retrieval results. We compute the NDCG quality for each query and take their average values over all testing queries to obtain the overall performance. Following previous works [34, 59], we compute the NDCG quality for different values of k to obtain the NDCG curves.

Mean Average Precision (mAP): At each rank position in the retrieval result, the precision and recall value are computed according to the spatial-semantic relevance score (Equation 5) by thresholding with a threshold T (we use $T = 0.3$ in our experiments). The average precision is the area under the resulted precision-recall curve. The overall mean average precision (mAP) is computed by accumulating the average precision values over all test queries.

Spearman Rank Correlation: For each query, we obtain the ground-truth ranking for all database images using their relevance scores defined in Equation 5. The quality of each method can then be assessed by the Spearman rank correlation [35, 40] between the ground-truth ranking and the predicted ranking obtained according to the cosine similarity between the query's synthesized feature and the database image features.

Figure 4 compares the spatial-semantic retrieval performance of different methods. As expected, the text-based approach does not perform well for this task as it cannot take into account the spatial information. By constructing visual features from relevant exemplars and capturing spatial information, the exemplar-based approach [59] successfully improve the retrieval performance compared to the text-based approach. Leveraging the deep features from GoogLeNet can further improve the performance.

The Image-based approaches demonstrates good performance as it uses the known image features from the ground-truth image. The results indicate that the features extracted from the convolutional layers perform better than the ones from the later layers. This is due to the spatial information retained at this layer, making it more suitable for our target task of spatial-semantic image search.

Our method performs comparably to the known image features when considering the top-ranked search results and gradually performs better than the image features when the longer rank list is considered. The features extracted from the deep network such as GoogLeNet can faithfully capture both visual and semantic information in the image, which helps retrieve highly relevant images. However, as more images are considered, images with high visual similarity but less spatially and semantically relevant may also be ranked high. Our method explicitly learns the feature synthesis model for the spatial-semantic retrieval task and outperforms other methods according to all evaluation metrics. The results also demonstrate the effects of different loss functions in our framework. Incorporating discriminative loss helps synthesize a more discriminative feature, which improves the performance compared to using only similarity loss. The performance is further improved when all three loss functions are applied. Figure 1 and Figure 3 show retrieval results of our methods for example queries with different concepts and configurations.

4.2. Subjective Evaluation

In addition to the quantitative evaluation based on objective relevance scores and automatically generated test queries, we also investigate the performance of our method on queries constructed by human users through a user study that lets participants create their own spatial-semantic queries and rate the relevance of the retrieval results.

In our study, we recruit 14 participants. Each participant is asked to performs six search sessions. In each session, we let the participant construct the query canvas for a specific target scene she wants to search and then rate the relevance of the search results returned by our method along with two baseline methods, including the text-based baseline, and the exemplar-based baseline with GoogLeNet feature.

Without a specific content to start with, we found it difficult for users to imagine a realistic scene for which relevant images can be found, especially when the database is not too large. Therefore, during the query construction stage, we prompt each participant with a caption obtained from the MS-COCO caption set. That caption serves as a prompt that helps participants easily imagine a realistic scene while arranging the spatial query as they want. We limit the randomly selected prompt captions to those that are at most 15-words long and contain at least one concept among our 269 trained concepts.

After constructing each query, the user is shown the top-20 search results retrieved by all three algorithms, presented in a random order. The user scores each result from 1 to 5 (with 5 be the highest relevance) indicating how relevant it is to the user intent.

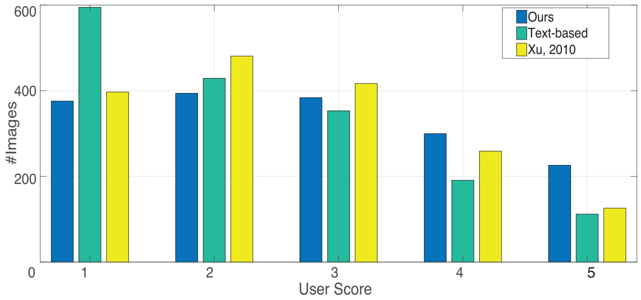


Figure 5: Our spatial-semantic image search method obtains significantly more search results with the high scores, indicating its ability to retrieve results that satisfy the user intent both semantically and spatially.

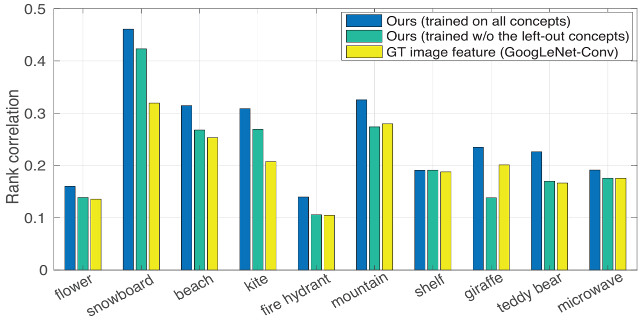


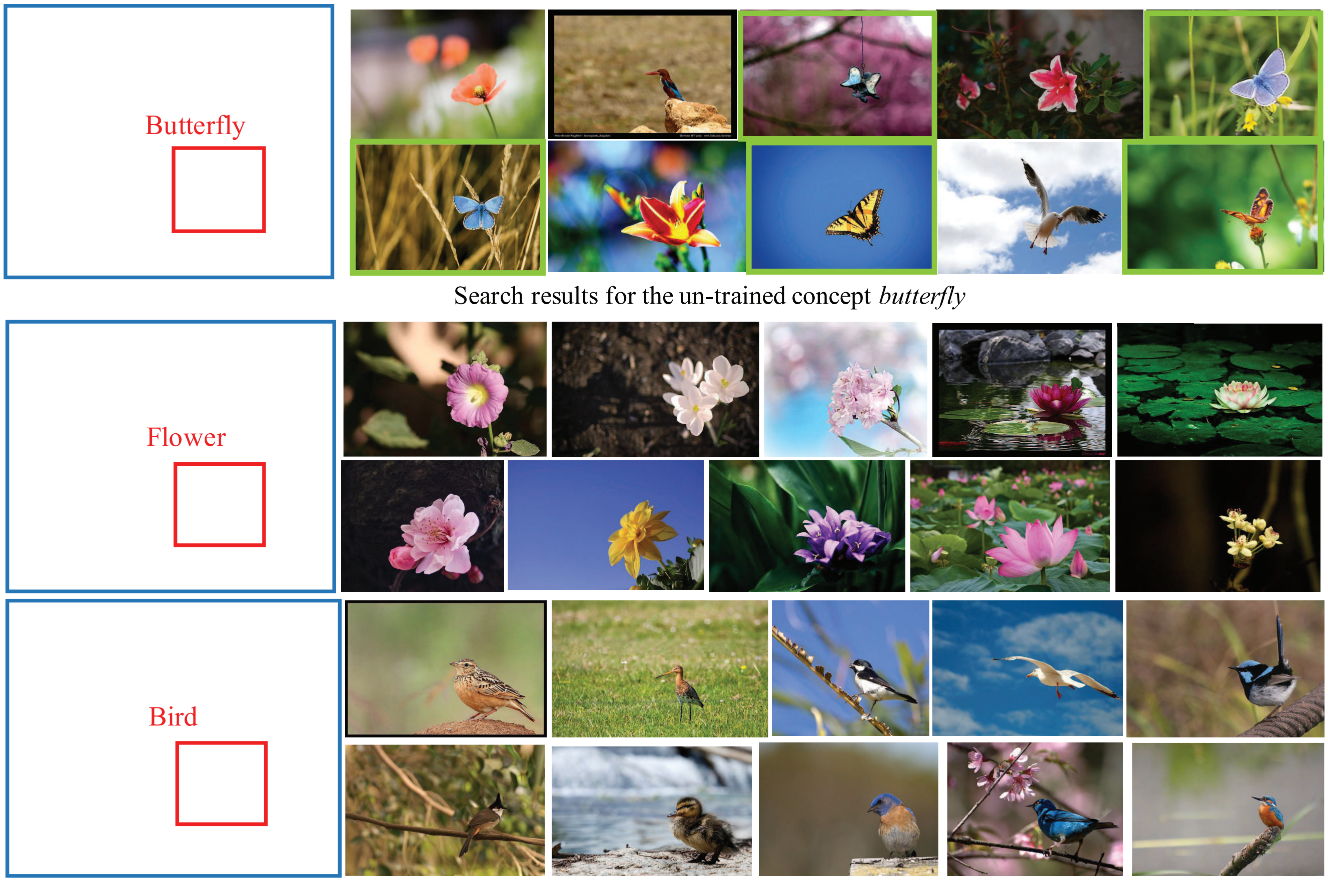
Figure 6: The performance of our model on untrained concepts, while lower than the model trained with all concepts, is in general comparable or better than the ground-truth image features, indicating we can synthesize useful features even for untrained concepts.

Figure 5 depicts a histogram of relevance score values for each of the algorithms. The results indicate that our spatial-semantic image search method obtains significantly more search results with the high relevance scores, which reflects its ability to retrieve relevant images that satisfy the user intent both semantically and spatially.

4.3. Generalization to New Concepts

Our feature synthesis model learns the transformation from each visual concept to the corresponding visual features using spatially and semantically labeled data. It therefore relies on the availability of the concepts in the training data. In this section, we investigate how our model performs when given the hitherto unseen concepts.

Figure 7 provides an example of the query with an untrained concept. In this case, the concept *butterfly* is not part of the concept list used to train the model. The example indicates that our method is able to leverage the knowledge learned from the related trained concepts to synthesize the useful features to retrieve relevant images. Note that while our model was never trained with the *butterfly* concept, it can somewhat leverage the features learned from semanti-



Search results for the un-trained concept *butterfly*

Search results for the related trained concepts: *flower* and *bird*

Figure 7: The untrained concept *butterfly* can be captured by related concepts such as *bird* and *flower* which are similar in the semantic $Word2Vec$ space. As a result, the model can synthesize the useful features from both of these concepts to retrieve several butterfly images (marked with green frames) in the top list.

cally relevant concepts, such as *flower* and *bird* which are close to *butterfly* in the $Word2Vec$ semantic space. With the synthesized feature, several butterfly images were retrieved in the top list. Searching directly with the trained concepts *flower* and *bird* (Figure 7), on the other hand, returned mostly different results. This suggests that our model can combine the knowledge from multiple trained concepts to represent the novel ones instead of merely copying features from the closest trained concept.

To further investigate how our model performs on untrained concepts, we randomly leave out ten selected concepts from our original concept list. We then train our feature synthesis model from the remaining set and test on queries containing the left-out concepts. Figure 6 compares the rank correlation from three methods: our original feature synthesis model trained with all concepts, our model trained without the 10 left-out concepts, and the image-based method with the ground-truth GoogLeNet feature. The result indicates that the performance of our model on untrained concepts, while lower than the model trained with all concepts, is mostly comparable or better than the ground-truth image features, indicating we can synthesize useful visual information even for untrained concepts.

5. Conclusion

This paper presents a data-driven approach to spatial-semantic image search which learns to synthesize visual features directly from the user canvas query using a visual feature synthesis model based on convolutional neural networks. A dedicated training framework with three loss functions is developed to train the feature synthesis model so as to optimize the the spatial-semantic retrieval performance in the visual feature space. Experiments on the combination of the MS-COCO and Visual Genome datasets show that our method can learn an effective representation from the query, which improves the retrieval performance compared to other baseline and state-of-the-art methods. By explicitly learning the mapping from the spatial-semantic representation to the visual representation, our model can exploit the relationship in the semantic space, which enables generalization to novel concepts. In future work, we plan to augment our current framework with additional information such as high-level attributes and hand-drawn sketches to allow more fine-grained search and refinement.

Acknowledgments. This work was partially done when the first author was an intern at Adobe Research. This work was supported in part by NSF IIS-1321119.

References

[1] A. Babenko and V. S. Lempitsky. Aggregating local deep features for image retrieval. In *IEEE International Conference on Computer Vision*, 2015. 2

[2] A. Babenko, A. Slesarev, A. Chigorin, and V. S. Lempitsky. Neural codes for image retrieval. In *European Conference on Computer Vision*, 2014. 2

[3] S. Bell and K. Bala. Learning visual similarity for product design with convolutional neural networks. *ACM Trans. Graph.*, 34(4), July 2015. 2

[4] Y. Cao, H. Wang, C. Wang, Z. Li, L. Zhang, and L. Zhang. Mindfinder: Interactive sketch-based image search on millions of images. In *ACM International Conference on Multimedia*, 2010. 2

[5] F. Carrara, A. Esuli, T. Fagni, F. Falchi, and A. M. Fernández. Picture it in your mind: Generating high level visual representations from textual descriptions. *CoRR*, abs/1606.07287, 2016. 2

[6] S. H. Cooray and N. E. O'Connor. Enhancing person annotation for personal photo management applications. In *International Workshop on Database and Expert Systems Application*, 2009. 1

[7] B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, 1st edition, 2009. 6

[8] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2), May 2008. 1, 2

[9] E. L. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems*, pages 1486–1494, 2015. 2

[10] J. Dong, X. Li, and C. G. M. Snoek. Word2visualvec: Cross-media retrieval by visual feature prediction. *CoRR*, abs/1604.06838, 2016. 2

[11] A. Dosovitskiy, J. T. Springenberg, and T. Brox. Learning to generate chairs with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 3

[12] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollr, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From captions to visual concepts and back. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 1

[13] J. Flynn, I. Neulander, J. Philbin, and N. Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3

[14] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems 26*, pages 2121–2129. 2013. 1, 2

[15] E. D. Gelasca, J. D. Guzman, S. Gauglitz, P. Ghosh, J. Xu, E. Moxley, A. M. Rahimi, Z. Bi, and B. S. Manjunath. Cortina: Searching a 10 million + images database. Technical report, 2007. 1

[16] P. Ghosh, S. Antani, L. R. Long, and G. R. Thoma. Review of medical image retrieval systems and future directions. In *International Symposium on Computer-Based Medical Systems*, 2011. 1

[17] A. Gordo, J. Almazán, N. Murray, and F. Perronnin. LEWIS: latent embeddings for word images and their semantics. In *IEEE International Conference on Computer Vision*, 2015. 2

[18] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. Deep image retrieval: Learning global representations for image search. In *European Conference on Computer Vision*, 2016. 2

[19] A. Gordo, A. Gaidon, and F. Perronnin. Deep fishing: Gradient features from deep nets. In *Proceedings of the British Machine Vision Conference*, 2015. 2

[20] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. In *International Conference on Machine Learning*, pages 1462–1471. JMLR Workshop and Conference Proceedings, 2015. 2

[21] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, 2015. 4

[22] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015. 4

[23] M. Jain, J. C. van Gemert, T. Mensink, and C. G. M. Snoek. Objects2action: Classifying and localizing actions without any video example. In *IEEE International Conference on Computer Vision*, December 2015. 3

[24] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, Oct. 2002. 6

[25] L. Jiang, S.-I. Yu, D. Meng, Y. Yang, T. Mitamura, and A. G. Hauptmann. Fast and accurate content-based semantic search in 100m internet videos. In *International Conference on Multimedia*, 2015. 1

[26] Y. Jing, D. Liu, D. Kislyuk, A. Zhai, J. Xu, J. Donahue, and S. Tavel. Visual search at pinterest. In *ACM International Conference on Knowledge Discovery and Data Mining*, pages 1889–1898, 2015. 1

[27] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Trans. Graph.*, 35(6):193:1–193:10, Nov. 2016. 3

[28] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 4

[29] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 3, 5

[30] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1), 2006. 1

[31] X. Li, S. Liao, W. Lan, X. Du, and G. Yang. Zero-shot image tagging by hierarchical semantic embedding. In *International ACM Conference on Research and Development in Information Retrieval*, pages 879–882, 2015. 2

[32] Y. Li, H. Su, C. R. Qi, N. Fish, D. Cohen-Or, and L. J. Guibas. Joint embeddings of shapes and images via cnn image purification. *ACM Trans. Graph.*, 34(6), Oct. 2015. 2

[33] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision*, 2014. 3, 5

[34] C. Liu, D. Wang, X. Liu, C. Wang, L. Zhang, and B. Zhang. Robust semantic sketch based specific image retrieval. In *IEEE International Conference on Multimedia and Expo*, 2010. 2, 6

[35] T.-Y. Liu. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331, Mar. 2009. 6

[36] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004. 2

[37] C. Lynch, K. Aryafar, and J. Attenberg. Images don't lie: Transferring deep visual semantic features to large-scale multimodal learning to rank. In *International Conference on Knowledge Discovery and Data Mining*, pages 541–548, 2016. 1

[38] S. Lyu, D. Rockmore, and H. Farid. A digital technique for art authentication. *Proceedings of the National Academy of Sciences of the United States of America*, 101(49), 2004. 1

[39] E. Mansimov, E. Parisotto, L. J. Ba, and R. Salakhutdinov. Generating images from captions with attention. *CoRR*, abs/1511.02793, 2015. 3

[40] M. Melucci. On rank correlation in information retrieval evaluation. *SIGIR Forum*, 41(1):18–33, June 2007. 6

[41] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. 2013. 2, 3

[42] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-Shot Learning by Convex Combination of Semantic Embeddings. In *ICLR*, 2014. 2

[43] M. C. Oliveira, W. Cirne, and P. M. de Azevedo Marques. Towards applying content-based image retrieval in the clinical routine. *Future Gener. Comput. Syst.*, 23(3):466–474, 2007. 1

[44] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros. Context encoders: Feature learning by inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3

[45] M. Paulin, M. Douze, Z. Harchaoui, J. Mairal, F. Perronnin, and C. Schmid. Local convolutional features with unsupervised training for image retrieval. In *IEEE International Conference on Computer Vision*, 2015. 2

[46] M. Paulin, J. Mairal, M. Douze, Z. Harchaoui, F. Perronnin, and C. Schmid. Convolutional patch representations for image retrieval: an unsupervised approach. *CoRR*, 2016. 2

[47] Y.-Z. S. T. X. T. M. H. Qian Yu, Feng Liu and C. C. Loy. Sketch me that shoe. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[48] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015. 2

[49] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In *NIPS*, 2016. 3

[50] S. Reed, Z. Akata, B. Schiele, and H. Lee. Learning deep representations of fine-grained visual descriptions. In *IEEE Computer Vision and Pattern Recognition*, 2016. 2

[51] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text-to-image synthesis. In *International Conference on Machine Learning*, 2016. 3

[52] K. Rodden and K. R. Wood. How do people manage their digital photographs? In *Conference on Human Factors in Computing Systems*, CHI '03, 2003. 1

[53] P. Sangkloy, N. Burnell, C. Ham, and J. Hays. The sketchy database: Learning to retrieve badly drawn bunnies. *ACM Trans. Graph.*, 35(4), July 2016. 1, 2, 4

[54] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 4

[55] N. Sebe, M. S. Lew, X. Zhou, T. S. Huang, and E. M. Bakker. *The State of the Art in Image and Video Retrieval*. 2003. 1

[56] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000. 1

[57] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li. Deep learning for content-based image retrieval: A comprehensive study. In *ACM International Conference on Multimedia*, 2014. 1, 2

[58] F. Wang, L. Kang, and Y. Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[59] H. Xu, J. Wang, X.-S. Hua, and S. Li. Image search by concept map. In *International ACM Conference on Research and Development in Information Retrieval*, pages 275–282, 2010. 2, 5, 6

[60] J. Yang, S. E. Reed, M. Yang, and H. Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *Advances in Neural Information Processing Systems*, pages 1099–1107, 2015. 3

[61] Q. Yu, Y. Yang, Y. Song, T. Xiang, and T. M. Hospedales. Sketch-a-net that beats humans. In *Proceedings of the British Machine Vision Conference*, 2015. 2

[62] E. Zavesky and S.-F. Chang. Cuzero: Embracing the frontier of interactive visual search for informed users. In *ACM International Conference on Multimedia Information Retrieval*, pages 237–244, 2008. 2

[63] E. Zavesky, S.-F. Chang, and C.-C. Yang. Visual islands: Intuitive browsing of visual search results. In *International Conference on Content-based Image and Video Retrieval*, pages 617–626, 2008. 2

[64] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In *European Conference on Computer Vision*, 2016. 3