# Cross Domain Image Matching using Shape information

Neha Yadav
College of Information and Computer Sciences
University of Massachusetts Amherst
nyadav@umass.edu

Ashish Singh
College of Information and Computer Sciences
University of Massachusetts Amherst
ashishsingh@cs.umass.edu

## Abstract

*Images of same objects across different visual domains like sketches, paintings, photographs in varied environment, tend to show large difference at raw pixel level.This makes it difficult to match such images. We propose a simple and efficient method to compute cross domain image similarity by utilizing standard object detection framework to find relative importance of features. These weighted features are then used to find image matches, based on visual and contextual similarity. For robust experimentation analysis, we test our approach on UFU-DDD[6], PASCAL-VOC[18], INRIA Holidays[19], Random Flickr images[20] and Sketchy datasets[21].*

## 1. Introduction

Computation of visual similarity across different images is one of the most fundamental tasks of computer vision and pattern recognition, which finds application in many subtasks like Image retrieval, Object detection and Pattern matching. However, this task becomes more challenging and difficult when image to be matched are from different domains.

With present day imaging technologies, it is possible to generate large amount of images containing same object, across different domains (sketches,true photographs, paintings, computer generated pictures) with varied illumination and rendering effects. These images, although being similar at a higher scene level, tend to have very different representation at pixel level.As a result, most of the traditional pixel-wise matching schemes are highly ineffective in accounting for such small perceptual difference at higher visual level.

Finding such visually similar images are critical to many applications like Sketch-Photo face verification, Scene completion, Image retrieval based on sketch input and CG2Real application. To address all such applications, the visual similarity comparison scheme must be invariant to domain-specific information. This further makes the task of cross-domain matching challenging.

Figure 1. Example of expected outcome of Cross Domain Image Matching (a) represents the query; (b) the top 4 predictions

With regard to above mentioned aspects and challenges, our objective in this project is to develop a generalized framework for finding image matches, which have similar salient attributes on a visual level. Not only our framework should be able to identify salient aspects of the object, but should be robust towards small differences that arises due to cross-domain nature of the query images.

Research work done in this field can be broadly categorized into three categories-

- Designing local image descriptors[(self-similarity (SSIM)[1], symmetry [2] GIST[3] ] which are invariant to redundant information and aims at capturing salient aspects of the image.These descriptors encode the unique attributes of the image based on color, shape and texture information, which is then used to compare two images for similarity/dissimilarity using distance metric function.

- Assigning weights to basic image descriptors (SIFT [4], HoG [5], wavelets, etc.) based on globally salient regions of the image, which best defines an object in the image. This can be done in two ways:

  – By using visual attention models[6], which uses intensity contrast and center surround axioms to compute salient regions of an image in a bottom-up approach. The salience image generated is

used to weigh important dimensions of image descriptors.

- By using data-driven paradigm[7] to learn weights over image descriptors by training a linear classifier at query time.

This approach significantly improves performance over uniform-weighted image descriptors by assigning relative importance to feature dimensions based on unique regions of the image, which is common for an object irrespective of the image domain.

Inspired by [7], we propose a generalized image matching framework, utilizing data-based learning for computing weighted image descriptor. As concurred from our initial survey of the related works, we believe that data-driven paradigm assures generalization of image matching task for cross-domain images.By learning weights over image descriptors, we are computing a general template of the object, which successfully captures the salient representation of the image, thus alleviating the need for domain specific information about the object for image matching.
As delineated in [7], we train a linear classifier to learn weights over salient feature dimensions that best discriminates the image at a higher visual level. However, unlike [7], instead of using the linear classifier for classifying objects into classes, we use our model to detect the presence or absence of the object. We then compute similarity metric for all the positive results,to get top matches for our query image.
Extensive experimentation over different domains and comparisons with other approaches demonstrates that our proposed approach performs at par the state of the art methods [6], and even outperform some approaches for specific domains, with respect to accuracy and efficiency.
The remainder of this paper is organized as follows. Sections 2 reviews background and related work. Sect. 3 describes our methodology adopted. Extensive experimental results will be provided in Sect. 4. Discussion regarding the limitations and future work is provided in Sect. 5.

## 2. Related Work

Matching visually similar images is a broad research area under computer vision and pattern recognition, with many methodologies proposed with the global aim of finding similar images or sub-patches. Some of the methods, aimed specifically at content based image retrieval uses image descriptors[8] along with textual information to retrieve 'semantically' similar images.
With respect to image matching across specific domains, many different approaches have been proposed for different applications like: sketches to photographs [9,10], photos under different lighting conditions [11], CG images to photographs [12] and paintings to photographs [13]. However, these methods are domain specific and and cannot be employed as a generalized framework for cross-domain image matching.
For developing a generic approach towards this problem statement, [1] proposed a new image descriptor for computing self-similarity metric, while [6[[7][17] learns weights over image descriptors based on uniqueness described by each feature. Of the above mentioned approaches, [7][17] learns weights using linear classifiers ([7] uses a single image descriptor, while [17] uses multiple image descriptors for image representation.), while [6] uses visual attention model based on intensity and color contrast to compute saliency based weights.

## 3. Approach

To efficiently compute image matches across different domains, we follow a generalized data-driven approach[7], based on which, we learn the optimum set of weights over image descriptor representing our query image, that best discriminates the positive set (query image and its deformed representations) from the negative set. Following the above paradigm, we use Histogram of Oriented Gradients(HoG) template[5] with Support vector machine with linear kernel to learn the optimum weights. We model the learning problem in an exemplar fashion[14], wherein given a large set of negative data, SVM can generalize over a single positive image.
Based on the above setup, we train our linear classifier for each query, in query time, to generate a set of optimum weights for detecting presence or absence of the query object. We further use hard negative mining to optimize the above. After getting a subset of possible similar matches, we rank the given set of images using Structural similarity metric[15].

### 3.1. Image Representation

For cross-domain image matching, shape of the object is one of the most important attribute, as irrespective of the domain representation of a particular object, its structural information should remain approximately same.
Following the above assumption, we utilize Histogram of Oriented Gradients(HoG) template[5] as our image descriptor. HoG features aptly encodes the spatial representation of objects and is invariant to minor changes in alignment of objects.For our approach, we use rigid grid like representation of HoG template with $\approx$ 200 cells, so as to limit the dimensionality of the feature vector to roughly $5K$.

### 3.2. Data Driven exemplar learning framework

As introduced in [7] the problem statement of learning optimal weights to highlight salient feature dimensions can

be formulated as optimization task of minimizing following objective function:

$$L(w_q, b_q) = \sum_{x_i \in I_p \cup I_q} H(w_q^T x_i + b_q) + \sum_{x_j \in I_n} H(-w_q^T x_j - b_q) + \lambda ||w_q||^2$$

where, $w_q$ represents the learned weight vector over $x_i$, the HoG feature template of the images from positive set $I_p \cup I_q$ and negative set $I_n$.

In our framework, the positive set consists of the query image $I_q$ and transformed set of the query image $I_p$ as extra positive data points, formed by applying small transformations(scaling, shifting, aspect ratio). By adding extra positive data points, we intend to form a deformable set of the query image, so as to make our framework robust to small misalignments and scaling. For negative set $I_n$, we subsample images not containing the object represented in the query image.

Based on the above objective function, we use linear SVM with regularization parameter $\lambda = 100$ and loss function as standard hinge loss function $H(x) = max(0, 1-x)$ for detecting the presence or absence of the object in test images. Thus, for optimal weight vector $w_q$, our object detection task can be summarized as

$$\text{F}(\text{I}_q, I_i) = \begin{cases} 1, & \text{if } x_i w_q + b_q \geq Threshold \\ 0, & \text{otherwise} \end{cases}$$

where, $x_i$ is the HoG template of the test image $I_i$

To further optimize our results, we incorporate hard negative mining[16] in our pipeline.This significantly improves our results by constraining the negative set to incorporate only hard negatives. Based n experimentation results, we set an upper bound of 10 iterations to compute hard negatives in the negative set.

After obtaining an initial set of similar images, we further refine this set of probable matches by comparing them with query image using Structural Similarity Index[15] (SSIM). This provides us with a similarity score of each probable match with respect to query image, which is used to rank the obtained matches, with the best match getting a score $\approx 1$. This allows to obtain the best set of similar images without incorporating window based search, which substantially increases the computational speed.

# 4. Experiments

In this section, we perform a number of matching experiments across multiple visual domains. In all experiments, the images were resized to 200 x 200 pixels. We first describe the features used in the experiments, how to generated training and testing data and then present the experimental results.

| Method | Accuracy |
|---|---|
| UFU-DDD | 0.78 |
| UFU-DDD+ 100 RFI | 0.81 |
| UFU-DDD+ 1000 RFI | 0.81 |
| UFU-DDD+ 5000 RFI | 0.82 |
| UFU-DDD+ 20000 RFI+PascalVOC | 0.85 |

Table 1. UFU-DDD: Image to Image Matching

| Method | Accuracy |
|---|---|
| UFU-DDD | 0.67 |
| UFU-DDD+ 100 RFI+INRIA | 0.67 |
| UFU-DDD+ 1000 RFI+INRIA | 0.67 |
| UFU-DDD+ 5000 RFI +INRIA | 0.68 |
| UFU-DDD+ 20000 RFI+INRIA+PascalVOC | 0.69 |

Table 2. UFU-DDD: Painting to Image Matching

| Method | Accuracy |
|---|---|
| Sketchy | 0.55 |
| Sketchy+ 100 RFI+INRIA | 0.55 |
| Sketchy+ 1000 INRIA | 0.56 |
| Sketchy+ 5000 RFI+INRIA | 0.56 |
| Sketchy+ 20000 RF+INRIAI+PascalVOC | 0.61 |

Table 3. Sketchy Database: Sketch to Image Matching

## 4.1. ImageFeature

In our experiments, we first resize the query image heuristically to limit its feature dimensionality ˜ 5000. **HOG:** The HOG feature provides excellent performance for object and human detection tasks. The HOG descriptors are densely extracted with following parameters: *Number of orientation bins* : 9, *Size (in pixels) of a cell*:$[20, 20]$, *Number of cells in each block.*:$[3, 3]$

## 4.2. DataSet Generation

We automated the process of generating the positive and negative dataset.

### 4.2.1 Training DataSet

Our negative data set included images form random flicker images. Postive training data included image from: Sketchy database(sketch and photo from 5 Classes) and UFU-DDD

- **Positive Training Data**: We select a Random Query Image from the specified class of trainning set(UFDD,Sketchy) and generate the transformations of the image to create positive training set.

  *Transformations* We trained our classifier using a single positive Query Image and negative training data,

Figure 2. Top Results of Painting to Image Matching

the classifier was not able to define a margin separating positives and negatives in the test data. We observed classfier was classifying almost everything as negative. To assign more weights to positives and help classfier in define a proper margin to segregrate negative and positive testing data we added transformation of the image. We generated random parameters to perform transformations on the image. This was done to add robustness to model.

- *Cropping*: Crop image around centre at different radii by resizing image to a bigger size while maintaing the aspect ratio or on the original image. To do so we generated list of equally spaced integers and iteratively crop image using those parameter. Cropped image was padded around the center if requiredto match it's size to original Image size. This was required to have same sized feature descriptor for all images.
- *Modifying Aspect Ratio*: Change one dimension of Image, calculate the aspect ratio and add a constant to it and use it calcuate the other dimension of the Image. Image is resized and the then padded or cropped around center to match it's size to original Image size.
- *Translations*: We generated some random translations of the image using affine transformation. The translation parameters were randomized in each iteration while generating the translated image.

- **Negative Training Data**: We start with an initial cache of examples extracted from random flickr image at random and alternated between training a model and updating the cache. In each iteration we removed easy examples from the cache and added new hard examples along with random new negative examples from random flickr image dataset.

#### 4.2.2 Testing DataSet

We used UFU-DDD, Sketchy and PascalVOC(2007),INRIAHolidays(INH),Random Flickr Images(RFI) to generate our test data. PascalVOC(2007 training dataset was labeled. Since UFU-DDD, Sketchy is not labeled so we generated labels for testing data created using images from these sets.

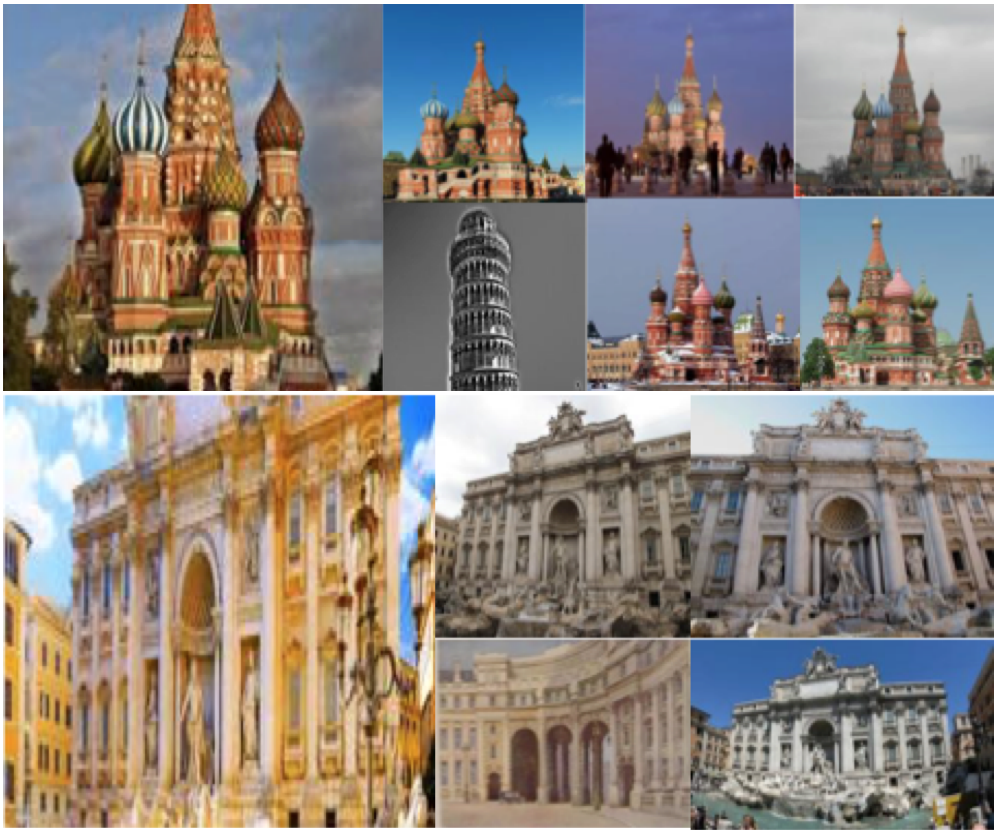

Figure 3. Top Results of Image to Image Matching

- *UFDD*: For Query Image of given class, we copied all positive instance of the Query Class, all images from UFDD test (paintings+Sketch+photo) as one testSet. We then add more random negative instances form flickr images, Pascal Voc and INRIAHolidays(INH) Data set and generated labels for the testSet.

- *Sketchy* For a Query Image of given class, we copied all positive instance of the Query Class labeled them as one.Copied instances from other classes,labeled them one. We then add more random negative instances form flickr images(RFI),INRIAHolidays(INH), Pascal Voc Data set to increase number of negative instances in testing data.

## 4.3. Results

In order to evaluate the performance of our approach, we conducted experiments, each one using a specific domain as the query image. The experiments are run using the UFU-DDD and Sketchy database. We also used images from the random Flickr Images(RFI) database as negatives, in 5 different versions:UFU-DDD/Sketchy, UFU-DDD/Sketchy + 100 (RFI+INH); UFU-DDD/Sketchy + 1,000 (RFI+INH); UFU-DDD/Sketchy + 5,000 (RFI+INH),UFU-DDD/Sketchy + 10,000 (RFI+INH)+10,1000 PASCALVOC.

We had also generated true lables for our testing Data, hence we created a text file containing the predicted images and their corresponding true labels.Since we are applying SSIM on the predicted files to generate scores, we can also refine the results by filtering the images whose true labels matches the predicted label and then measuring the similarity against query image. This is helpful in predicting image similarity amongst those labeled as positvies.

### 4.3.1 Image to Image Matching

The aim of this task is to match photos taken over different ages, seasons, weather or lighting conditions. Our testing set also contains few painting and sketches of the Query Image. We observed that adding negatives negative instances form flickr images(RFI),INRIAHolidays(INH), Pascal Voc imported our accuracy by approx $(3-5\%)$ which shows the robustness of our method. Figure **??** shows some queries and the corresponding top matches for our approach

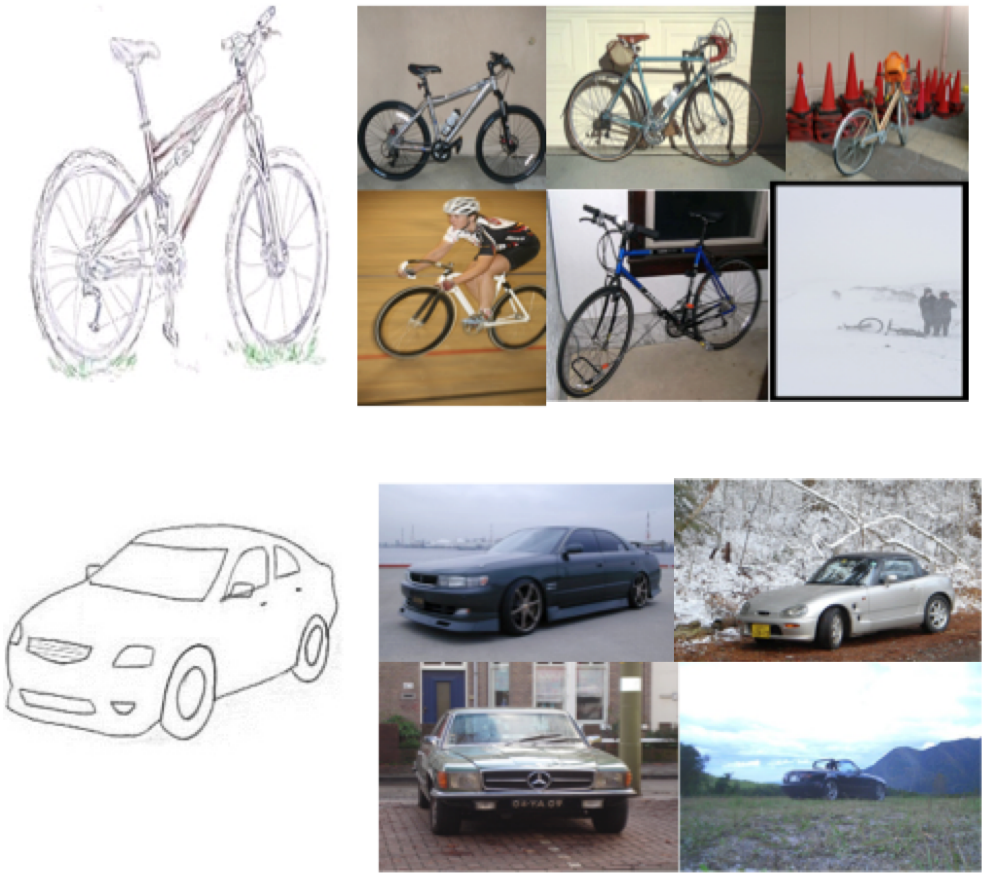

Figure 4. Top Results of Sketch to Image Matching

#### 4.3.2 Painting to Image Matching

To test our approach on another cross domain image matching, we measured the performance of our classifier on matching paitings to images. Retrieval of images similar to paintings is an extremely difficult problem due to presence of strong local gradients due to brush strokes.Figure 2 shows some queries and the corresponding top matches for our approach

#### 4.3.3 Sketch to Image Matching

Matching sketches to images is a difficult cross-domain visual similarity task, as the majority of sketches tend to make use of complex lines rather than just simple contour(s).In addition, sketches also display local deformations and strong abstraction with respect to the real scene. Figure 4 shows some queries and the corresponding top matches for our approach.

## 5. Conclusion

Through this project, we presented a generalized framework for finding similar image matches invarient to domain of the respective images. We investigated data-driven approach towards computing visual similarity between a query and test images of same object in different domain, by learning weights over feature dimesions of the image descriptor, so as to highlight unique aspects of the image. For learning task, we utilize Histogram of oriented gradients along with Support vector machines, modeled in exemplar paradigm by learning weight vector of a single query image, constrained by large number of negative examples.
Our proposed methodology was able to match images from different domains like sketches, paintings and images with varied background conditions with original images with substantial accuracy. By introducing transformed set of query images as extra positive data and Structural similarity index as a measure for similarity, we were able to design a more robust model of the exsisting framework.

However, our methodolygy failed to produce good matches for certain extreme cases, where the query image was highly abstract in nature (Sketches and paintings) and when the same object had very different alignment. Moreover, due to optimization techniques like hard negative mining and computationally complex nature of our approch, the overall time taken to compute similarity is very large, rendering it unviable to be used for certain real time applications.
For improvement of the current model, work can be done on improving the time taken by the model by combining bottom up saliency generation with the exsisting model for fast salient region estimation. Differnt learning techniques like approximate nearest neighbour methods can be used as an alternative to linear SVM

## 6. References


[1]Oliva, A., and Torralba, A. Building the gist of a scene: the role of global image features in recognition. Progress in Brain Research. 155, 23- 36, 2006.

[2] Shechtman, E., and Irani, M. Matching local self-similarities across images and videos. CVPR, 18, 2007

[3]Hauagge, D.C., Snavely, N.: Image matching using local symmetry features. In: IEEE Conference on Comp.

[4]Lowe, D. G. Object Recognition from local scaleinvariant features. ICCV, 1999

[5] Dalal, N., and Triggs, B. Histograms of oriented gradients for human detection.CVPR, 2005

[6] E. V. Melo, S. de Amo and D. Guliato. Cross-domain image matching improved by visual attention. Journal of WSCG, v. 22, 2014.

[7]Shrivastava, A., Malisiewicz, T., Gupta, A., Efros, A.A.: Datadriven visual similarity for cross-domain image matching. ACM Trans. Graph. 30(6), 154 (2011) (SIGGRAPH Asia).

[8] DATTA, R., JOSHI, D., LI, J., AND WANG, J. Z. 2008. Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv..

[9]Cao, Y., Wang, C., Zhang, L., Zhang, L.: Edgel index for largescale sketch-based image search. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 761768 (2011)

[10]Chen, T., Cheng, M.M., Tan, P., Shamir, A., Hu, S.M.: Sketch2Photo: Internet image montage. ACM Trans. Graph. 28(5), 124 (2009)

[11]5. Chong, H.Y., Gortler, S.J., Zickler, T.: A perception-based color space for illumination-invariant image processing. ACM Trans. Graph. 27(3), 61 (2008) (SIGGRAPH)

[12]16. Johnson, M.K., Dale, K., Avidan, S., Pfister, H., Freeman, W.T., Matusik, W.: CG2Real: improving the realism of computer generated images using a large collection of photographs. IEEE Trans. Vis. Comput. Graph. 17(9), 12731285 (2011)

[13]25. Russell, B.C., Sivic, J., Ponce, J., Dessales, H.: Automatic alignment of paintings and photographs depicting a 3D scene. In: 3rd International IEEE Workshop on 3D Representation for Recognition (3dRR) (2011)

[14] MALISIEWICZ, T., GUPTA, A., AND EFROS, A. A. 2011. Ensemble of exemplar-svms for object detection and beyond. In ICCV.

[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error measurement to structural similarity," IEEE Transactions on Image Processing, accepted, May 2003.

[16] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan. Object Detection with Discriminatively Trained Part Based Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, No. 9, September 2010.

[17] Sun, G., Wang, S., Liu, X., Huang, Q., Chen, Y., and Wu, E. Accurate and efficient cross-domain visual matching leveraging multiple feature representations. The Visual Computer, 29, 565-575, 2013.

[18] EVERINGHAM, M., GOOL, L. V., WILLIAMS, C. K. I., WINN, J., AND ZISSERMAN, A., 2007. The PASCAL Visual Object Classes Challenge.

[19] JEGOU , H., DOUZE, M., AND SCHMID, C. 2008. Hamming embedding and weak geometric consistency for large scale image search. In ECCV.

[20] HAYS, J., AND EFROS, A. A. 2008. im2gps: estimating geographic information from a single image. In CVPR.

[21] Patsorn Sangkloy and Nathan Burnell and Cusuh Ham and James Hays,The Sketchy Database: Learning to Retrieve Badly Drawn Bunnies,ACM Transactions on Graphics (proceedings of SIGGRAPH),2016.