Learning to Compose with Professional Photographs on the Web

Anonymous Authors(s)
Paper ID: 37 (Supplemental Material)

ACM Reference format:
Anonymous Authors(s). 2017. Learning to Compose with Professional Photographs on the Web. In *Proceedings of ACM Multimedia conference, Mountain View, California USA, October 2017 (MM'17),* 8 pages.
DOI: 10.475/123_4

In this document, we describe the procedures of training the baselines in more details and provide more evaluation results to supplement the exposition of the paper. Additionally, more results of image cropping are also presented.

1 BASELINE ALGORITHMS

eDN. This baseline method serves as the representative of the attention-based image cropping algorithms. In this baseline system, we employ the saliency map generated by [11], which is one of the best performing algorithms in the MIT Saliency Benchmark [1], to compute the attention score for each candidate window. The attention score of a crop is defined by the average intensity value of the corresponding region in the sacliency map. We use the implementation provided by the authors and adopt the default settings to generate the saliency maps. Several examples of the saliency maps generated by [11] are illustrated in Figure 1.

AlexNet_finetune. This baseline method serves as the representative of the aesthetics-based image cropping algorithms. It is obtained by fine-tuning AlexNet [6] with the AVA dataset [9] for binary aesthetics classification. The AVA dataset contains more than 250,000 images, which are originally downloaded from the `dpchallenge` website, and each image is associated with about 200 ratings ranging from 1 to 10. Following the experimental configurations of [5, 7, 8], we split the AVA dataset into about 230,000 training and 20,000 test images. When fine-tuning AlexNet, the training set is divided into two classes (high- and low-quality images) with an average score threshold of 5. The resulting model achieved 72.67% accuracy in the test set.

RankSVM+AVA. This baseline method is a ranking-based image cropper. Following [2], we adopt ranking SVM [4] as a baseline of image croppers. To train this baseline, we selected the highest 26,000 and lowest 26,000 ranked images from AVA [9]. One image is randomly selected from each pool to form a ranking constraint without replacement. The SVMs are trained by using *deep convolution activation* features [3]. Specifically, the training images are fed into AlexNet (pre-trained with ImageNet dataset [10]) and the activation of the last fully-connected layer are extracted as the aesthetic features (**DeCAF**$_7$), which are of 4,096-dimension. We exploit the SVM^{rank}[1] library to train our models. 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 grid search.

[1]https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

MM'17, Mountain View, California USA
2017. 123-4567-24-567/08/06…$15.00
DOI: 10.475/123_4

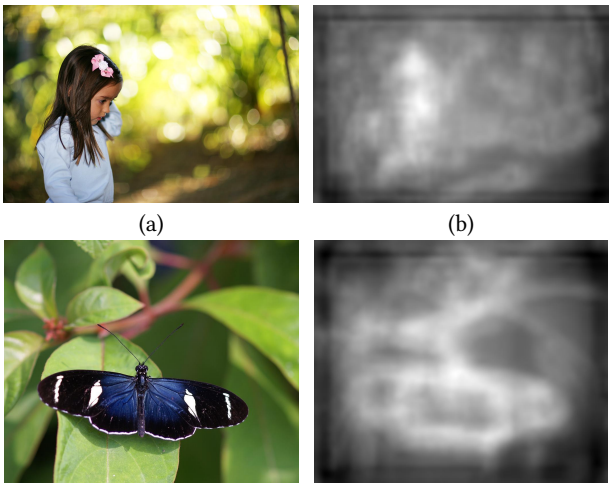


Figure 1: Examples of saliency map generation, (a)(c) source image, (b)(d) saliency map [11]. Best viewed in color.

RankSVM+FCDB. is another variant of ranking-based image cropper. The training procedure of this baseline is the same as RankSVM+AVA except that the training data is from FCDB [2]. The training/test set of FCDB includes 3,065/348 images, each of which are associated with 10 pairs of random crops. The relative aesthetic ranking of each image pair is rated by 5 Mechanical Turk workers. We adopt the training set of FCDB to train our SVM rankers, *i.e.*, 30,650 image pairs in total.

2 ADDITIONAL RESULTS

2.1 Comparison of Different Sampling Strategies

We conduct experiments to investigate the influence of different crop sampling configurations. To this end, we train VFN (SPP-Max) with crop pairs generated by the procedure described in Section 3.1 (main text) while varying a parameter, *e.g.*, scale of crops and degree of randomization. We then evaluate the corresponding cropping accuracy on FCDB. Note that we used roughly 17,000 professional images to conduct the experiments presented in this section.

The first experiment compares the effect of training VFN using various scales of crop pairs. Table 1 and 2 summarize the performance of VFNs trained by a single scale of *border* and *square* crops, respectively. One can see that generally VFNs trained by smaller

Setting	IoU	Disp.	α-recall
$s = 0.5$	0.639	0.0959	**28.4483**
$s = 0.6$	**0.642**	**0.0949**	25.8621
$s = 0.7$	0.6292	0.0988	24.4253
$s = 0.8$	0.6033	0.1087	23.2759

Table 1: Performance evaluation of VFN using border crops of different scales. The best results are highlighted in bold.

Setting	IoU	Disp.	α-recall
$s = 0.5$	**0.6267**	**0.0988**	23.8506
$s = 0.6$	0.6216	0.1012	**24.4253**
$s = 0.7$	0.5989	0.1075	20.977
$s = 0.8$	0.5795	0.1139	19.5402

Table 2: Performance evaluation of VFN using square crops of different scales. The best results are highlighted in bold.

crops performs better. Intuitively, this phenomenon could be accounted for that larger crops are closer to the original image and thus less likely to destroy the original composition. Similar results can also be observed from VFNs trained by square crops of single scale (please refer to supplemental material). This justifies the reason to generate crops of s ={0.5, 0.6} as described in Section 4.1 (main text).

In the second experiment, VFNs are trained with the entire dataset (both border and square crops with s ={0.5, 0.6}). However, the generated crops undergo various degrees of randomization. As shown in Table 3, slightly randomizing crop generation improves the performance of VFN because it enriches the diversity of training data. On the other hand, increasing the degree of randomization does not necessarily further enhance the performance. This phenomenon could be explained as follows: although the spatial placement of border and square crops is designed to minimize mutual overlap, large random perturbation increases the chance for the generated crops to largely coincide with each other, causing redundancy in the training data. In the remaining experiments, we empirically limit the random shift imposed on a crop to be no greater than 20% of the *movable* range in the image.

Our third experiment evaluates the impact of the size of training data. We randomly selected 40%, 60% and 80% of training images and use the corresponding crop pairs to train VFN. As shown in Table 4, although it is reasonable that the VFN trained on all training images achieves the highest performance, it is worth noting that the other models achieve comparable performance with significantly less training data. VFN can be regarded as fine-tuning a pre-trained AlexNet with a ranking loss, which reduces the necessity of massive training data. Thanks to the low cost of expanding the training data, the performance of VFN can potentially be further boosted by adding more training images.

2.2 More Examples

Figure 2 to 7 show more examples of comparing the cropping results of various baseline methods with VFN on the FCDB dataset [2].

Setting	IoU	Disp.	α-recall
No randomization	0.6608	0.0906	31.0345
Small randomization	**0.6741**	**0.0868**	**33.046**
Large randomization	0.6617	0.0899	28.7356

Table 3: Performance evaluation of VFN using crops of various degrees of randomization. The best results are highlighted in bold.

Setting	IoU	Disp.	α-recall
40% training images	0.6391	0.0968	24.71
60% training images	0.6523	0.0923	28.609
80% training images	0.6595	0.0909	29.3103
All training images	**0.6741**	**0.0868**	**33.046**

Table 4: Performance evaluation of VFN using crops of various amounts of training images. The best results are highlighted in bold.

Note that in many examples, VFN selects the ground truth as the best crop.

We also briefly describe the procedure to generate the view finding examples in panoramic scenes shown in Figure 6 of the main text. In these examples, the best view is selected among a set of sliding windows. Denote the width and height of a panoramic image as W and H, respectively. We set the width and height of a crop C as the fraction of the original size, *i.e.*, $w = sW, h = sH, s \in \{0.4, 0.5, 0.6, 0.7\}$, resulting in 16 different aspect ratios. Each crop window slides with a horizontal and vertical step of $0.05W$ and $0.05H$, respectively. There are roughly 2,100 search windows in total. The best crop windows are shown in yellow rectangles in the left column of Figure 6.

REFERENCES

[1] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. MIT Saliency Benchmark. (????). http://saliency.mit.edu/index.html

[2] Yi-Ling Chen, Tzu-Wei Huang, Kai-Han Chang, Yu-Chen Tsai, Hwann-Tzong Chen, and Bing-Yu Chen. 2017. Quantitative Analysis of Automatic Image Cropping Algorithms: A Dataset and Comparative Study. In *WACV*.

[3] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2013. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *ICML*. 647–655.

[4] Thorsten Joachims. 2006. Training Linear SVMs in Linear Time. In *ACM KDD*. 217–226.

[5] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. 2016. Photo Aesthetics Ranking Network with Attributes and Content Adaptation. In *ECCV*.

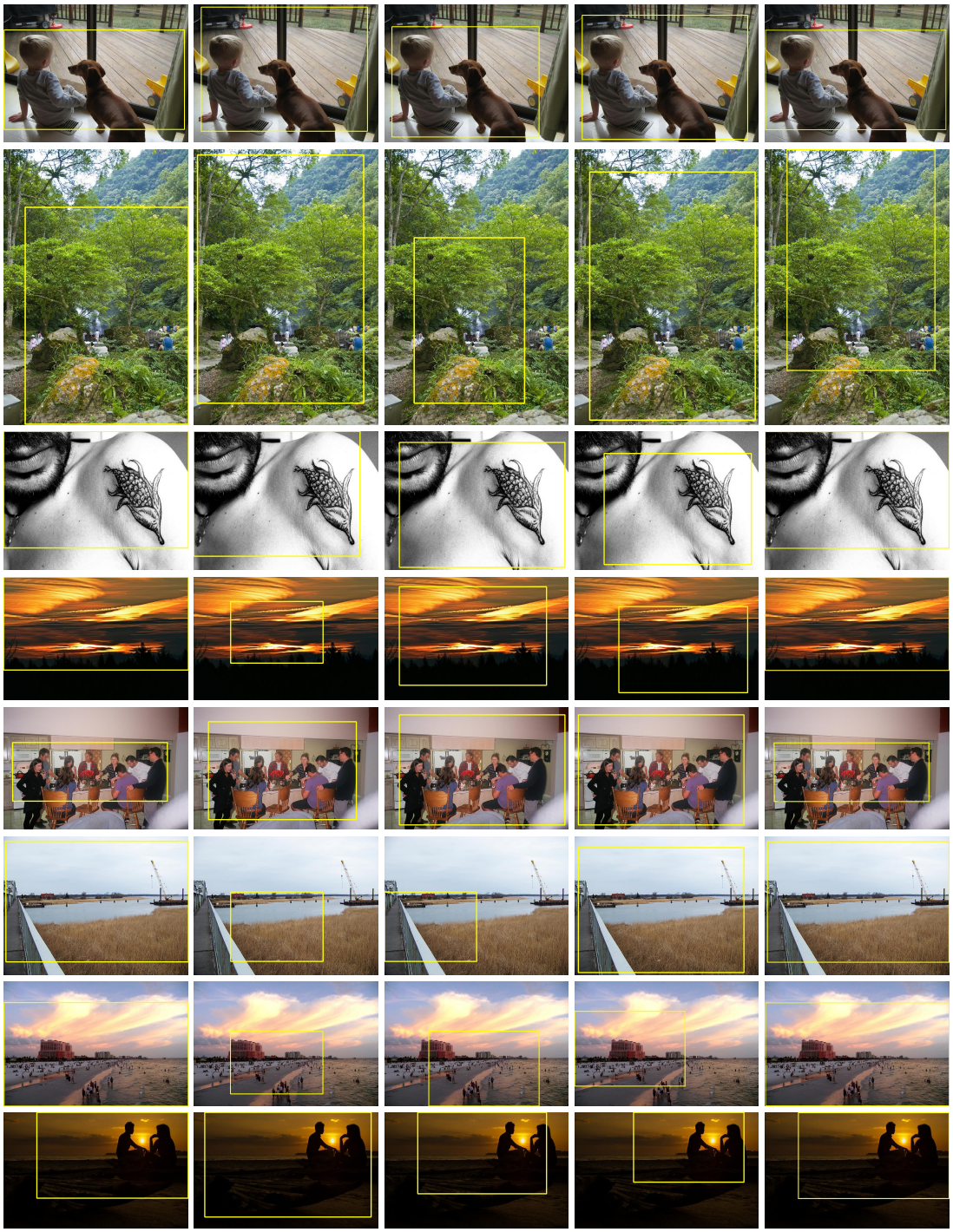
[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*. 1106–1114.

[7] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z. Wang. 2014. RAPID: Rating Pictorial Aesthetics Using Deep Learning. In *ACM Multimedia*. 457–466.

[8] Xin Lu, Zhe Lin, Xiaohui Shen, Radomír Mech, and James Zijun Wang. 2015. Deep Multi-patch Aggregation Network for Image Style, Aesthetics, and Quality Estimation. In *ICCV*. 990–998.

[9] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A Large-Scale Database for Aesthetic Visual Analysis. In *CVPR*. 2408–2415.

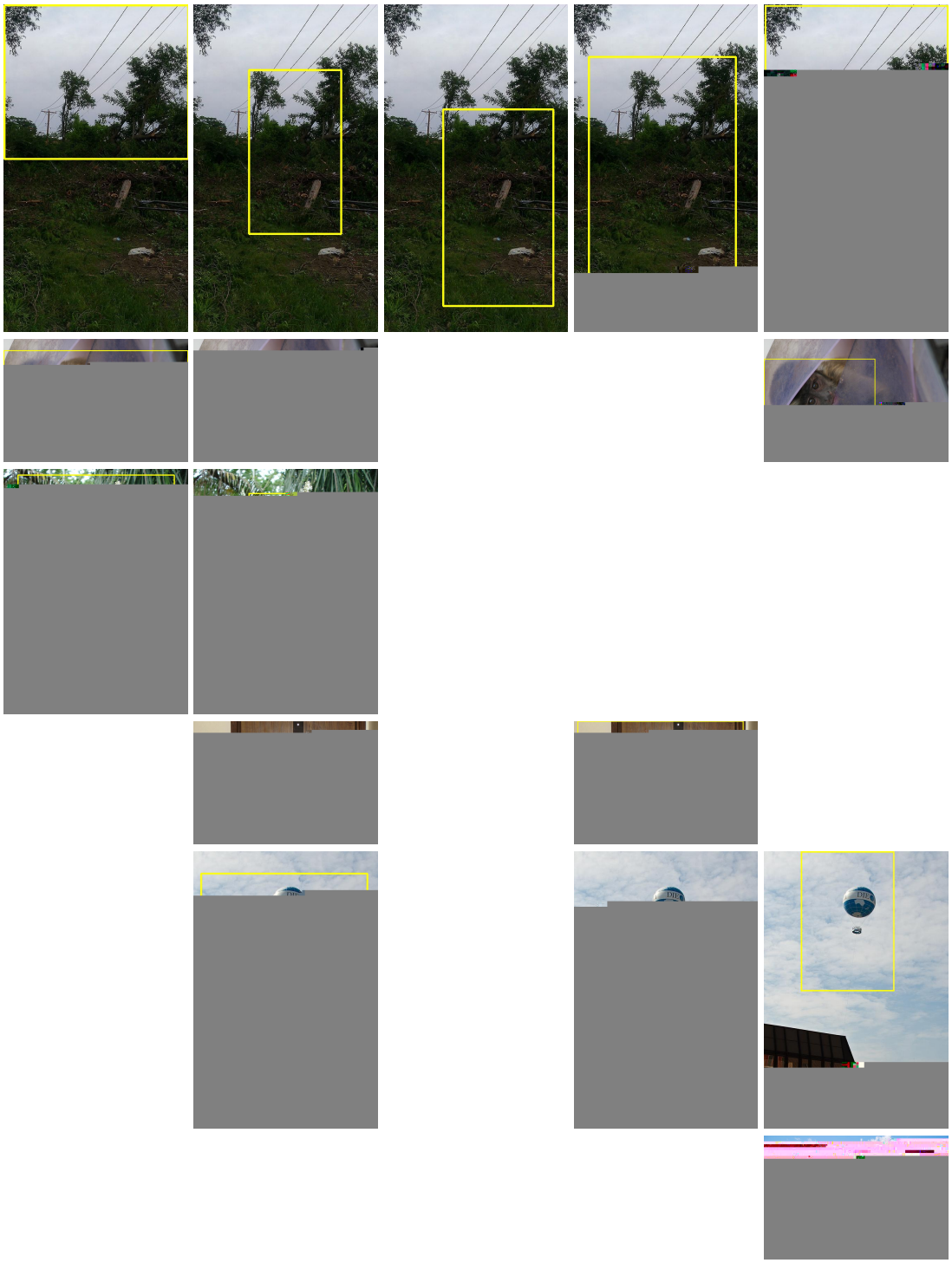
[10] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252.



(a) Ground Truth (b) eDN (c) AlexNet_finetue (d) RankSVM+FCDB (e) VFN

Figure 2: Image cropping examples. The best crop windows determined by various baselines are drawn as yellow rectangles. The test images are from FCDB [2]. Best viewed in color.

[11] Eleonora Vig, Michael Dorr, and David Cox. 2014. Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images. In *CVPR*. 2798–2805.



(a) Ground Truth (b) eDN (c) AlexNet_finetue (d) RankSVM+FCDB (e) VFN

Figure 3: Image cropping examples. The best crop windows determined by various baselines are drawn as yellow rectangles. The test images are from FCDB [2]. Best viewed in color.

(a) Ground Truth (b) eDN (c) AlexNet_ netue (d) RankSVM+FCDB (e) VFN

Figure 4: Image cropping examples. The best crop windows determined by various baselines are drawn as yellow rectangles. The test images are from FCDB [2]. Best viewed in color.

(a) Ground Truth (b) eDN (c) AlexNet_ netue (d) RankSVM+FCDB (e) VFN

Figure 6: Image cropping examples. The best crop windows determined by various baselines are drawn as yellow rectangles. The test images are from FCDB [2]. Best viewed in color.