

guidelines, respectively. However, in spite of its nature of a ranking problem, to the best of our knowledge none of the existing researches have adopted the *learning-to-ranking* approaches to accomplish this task, which is proven to be useful and widely used in many information retrieval systems. The main goal of this work is thus to study the effectiveness of applying ranking algorithms on image cropping and view finding problems.

We believe that the ability of ranking pairwise views *in the same context* is essential for evaluating photo cropping techniques. Therefore, we build a new dataset consisting of 1,743 images with human labeled crop windows and 31,430 pairs of subviews with visual preference annotations. To obtain quality annotations, we carefully designed an image collection and annotation pipeline which extensively exploited a crowd-sourcing platform to validate the annotated images. We conduct extensive evaluation on traditional approaches and a variety of machine learned rankers trained on the AVA dataset [27] and our dataset with various aesthetic features [26, 9]. Experimental validations show that ranking based image croppers consistently achieve higher cropping accuracy in both the image cropping dataset [37] and our dataset. Additionally, it also suggests that ranking-based algorithms still have great potential to further improve their performance on automatic image cropping with more effective features. The dataset presented in this work is publicly available¹.

2. Previous Work

2.1. Aesthetic Assessment and Modeling

The main goal of aesthetic visual analysis is to imitate human interpretation of the beauty of natural images. Traditionally, aesthetics visual analysis mainly focuses on the binary classification problem of predicting high- and low-quality images [7, 8, 24]. To this end, researchers design various features to capture the aesthetic properties of an image compliant with photographic rules or practices, such as the rule of thirds and visual balance. For example, the spatial distribution of edges is exploited as a feature to model the photographic rule of “simplicity” [17]. Some photo *recomposition* techniques attempt to enhance image composition by rearranging the visual elements [1], applying crop-and-retarget operations [21] or providing on-site aesthetic feedback [38] to improve the aesthetics score of the manipulated image.

Instead of using “hand-crafted” features highly related to the best photographic practices, Marchesotti *et al.* show that generic image descriptors previously used for image classification are also capable of capturing aesthetic properties [26]. In [12], Isola *et al.* show that the *memorability*

of images is predictable by using global image descriptors. In recent years, deep convolutional neural network (DCNN) has been proven to gain tremendous success in various visual recognition tasks and several works also exploited it as the machinery to learn effective features for aesthetics prediction [15, 22, 23]. In [16], Karayev *et al.* compare the performance of different image features for style recognition and show that CNN features generally outperform other features even when trained on object class labels.

2.2. Photo Cropping and View Finding Methods

Generally, automatic photo cropping techniques can be categorized into two lines of researches: attention-based and aesthetics-based approaches. The basic principle of attention-based methods is to place the crop window over the most visually significant regions in an image according to certain attention scores, *e.g.* saliency map [32, 30], or by resorting to eye tracker [29], face detector [42] to find the regions of interest. In [25], a classifier trained on an annotated database for saliency prediction is used to facilitate image thumbnail extraction. Recently, Chen *et al.* [5] conduct a complexity study of several different formulations of optimal window search under the attention-based framework. Although the attention-based approaches can usually determine a crop receiving the most human attention, the cropped images are not necessarily visually pleasing due to little consideration of image composition.

The aesthetics-based methods accomplish the cropping task mainly by analyzing the attractiveness of the cropped image with the help of a quality classifier [28, 10], and are thus closely related to photo quality assessment [7, 8, 24]. Recently, Yan *et al.* [37] proposed a cropping technique that employs features designed to capture the changes between the original and cropped images. In [4], finding good subviews in a panoramic scene is achieved by analyzing the structural features and layout of visual saliency learned from reference images of professional photographs. Several other related works achieve view finding by learning aesthetic features based on position relationships between regions [6] or image decomposition [31].

2.3. Datasets for Computational Aesthetics

Datasets play an important role for computer vision researches since they provide a means to train and evaluate algorithms. There are already several publicly available databases containing aesthetic annotations, such as [7, 17, 24, 27]. Among them, AVA [27] is a large-scale dataset which takes advantage of community-shared data (*e.g.* dpchallenge) to provide a rich collection of aesthetic and semantic annotations. Despite all these efforts, there is still not a standard benchmark for evaluating automatic photo cropping algorithms. In [37], the authors built a dataset consisting of 950 images, which are divided into

¹<https://github.com/yiling-chen/flickr-cropping-dataset>

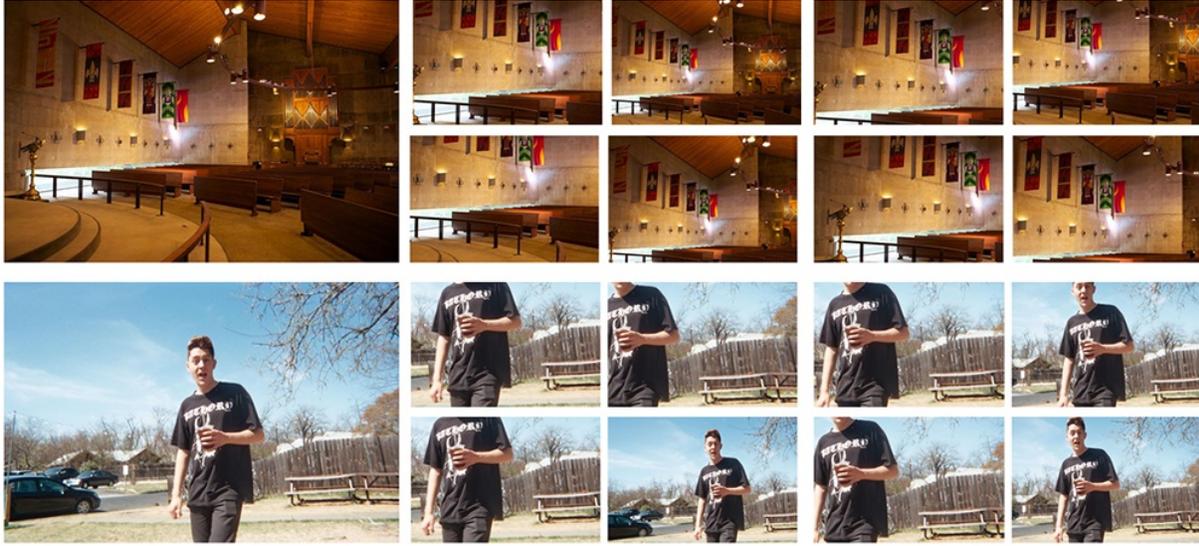


Figure 2. Examples of crop pair generation. On the left are the source images and four corresponding crop pairs are shown on the right. Each pair of crop windows were randomly generated with the guidance of a saliency map to prevent from too much unimportant contents. The aesthetics preference relationship between the crop pairs were determined by the ranking results from AMT workers.

seven categories and individually cropped by three professional experts. We see two deficiencies of this dataset. First, the selected images are from a database originally for photo quality assessment. Some images of professional quality and compositions are also included and cropped. These crops may not faithfully reflect the unwanted regions of the images. Second, many images in the database are *iconic* object or scene images which were taken from a canonical perspective, particularly in the animal, architecture, static categories, which may lack non-canonical views and contextual information. To provide a more general benchmark, we choose to build a new dataset from scratch with a carefully designed image collection and annotation pipeline.

3. Dataset Construction

In this section we describe how the candidate images are selected and the design principles of the annotation pipeline.

3.1. Design Principles

While designing the image annotation procedure, a pilot study was carried out among the authors. We randomly downloaded a small number of test images and had the authors to annotate the ideal crop windows individually. Several observations were obtained after the pilot study.

1. Photo cropping is sometimes very subjective. Particularly, it is extremely difficult to define an appropriate crop for photos of both professional and poor quality since there are no “obvious” answers.
2. Most online images are post-processed which means

that most unwanted regions had been already cut away before they were uploaded. Therefore, it is essential to search for “raw” images for annotation.

3. Sometimes people do agree others’ crops are good even though they are different from their own crops. To obtain quality crops, we decided to resort to a crowdsourcing platform to review all the cropping annotations and adopt only the highly ranked ones as final results.

In manual image cropping, human typically iterates the procedure of moving and adjusting the position, size and aspect ratio of the crop window, and examining the visual quality before and after the manipulation until an ideal crop window is obtained. It is essentially a problem of ranking a number of pairwise subviews that are visually similar. Inspired by the aforementioned process, we also build annotations indicating such preference relationships. We believe that this type of data will be beneficial for researchers to more faithfully evaluating the performance of image cropping techniques.

3.2. Image Collection

In the image collection stage, we aimed to collect as many non-iconic images as possible for better generalization capability [33]. Following the strategy suggested in [20], we chose Flickr as our data source, which tends to have fewer iconic images. In addition, we searched Flickr with many combinations of a pre-defined set of keywords, by which more non-iconic images with richer contextual information are more likely to be returned. The above pro-

cess resulted in an initial set of candidate images consisting of 31,888 images, which were then passed through a data cleaning process. We employed workers on Amazon Mechanical Turk (AMT) to filter out inappropriate images, such as collage, computer-generated images or images with post-processed frames. Particularly, we also asked the AMT workers to pick the photos of excellent quality, because they are potentially not necessary for cropping and thus not suitable for annotation. After data cleaning, 18,925 images remained to enter the next stage of data annotation.

3.3. Image Annotation

We collected two types of annotation through crowdsourcing in our dataset.

- Cropping annotation:** We built a web-based interface for performing the image cropping tasks. The users were recruited from the photography club in our university by invitation. In our task design, we allowed users to skip the images which were judged to be unnecessary for cropping. We eventually retrieved 3,413 cropped images after the human labeling process was finished. For validation, we grouped pairs of cropped image and its corresponding source image as Human Intelligence Tasks (HITs) and assigned each of them to 7 distinct workers on AMT. It is worth noting that a qualification test consisting of 10 pictorial ranking questions was given to each worker. Only the workers who correctly answered at least 8 questions were allowed to take the HITs. For each HIT, the order that the source and cropped image appeared in the HIT was randomized and the workers were asked to pick the more preferable one in terms of their aesthetics. In total, 1,743 out of the 3,413 cropped images were ranked as preferable by at least 4 workers and they constitute the final cropping annotation of our dataset.
- Ranking annotation:** Besides the cropping annotation, we want to enrich the dataset with pairwise ranking relationships between subviews in the same image. For each image with human labeling, 10 pairs of crop windows were randomly generated and then ranked by 5 workers with a similar process as the cropping annotations. To prevent the crop windows from containing too much unimportant contents, we utilized a saliency map [34] to guide crop selection. The size of crop windows varied to imitate the effect of zoom in/zoom out and each pair of crop windows possessed sufficient overlapping. Figure 2 illustrates some examples of the generated crop pairs for ranking. We eventually obtained a collection of totally 34,130 pairs of crop windows with aesthetics preference information. Note that the human cropped images and the corresponding

source images can also be treated as ranking annotations.

To summarize, our dataset is composed of 3,413 cropped images and 34,130 crop pairs generated from the corresponding images. All the source/crop and crop/crop pairs were reviewed by a number of human workers to derive the pairwise aesthetics relationships as the ranking annotation. Finally, 1,743 out of the 3,413 human cropped images were selected as the final cropping annotation of our dataset.

4. Algorithmic Analysis

In this section, we first describe the experimental settings and baseline algorithms, and then demonstrate the experimental validation results.

4.1. Experimental Settings

4.1.1 Datasets

We adopt both the AVA [27] and our dataset to train various image croppers to be compared in this study. The average aesthetic score associated with each image in AVA are used to select a set of high and low quality images to train a photo quality classifier (Section 4.2.2). Additionally, we also exploit the aesthetic scores to form relative ranking constraints to train ranking-based image croppers (Section 4.2.3).

For our dataset, we split the cropping and ranking annotations into training and test set with a roughly 4:1 ratio. Specifically, 348 out of the 1,743 images with highly ranked crops are adopted as ground truth for evaluating the performance of image croppers. The ranking annotations are also used to train ranking-based image croppers (Section 4.2.3).

Finally, the image cropping annotations in [37] is also used to evaluate the performance of image croppers.

4.1.2 Evaluation Protocol

For fair comparison, we take the strategy of evaluating all baseline algorithms on a number of sliding windows. For simplicity, we set the size of search window to each scale among $[0.5, 0.6, \dots, 0.9]$ of the original image and slide the search window over a 5×5 uniform grid. The optimal crop windows determined by image croppers are compared to the ground truth to evaluate their performance.

We adopt the same evaluation metrics as in [37], *i.e.*, *average overlapped ratio* and *average boundary displacement error* to measure the *cropping accuracy* of image croppers. The average overlapped ratio is computed by

$$\frac{1}{N} \sum_{i=1}^N \text{area}(W_i^g \cap W_i^c) / \text{area}(W_i^g \cup W_i^c), \quad (1)$$

where W_i^g and W_i^c denote the ground-truth crop window and the crop window determined by the baseline algorithms

for the i -th test image, respectively. N is the number of test images. The boundary displacement error is given by

$$\sum_{j=\{l,r,b,u\}} \|B_j^g - B_j^c\|/4,$$

where B_i^g and B_i^c denote the four corresponding edges between W_g and W_c . Note that the boundary displacements have to be normalized by the width or height of the original image.

We optionally report the *swap error* evaluated on the test set of AVA and our dataset. It is the ratio of swapped pairs averaged over all queries, which measures the *ranking accuracy* of image croppers to correctly rank pairwise subviews.

4.1.3 Aesthetic Features

For all the learning-based image croppers, we adopt the “deep” activation features [9] to accomplish aesthetics prediction as suggested in [16]. For feature extraction, we exploit the implementation of AlexNet [18] provided by the Caffe library [13]. Each training sample is resized to 227-by-227 pixels and forward-propagated into the network. The activations of the last fully-connected layer are retained as the aesthetic features (DeCAF₇), which are of 4,096-dimension.

We optionally train the ranking-based image croppers with generic image descriptors [26] to inspect the performance variations. Specifically, Fisher vectors of SIFT descriptors with spatial pyramid (SIFT-FV) and Fisher vectors of color descriptors with spatial pyramid (Color-FV) are considered. For SIFT-FV and Color-FV, the cardinality of visual words is 256, and the image descriptor is constructed by concatenating the features extracted from 8 sub-image layouts: “1×1” (whole image), “3×1” (upper, center, bottom), “2×2” (quadrant). The feature points are densely evaluated every 4 pixels, resulting in 262,144-dimension feature vectors.

4.2. Baseline Algorithms

4.2.1 Attention-Based Methods

The first category of methods to be compared are the extension of the *attention-based* photo cropping methods [32, 30], which take advantage of the saliency map accompanying the original image to search for an optimal crop window with the highest average saliency. Instead of the outdated saliency detection methods used in the previous works, we adopt two state-of-the-art methods, *i.e.*, BMS [39] and eDN [34], with leading performance on the CAT2000 dataset from MIT Saliency Benchmark [3]. In addition to the aforementioned search strategy (MaxAvg), we further implement another search criterion, which maximizes the difference of average saliency between the crop

Method	Overlap	Disp.
RankSVM [14]	0.6019	0.1060
RankNet [2]	0.6015	0.1058
RankBoost [11]	0.5017	0.1383
LambdaMART [36]	0.5451	0.1217

Table 1. Benchmarking of various learning-to-rank algorithms. DeCAF₇ feature is used to train the image rankers. The cropping accuracy is evaluated on the 348 test images of our dataset. The best results are highlighted in bold.

window and the outer region of the image (MaxDiff). The saliency maps are generated by the implementation of the original authors with the default parameter settings.

4.2.2 Aesthetics-Based Method

The second category of comparison techniques represent the research line of *aesthetics-based* methods, which exploit a quality classifier that measures whether the cropped region is visually attractive to users [28, 10]. Instead of the low-level features used in the previous works, we adopt the more advanced DeCAF₇ features [9] to achieve aesthetics recognition. A total of 52,000 images with the highest and lowest aesthetics scores are selected from the AVA dataset [27] as the training (67%) and testing (33%) samples. We thus train a binary SVM classifier with RBF kernels, which predicts a photo as high or low quality. The parameters of the classifier are obtained through 5-fold cross validation on the training set and the testing accuracy achieved 80.27%. To use the binary classifier as an image cropper, we take advantage of the method described in [19] to compute the posterior class probability as the aesthetics score to pick the best crop among all candidate windows.

4.2.3 Ranking-Based Methods

The third category of comparison techniques are a family of aesthetics-aware image rankers. To choose an appropriate training algorithm, we have test several pairwise learning-to-ranking algorithms to train the image rankers, including RankSVM [14], RankNet [2], RankBoost [11] and LambdaMART [36]. We exploit the implementation of the above algorithms provided by SVM^{rank2} and RankLib³ libraries for our experiments. The image rankers are trained by using the training set of our dataset with many different configurations of the individual algorithms. The best-performing models are determined by 5-fold cross validation. As summarized in Table 1, RankSVM and RankNet achieve very

²https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

³<https://people.cs.umass.edu/~vdang/ranklib.html>

competitive performance in terms of cropping accuracy. However, since RankNet rankers take much longer time to train, we thus choose RankSVM as the training method for the rest of the experiments in this study. Specifically, all the SVM rankers are trained with a linear kernel and use L1-norm penalty for the slack variables. The loss is measured by the total number of swapped pairs summed over all queries. The parameter C , which controls the trade-off between training error and margin, is determined via 5-fold cross validation.

4.3. Evaluations and Analysis

1) *Comparison of traditional methods:* As shown in Table 2, the first five rows summarize the performances of the four variants of attention-based methods and the aesthetics-based method. One can see that the search strategy of **MaxDiff** consistently outperforms **MaxAvg** for either type of saliency maps. The possible reason is that **MaxDiff** tends to include more salient regions into the crop window in order to lower the total saliency score of the outer region. Unlike **MaxAvg** which usually only concentrates on a single salient region, **MaxDiff** is more likely to obtain a crop window that forms a good composition.

The performance of attention-based methods are highly dependent on the underlying saliency detection scheme. Although **eDN** [34] and **BMS** [39] possess comparable performance in [3], their performance greatly varied in image cropping. It suggests that a standard benchmark is essential to choose the best saliency detection method for automatic image cropping. A hybrid method that optimizes the compositional layout of salient objects might be less sensitive to the selection of saliency maps, such as [42]. In general, attention-based methods performed poorly in determining the aesthetics preferences between crop pairs (45.34% – 63.66% swap error). We believe that this phenomenon could be accounted for the lack of aesthetics considerations in this family of methods. Note that the swap errors are calculated by the attention scores received by the crop pairs.

Comparing with attention-based methods, the aesthetics-based method (**SVM+DeCAF₇**) achieved better performance in all evaluation metrics. However, although the SVM classifier showed good capability of predicting high- and low-quality images, it did not perform well in ranking pairwise views (*i.e.*, 42% swap error), resulting in moderate performance in image cropping accuracy.

2) *Comparison of various aesthetic features:* The 9-th to 11-th rows of Table 2 compare the performance of image rankers trained by different aesthetic features using our new dataset. **DeCAF₇** achieves the best accuracy in all metrics. This result is consistent with the findings reported by [16], *i.e.*, **DeCAF₇** generalizes well to other visual recognition tasks even though the DCNN was trained for object

Method	Overlap	Disp.	Swap
eDN (MaxAvg)	0.3573	0.1729	0.6366
eDN (MaxDiff)	0.4857	0.1372	0.4534
BMS (MaxAvg)	0.3427	0.1815	0.5775
BMS (MaxDiff)	0.3905	0.1674	0.4962
SVM+DeCAF₇	0.5154	0.1325	0.4201
AVA 1-1+DeCAF₇	0.5223	0.1294	0.1317
AVA 2-2+DeCAF₇	0.5069	0.1346	0.2379
AVA 5-5+DeCAF₇	0.4931	0.1384	0.2775
Our+SIFT-FV	0.5917	0.1084	0.3068
Our+Color-FV	0.5042	0.1405	0.3692
Our+DeCAF₇	0.6019	0.1060	0.3225

Table 2. Summarization of performance evaluation. The middle two columns measure the cropping accuracy on the 348 testing images of our dataset. The best results are highlighted in bold.

classification. Although **SIFT-FV** achieves comparable cropping accuracy with **DeCAF₇**, the later obviously provides a much more compact feature representation of visual aesthetics.

3) *Comparison of different datasets:* In this experiment we examine the effectiveness of training image rankers on the AVA dataset [27]. Same as the aesthetics-based method, 52,000 images with the highest and lowest aesthetics scores are first selected. A configuration of **AVA n-n** means that we repeatedly select n images from the high- and low-ranked group, respectively, and generate all combinations of the selected images to form the ranking constraints. Note that the characteristics of pairwise ranking constraints formed by AVA and our dataset are very different since AVA differentiates the visual preferences *between* distinct images while our dataset ranks visually similar subviews *within* the same images.

Row 6-8 in Table 2 give the performances of three rankers trained on AVA using **DeCAF₇** feature. **AVA 1-1** performs best both in cropping and ranking accuracy. However, surprisingly, increasing ranking constraints (**AVA 2-2** and **AVA 5-5**) caused the performance to considerably drop instead. It indicates that only a sparse set of pairwise ranking constraints defined by the aesthetic scores are useful for image ranking and naively pairing images would not improve the ranking accuracy. Besides, although **AVA n-n+DeCAF₇** rankers generally outperform the traditional methods in ranking accuracy, it does not reflect on their cropping capability. For example, the cropping accuracy of **AVA 1-1+DeCAF₇** outperforms the best-performing traditional method (*i.e.*, **SVM+DeCAF₇**) with only an insignificant margin even though it has a much greater ranking accuracy. One possible reason is that the training data of AVA do not reflect the visual preference among visually similar views, which is essential for image cropping.

(a)		
Method	Overlap	Disp.
eDN (MaxDiff)	0.4636	0.1578
SVM+DeCAF ₇	0.5005	0.1444
AVA 1-1+DeCAF ₇	0.5142	0.1399
Our+DeCAF₇	0.6643	0.092

(b)		
Method	Overlap	Disp.
eDN (MaxDiff)	0.4399	0.1651
SVM+DeCAF ₇	0.4918	0.1483
AVA 1-1+DeCAF ₇	0.5034	0.1443
Our+DeCAF₇	0.6556	0.095

(c)		
Method	Overlap	Disp.
eDN (MaxDiff)	0.437	0.1659
SVM+DeCAF ₇	0.4882	0.1491
AVA 1-1+DeCAF ₇	0.4939	0.147
Our+DeCAF₇	0.6439	0.099

Table 3. Cross dataset validation. (a)-(c) summarize the cropping accuracy of the best performing image croppers of each category shown in Table 2, which are evaluated on the three different sets of annotations in the database of [37]. The best results are highlighted in bold.

Such observation can be further validated by comparing to the rankers trained on our dataset. **Our+DeCAF₇** achieves significant improvement in cropping accuracy using the same feature. It is also interesting to note that **Our+DeCAF₇** does not perform well in its ranking accuracy. The reason for the low ranking accuracy could be explained as follows: Since **DeCAF₇** is trained for the purpose of object recognition, it is thus very likely that the **DeCAF₇** features extracted from similar views containing the same “object” to be also similar. The same phenomenon can also be observed in other aesthetic features, *i.e.*, **SIFT-FV** and **Color-FV**. It suggests that there is still great potential to improve ranking-based image croppers by jointly learning the feature representation and semantically meaningful embedding of image similarity with DCNN [35] instead of directly using **DeCAF₇**.

4) *Cross-dataset validation*: In this experiment, we select the best performing image croppers from each category shown in Table 2 and directly apply them on the image cropping dataset of [37]. This dataset is composed of 950 images, which were annotated by three different users. Since this dataset contains only cropping annotations, it is thus only used to evaluate the cropping accuracy. A similar sliding window approach as described in Section 4.1.2 is adopted for evaluation. As shown in Table 3, **Our+DeCAF₇** consistently achieves the highest accuracy in all annotation sets, which further validates the effectiveness of ranking pairwise subviews in image cropping. Note

that higher cropping accuracy is reported in [37]. Since this is a comparative study, no optimization on the parameters of crop windows (*i.e.*, x , y , w and h) was performed for fair comparison. We believe that the performance of the ranking based image croppers can be further enhanced by incorporating appropriate crop selection procedures.

To summarize, the findings of this study lead to the most important insight of this work: *ranking pairwise views is crucial for image cropping*. To the best of our knowledge, all existing methods attempted to tackle this problem by visual saliency detection or learning an aesthetics-aware model from *distinct* images. However, according to our experimental study, these approaches do not necessarily perform well in differentiating pairwise views with substantial overlaps, which is crucial for image cropping. Figure 3 demonstrates several examples of comparing the ground truth and the best crop windows determined by various methods. To maximize the performance of machine learned image rankers, two possible directions can be considered: 1) adopting more effective feature representations learned from pairwise ranking of image subviews; 2) developing effective crop selection method to determine potentially good candidate windows for image ranking.

5. Conclusions

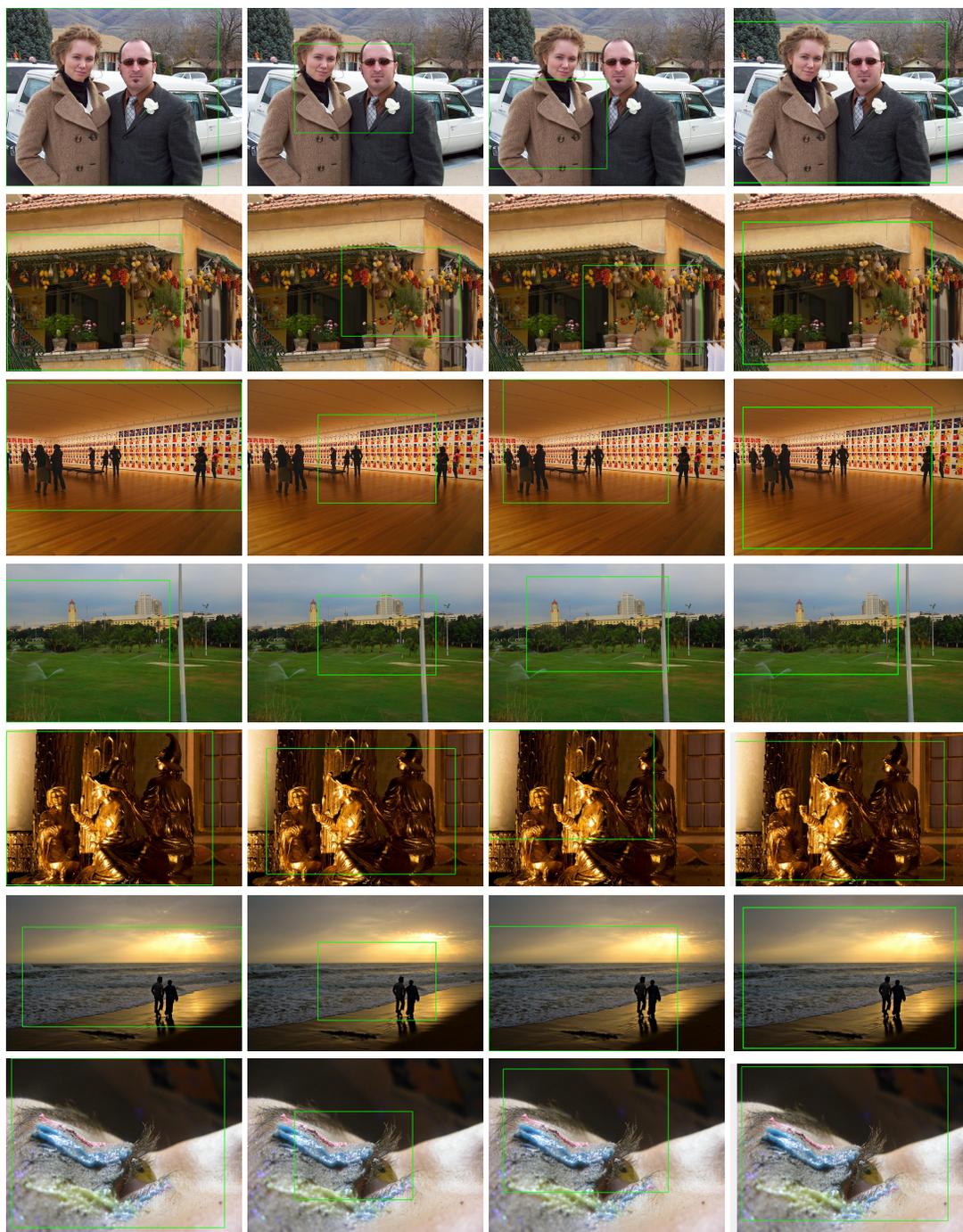
In this paper, we presented a new dataset which aims to provide a benchmarking platform for photo cropping and view finding algorithms. With carefully designed data collection pipeline, we were able to collect high quality annotations. One significant difference between our dataset and other databases is the introduction of pairwise view ranking annotations. Inspired by the procedure of iteratively comparing in manual image cropping, we argue that learning-to-rank approaches possess great potential in this problem domain, which have been overlooked by most previous researchers. We conducted extensive study on evaluating the performances of traditional image cropping techniques and several machine learned image rankers. The experimental results showed that image rankers trained on pairwise view ranking annotations outperform the traditional methods.

Acknowledgement

This work was supported in part by Ministry of Science and Technology, National Taiwan University and Intel Corporation under Grants MOST-105-2633-E-002-001 and NTU-ICRP-105R104045. T.-W. Huang and H.-T. Chen are partially supported by MOST 103-2221-E-007-045-MY3.

References

- [1] S. Bhattacharya, R. Sukthankar, and M. Shah. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *ACM Multimedia*, 2010. 2



(a) Ground Truth

(b) eDN(MaxDiff)

(c) SVM+DeCAF₇

(d) Our+DeCAF₇

Figure 3. Example image cropping results. The optimal crop windows determined by various baselines are drawn as green rectangles.

- [2] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *ICML*, pages 89–96, 2005. [5](#)
- [3] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. MIT Saliency Benchmark. [5](#), [6](#)
- [4] Y.-Y. Chang and H.-T. Chen. Finding good composition in

panoramic scenes. In *IEEE ICCV*, 2009. [1](#), [2](#)

- [5] J. Chen, G. Bai, S. Liang, and Z. Li. Automatic image cropping : A computational complexity study. In *IEEE CVPR*, pages 507–515, 2016. [2](#)
- [6] B. Cheng, B. Ni, S. Yan, and Q. Tian. Learning to photograph. In *ACM Multimedia*, pages 291–300, 2010. [1](#), [2](#)

- [7] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, pages 288–301, 2006. 2
- [8] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *IEEE CVPR*, pages 1657–1664, 2011. 2
- [9] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *ICML*, pages 647–655, 2013. 2, 5
- [10] C. Fang, Z. Lin, R. Mech, and X. Shen. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *ACM Multimedia*, pages 1105–1108, 2014. 1, 2, 5
- [11] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, Dec. 2003. 5
- [12] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. What makes a photograph memorable? *IEEE TPAMI*, 36(7):1469–1482, 2014. 2
- [13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, pages 675–678, 2014. 5
- [14] T. Joachims. Training linear SVMs in linear time. In *ACM KDD*, pages 217–226, 2006. 5
- [15] L. Kang, P. Ye, Y. Li, and D. Doermann. Convolutional neural networks for no-reference image quality assessment. In *IEEE CVPR*, pages 1733–1740, 2014. 2
- [16] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller. Recognizing image style. In *BMVC*, pages 1–20, 2014. 2, 5, 6
- [17] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *IEEE CVPR*, pages 419–426, 2006. 2
- [18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012. 5
- [19] H.-T. Lin, C.-J. Lin, and R. C. Weng. A note on Platt’s probabilistic outputs for support vector machines. *Machine Learning*, 68(3):267–276, 2007. 5
- [20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2014. 3
- [21] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or. Optimizing photo composition. *Computer Graphics Forum (Proc. of Eurographics ’10)*, 29(2):469–478, 2010. 2
- [22] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. RAPID: Rating pictorial aesthetics using deep learning. In *ACM Multimedia*, pages 457–466, 2014. 2
- [23] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *IEEE ICCV*, pages 990–998, 2015. 2
- [24] W. Luo, X. Wang, and X. Tang. Content-based photo quality assessment. In *IEEE ICCV*, pages 2206–2213, 2011. 2
- [25] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *IEEE ICCV*, pages 2232–2239, 2009. 1, 2
- [26] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *IEEE ICCV*, pages 1784–1791, 2011. 2, 5
- [27] N. Murray, L. Marchesotti, and F. Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *IEEE CVPR*, pages 2408–2415, 2012. 2, 4, 5, 6
- [28] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato. Sensation-based photo cropping. In *ACM Multimedia*, pages 669–672, 2009. 1, 2, 5
- [29] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen. Gaze-based interaction for semi-automatic photo cropping. In *ACM CHI*, pages 771–780, 2006. 2
- [30] F. Stentiford. Attention based auto image cropping. In *ICVS Workshop on Computation Attention and Applications*, 2007. 1, 2, 5
- [31] H.-H. Su, T.-W. Chen, C.-C. Kao, W. H. Hsu, and S.-Y. Chien. Preference-aware view recommendation system for scenic photos based on bag-of-aesthetics-preserving features. *IEEE Transactions on Multimedia*, 14(3-2):833–843, 2012. 1, 2
- [32] B. Suh, H. Ling, B. B. Bederson, and D. W. Jacobs. Automatic thumbnail cropping and its effectiveness. In *ACM UIST*, pages 95–104, 2003. 1, 2, 5
- [33] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *IEEE CVPR*, pages 1521–1528, 2011. 3
- [34] E. Vig, M. Dorr, and D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *IEEE CVPR*, pages 2798–2805, 2014. 4, 5, 6
- [35] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *IEEE CVPR*, pages 1386–1393, 2014. 7
- [36] Q. Wu, C. J. Burges, K. M. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Journal of Information Retrieval*, 13(3):254–270, June 2010. 5
- [37] J. Yan, S. Lin, S. B. Kang, and X. Tang. Learning the change for automatic image cropping. In *IEEE CVPR*, pages 971–978, 2013. 1, 2, 4, 7
- [38] L. Yao, P. Suryanarayan, M. Qiao, J. Z. Wang, and J. Li. Oscar: On-site composition and aesthetics feedback through exemplars for photographers. *International Journal of Computer Vision*, 96(3):353–383, 2012. 2
- [39] J. Zhang and S. Sclaroff. Saliency detection: A boolean map approach. In *IEEE ICCV*, pages 153–160, 2013. 5, 6
- [40] L. Zhang, Y. Gao, R. Ji, Y. Xia, Q. Dai, and X. Li. Actively learning human gaze shifting paths for semantics-aware photo cropping. *IEEE Transactions on Image Processing*, 23(5):2235–2245, 2014. 1
- [41] L. Zhang, M. Song, Q. Zhao, X. Liu, J. Bu, and C. Chen. Probabilistic graphlet transfer for photo cropping. *IEEE Transactions on Image Processing*, 22(2):802–815, 2013. 1
- [42] M. Zhang, L. Zhang, Y. Sun, L. Feng, and W. Ma. Auto cropping for digital photographs. In *ICME*, pages 2218–2221, 2005. 2, 6