Automated Crops Using Crowdsourced Data

Sally Ahn*
UC Berkeley

Soham Uday Mehta†
UC Berkeley

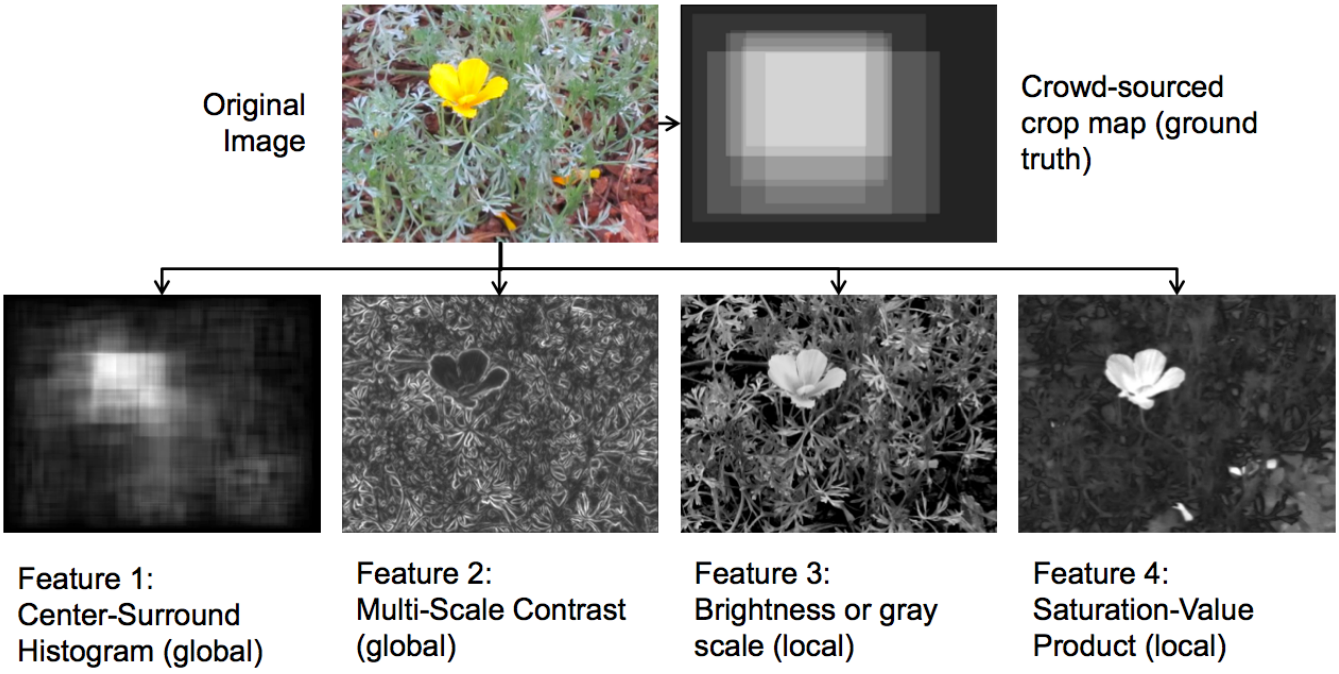

Figure 1: *Our data-driven approach for automating crops.*

Abstract

In this work, we take a data-driven approach to automate image cropping. We first gather data of real people's cropping tendencies by posting image cropping tasks on Mechanical Turk. Next, we explore this data and identify four features that captures both global and local aspects of people's cropping patterns: center-surround histogram, multi-scale contrast, brightness, and saturation-value product. We then train a model to learn the weights of these features on our data. Finally, we this model to automatically generate crops for other input images.

CR Categories: I.4.m [Image Processing and Computer Vision]: Miscellaneous;

Keywords: cropping, crowdsourcing, aesthetics, photographs

Links: DL PDF

1 Introduction

Image cropping is a frequent and relatively simple form of image manipulation that can become a tedious task for large sets of photos. Users crop images for a variety of reasons: to resize an image, emphasize a subject, remove distractions/unwanted elements from the image, improve the overall composition of the image, etc. Moreover, research based on user studies as well as the human visual system has shown that a strong correlation exists between the composition of an image and the aesthetic value perceived for that image [Savakis et al. 2000; Peters 2007; Obrador et al. 2010; Cerosaletti et al. 2011]. Many retargeting methods ([Suh et al. 2003; Santella et al. 2006; Luo 2007]) address these issues. However, while it is true that cropping can be an intermediate step in a retargeting algorithm, there are several distinctions that sets cropping apart from a retargeting subproblem. The main distinction is that the main con-

*e-mail: sallyahn@eecs.berkeley.edu
†e-mail: sohamumehta@berkeley.edu

straints of the retargeting problem is defined in terms of the output image dimensions, i.e. we want to resize the image while preserving salient regions of the image. By this formulation, the problem cannot be solved by cropping alone and various methods of image manipulation [Rubinstein et al. 2008; Wang et al. 2008; Rubinstein et al. 2009] are introduced to overcome the limitations of simple cropping or rescaling.

Often times, however, the user is less concerned about the exact dimensions of the final image, and more interested in emphasizing a subject or removing extraneous/undesired parts of an image *without altering the remaining content of the image*. For example, photographs for ads may need to be cropped to emphasize the subject while preserving the integrity of the resulting photograph. Thus, we are interested in analyzing the underlying psychology of people's cropping behavior and to automate the task of cropping.

2 Related Work

There has been much previous work in automated cropping algorithms. Suh et al. [2003] uses low-level saliency detection to automatically create thumbnail croppings. Such methods evaluate regions of photographs based on low-level features such as brightness, color, etc. As we mentioned earlier, such features often miss regions in the photograph that are of semantic importance. In another approach, Santella et al.[2006] finds "regions of interest" in the photograph by following the users' gaze for that photograph. This overcomes the loss of semantic information, but it places burden on the user by requiring inputs of their gazing pattern. Luo et al. [2007] presents a method for detecting the main subject, which it uses to create a belief map about the photograph content, and then finds the optimal window for that subject. However, this method requires a distinct subject, and many photographs, such as landscapes, lack a single subject.

In the field of psychology, there has been recent research on spatial aesthetics, which reveals that general composition "rules" like the golden rule of thirds–which many automated algorithms incorporate into their computation–lack scientific data supporting its claims on visual appeal [Palmer and Gardner 2008]. Their experiments with human subjects show that deeper principles such as "inward" and "center" bias provides a different and more accurate model of people's preferences in composition [Palmer and Gardner 2008]. Such findings inspired us to investigate whether we can gain better measures for evaluating the composition of a photograph by analyzing real humans' cropping patterns.

3 Methods

3.1 Collecting Data for Cropping Preferences

We gathered all of our data through Mechanical Turk, where we posted many independent *cropping tasks*. These tasks presented the worker with an embedded cropping interface and asked them to "crop the image...so that it looks the most visually appealing." Samples of photographs that croppers were asked to modify are shown in Figure 2.

We also posted corresponding *voting tasks* for each crop because Little [2010] and Bernstein [2010] showed that incorporating such peer evaluation yield more reliable data from Mechanical Turk. When a worker submitted his crop, our application automatically posted five voting tasks for that particular crop. Each of these five tasks displayed the original photograph and the cropped photograph (the order that these two images were presented was randomized) and asked the worker to select the "most visually pleasing" image. We removed the crops that were rejected by a majority of the voters; this allowed us to filter out "bad" crops from lazy or malicious workers. We ran two batches of this experiment on Mechanical Turk. At the end of our second experiment, we had cropping data for 65 different images, with about 25 crops per image. To aggregate the crops for an image, we superimposed the cropped regions and visualized them as a normalized heat map as shown in Figure 2.

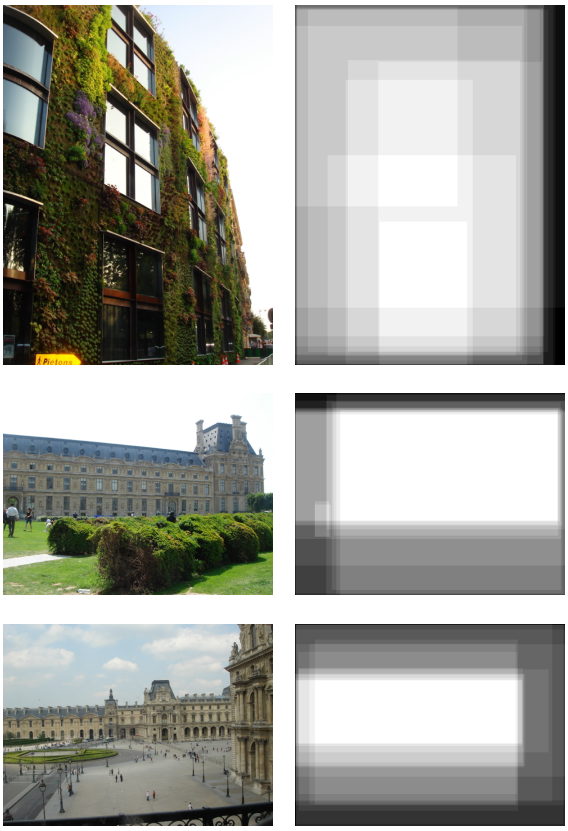

Figure 2: *Sample original photos (left) and their aggregated crops (right).*

3.2 Automatic Cropping

Using the crops obtained from Mechanical Turk, we use a data-driven model for automating cropping for an arbitrary image. We first needed to reduce the image to a set of features that covers both global and local information of the image in order to capture the "cropping function" well. We explored several features represented as a grayscale image normalized to [0,1] and compared to the ground-truth crop maps. Features that showed little or no correlation to the crop maps were discarded. We trained our final model with four features, which we describe below.

3.2.1 Center Surround Histogram

Liu et al. [Liu et al. 2007] introduces the *center surround histogram* feature that describes the saliency of an image region. It captures the values of the surrounding region of a pixel and can be viewed as a global feature in which each pixel depends on a large number of neighboring pixels. To derive a center surround histogram map, we first find the chi-square distance between the RGB histograms of a rectangle centered at the pixel, and a surrounding rectangular

contour of an equal area, for every pixel. We calculate these distances for different aspect ratios and scales of rectangles and select those that maximize the chi-square distance. The distance values of the selected rectangles are then summed and normalized to produce the final map. Although we saw some correspondence between this feature and our crop data, there were conflicting cases as well. Examples of these maps are shown in 3.

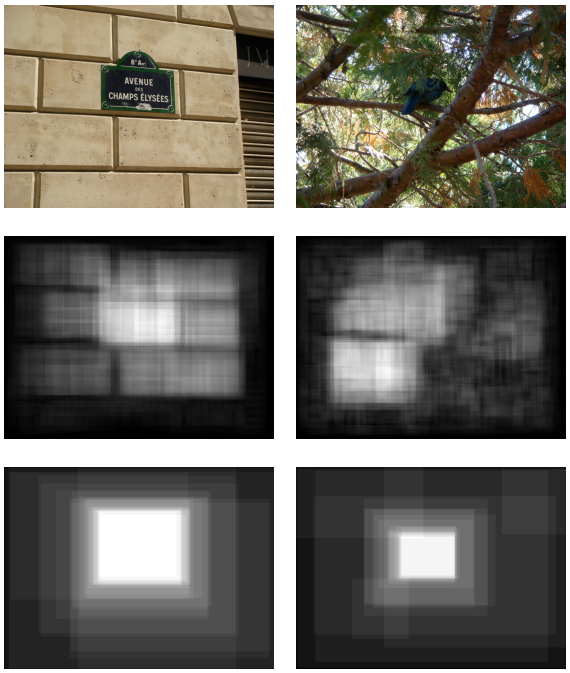

Figure 3: *From top to bottom, Original Image, its center surround histogram and its crop map. Left: Strong correlation with crops, Right: Weak correlation with crops*

3.2.2 Multi-Scale Contrast

By summing the contrast of the image at multiple scales, we capture more global information about the image. Our multi-scale contrast map is the sum over an image-pyramid of 5 images. Examples are shown in Figure 4.

3.2.3 Brightness and Saturation-Value Product

Our brightness map is simply a normalized grayscale image of the input image. For saturation, we chose to multiply the saturation channel with the value channel to remove noise from photos taken at low-light conditions. Examples of the latter are shown in Figure 5.

3.3 Learning Feature Weights

The task of the machine learning framework is to determine the relative importance of the four features discussed above in cropping.

Since each of the training images has a different set of feature maps, we use a Conditional Random Field (CRF) to model the dependence of the optimal crop on the features. We refer to [Murphy 2001] and [Klinger and Tomanek 2007] for the theory of CRFs. The underlying graph for the CRF is an Ising model-like grid of pixels, with edges only between adjacent pixels in the image.

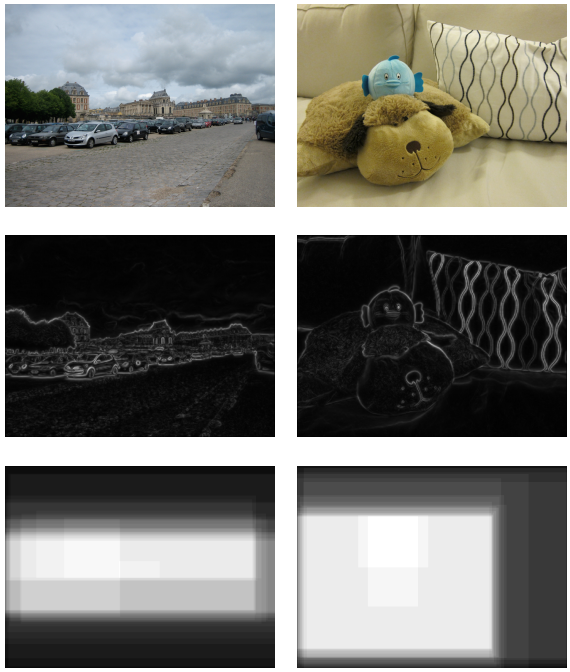

Figure 4: *From top to bottom, Original Image, its multi-scale contrast and its crop map. Left: Strong correlation with crops, Right: Weak correlation with crops*

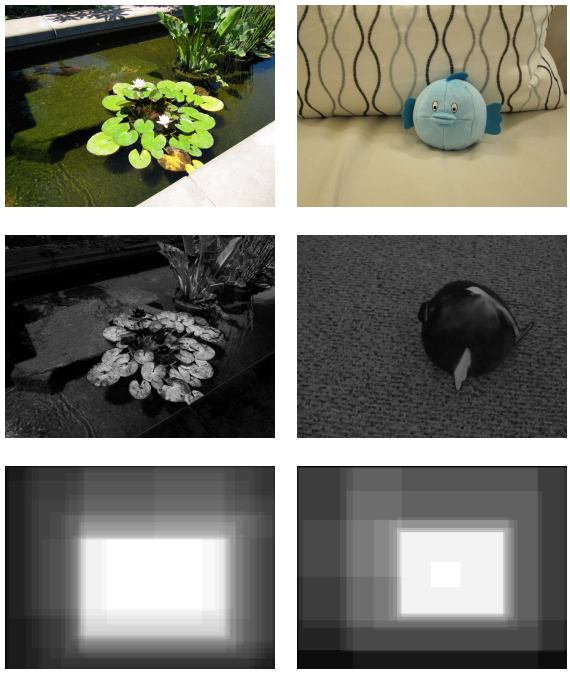

Figure 5: *From top to bottom, Original Image, its saturation and value product and its crop map. Left: Strong correlation with crops, Right: Weak correlation with crops*

The node states $c_x \in \{0, 1\}$ correspond to the pixels outside and inside the crop region respectively. The basic assumption is that the node potentials must be proportional to the feature values at that

node. Let $\{I^n\}_{n=1}^N$ be the set of training images and $\{f_k^n\}_{k=1}^K$ be the set of K normalized features where the superscript denotes the image index.

Then, we define node potentials as

$$\Psi_i^n(c_i = 0) = \sum_k w_k(1 - f_k^n(i))$$

and

$$\Psi_i^n(c_i = 1) = \sum_k w_k(f_k^n(i))$$

$\{w_k\}_{k=1}^K$ are the weights for the features. The primary goal is to estimate these weights.

We can also model the edge potentials as functions of the features at the two pixels, but it is not clear exactly how this can be modeled, especially since the crop boundaries (where the pixels change states) do not correspond well to the existing features. Hence, we simply modeled homogenous edge potentials of the form:

$$\Psi_{ij}^n(c_i = 0, c_j = 0) = p_1$$
$$\Psi_{ij}^n(c_i = 1, c_j = 0) = p_2$$
$$\Psi_{ij}^n(c_i = 0, c_j = 1) = p_2$$
$$\Psi_{ij}^n(c_i = 1, c_j = 1) = p_1$$

We did not implement estimation of $\{p_1, p_2\}$ due to time limitations. Instead, we chose these to be empirical estimates from the ground truth, with p_1^n equal to the fraction of edges with same values of the end pixels (since the ground truth image is a normalized gray scale map, we threshold it first to a binary image for this purpose) and p_2^n to be the fraction of edges with differing values of end pixels.

The expression for the conditional probability of the crop map C is

$$P(C^n|I^n; w, p) = \frac{1}{Z} exp\left(\sum_{i \in V} \Psi^n(c_i) + \sum_{(i,j) \in E} \Psi^n(c_i, c_j)\right)$$

The log likelihood for the entire data set $D = \{I^n, G^n\}_{n=1}^N$ is:

$$L(w, D) = log\left\{\prod_n P(G^n|I^n; w, p)\right\}$$

Since the ground truth G^n is a normalized probability map, we can write

$$p(c_x^n = 1|g_x^n) = g_x^n$$

and

$$p(c_x^n = 0|g_x^n) = 1 - g_x^n$$

We implemented a simple gradient descent scheme to estimate w.

$$w^* = argmax_w L(w, D)$$

For each step of the gradient descent we need to compute the node marginals $P(c_i|I; w)$ under the current model for each image. This was done using a loopy belief propagation algorithm from open-source Matlab Toolbox UGM [Schmidt 2007].

3.4 Predicting the best crop

Given a new image I we would like to predict the best crop of this image using the previously learned weights w. The problem is now

$$C^* = argmax_c P(C|I; w)$$

This is a very high dimensional (equal to number of pixels in the image) optimization problem. We resized the image, as in the learning phase, to 80x60. For maximum likelihood decoding we used Loopy decoding from the UGM toolbox.

The obtained most likely crop map (see Figures 6b, 6f, and 6j) is generally a connected white region of irregular shape. We can find a simple bounding box for this region, but it may be arbitrarily large, which is not useful. Hence, we try to find a small enough bounding box that covers about 90% of the pixels with $c_x^* = 1$. We first find the centroid of the raw crop map, and then in each iteration grow the current bounding box in one of four directions (by adding a row to the top or bottom or a column to the left or right) depending on which addition gives the maximum number of white pixels.

4 Results

Our resulting crops are shown in Figure 6. To evaluate, we define a "goodness" metric as follows the mean of ground truth values in the cropped region. This is reasonable because if a crop is larger than the most bright region in the ground truth map, then the resulting "goodness" score is small. With this definition, the scores for our images shown in 6 are shown in Table 1.

Input Image	Crop Goodness Score
Eiffel Tower (Figure 6a)	0.77
Flower (Figure 6e)	0.27
Fish (Figure 6i)	0.71

Table 1: *"Goodness" scores of our generated crops.*

We resized the images to 80x60 for a reasonable computation time. This size has 4800 nodes. The convergence was extremely slow, and the derivative reduced by one order of magnitude after 10 iterations. The learned (normalized) weights for one training set of 60 images are tabulated in Table 2.

Features	Weights
Center-Surround Histogram	0.36
Multi-Scale Contrast	0.22
Saturation-Value Product	0.24
Brightness	0.18

Table 2: *The learned (normalized) weights for one training set of 60 images*

5 Discussion

In summary, we collected cropping data for a number of images using a public domain crowd-sourcing platform. We devised a CRF model for image cropping based on four image features, both local and global. We trained our model with the obtained data, and tested it with new images. We obtained reasonable results.

We found that identifying features that apply to an arbitrary photograph to be extremely difficult. As we explain with the figures in the previous sections, for every feature we explored, there seemed

to be some correlation with people's cropping patterns, but some cases directly contracted the general trends. This suggests that we may need to classify photographs into categories that reduce some of the variance of such features. Moreover, Our model assumes a linear relationship among the features, which may be an oversimplification of the relationships among the features in the real world.

6 Future Work

For future work, we will be working on finding more robust image features for the model. Specifically, we will need to devise features that capture not only the salient object but also the region around it. Further, we expect the weights of the features to be slightly different for different categories of images. Therefore, we need to train the model separately for these categories. Some categories to explore are landscapes and subject vs. no subject. Also, we will be working on improving the model itself, by trying different forms of dependence of the node and edge potentials on the features. We intend to follow an approach similar to [Pietra et al. 1997].

References

BERNSTEIN, M. S., LITTLE, G., MILLER, R. C., HARTMANN, B., ACKERMAN, M. S., KARGER, D. R., CROWELL, D., AND PANOVICH, K. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, ACM, New York, NY, USA, UIST '10, 313–322.

CEROSALETTI, C. D., LOUI, A. C., AND GALLAGHER, A. C. 2011. Investigating two features of aesthetic perception in consumer photographic images: clutter and center. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, vol. 7865 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*.

KLINGER, R., AND TOMANEK, K. 2007. Classical Probabilistic Models and Conditional Random Fields. Tech. Rep. TR07-2-013, Department of Computer Science, Dortmund University of Technology, December. ISSN 1864-4503.

LITTLE, G., CHILTON, L. B., GOLDMAN, M., AND MILLER, R. C. 2010. Exploring iterative and parallel human computation processes. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, ACM, New York, NY, USA, HCOMP '10, 68–76.

LIU, T., SUN, J., NING ZHENG, N., TANG, X., AND YEUNG SHUM, H. 2007. Learning to detect a salient object. In *in: Proceedings of IEEE Computer Society Conference on Computer and Vision Pattern Recognition (CVPR*, CVPR, 1–8.

LUO, J. 2007. Subject content-based intelligent cropping of digital photos. *IEEE ICME*.

MURPHY, K. P. 2001. An introduction to graphical models. Tech. rep.

OBRADOR, P., SCHMIDT-HACKENBERG, L., AND OLIVER, N. 2010. The role of image composition in image aesthetics. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, 3185 –3188.

PALMER, S., AND GARDNER, J. 2008. Aesthetic issues in spatial composition: Effects of position and direction on framing single objects. *Spatial Vision*.

PETERS, G. 2007. Aesthetic primitives of images for visualization. In *Information Visualization, 2007. IV '07. 11th International Conference*, 316 –325.

PIETRA, S. D., PIETRA, V. D., AND LAFFERTY, J. 1997. Inducing features of random fields. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE 19*, 4, 380–393.

RUBINSTEIN, M., SHAMIR, A., AND AVIDAN, S. 2008. Improved seam carving for video retargeting. *ACM Transactions on Graphics 27*, 3 (Aug.), 1.

RUBINSTEIN, M., SHAMIR, A., AND AVIDAN, S. 2009. Multi-operator media retargeting. *ACM Transactions on Graphics 28*, 3 (July), 1.

SANTELLA, A., AGRAWALA, M., DECARLO, D., SALESIN, D., AND COHEN, M. 2006. Gaze-based interaction for semi-automatic photo cropping. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, ACM, New York, NY, USA, CHI '06, 771–780.

SAVAKIS, A. E., ETZ, S. P., AND LOUI, A. C. 2000. Evaluation of image appeal in consumer photography. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, B. E. Rogowitz & T. N. Pappas, Ed., vol. 3959 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, 111–120.

SCHMIDT, K., 2007. Ugm: Matlab code for undirected graphical models, June.

SUH, B., LING, H., BEDERSON, B. B., AND JACOBS, D. W. 2003. Automatic thumbnail cropping and its effectiveness. *Proceedings of the 16th annual ACM symposium on User interface software and technology - UIST '03 5*, 2, 95–104.

WANG, Y.-S., TAI, C.-L., SORKINE, O., AND LEE, T.-Y. 2008. Optimized scale-and-stretch for image resizing. *ACM Transactions on Graphics 27*, 5 (Dec.), 1.

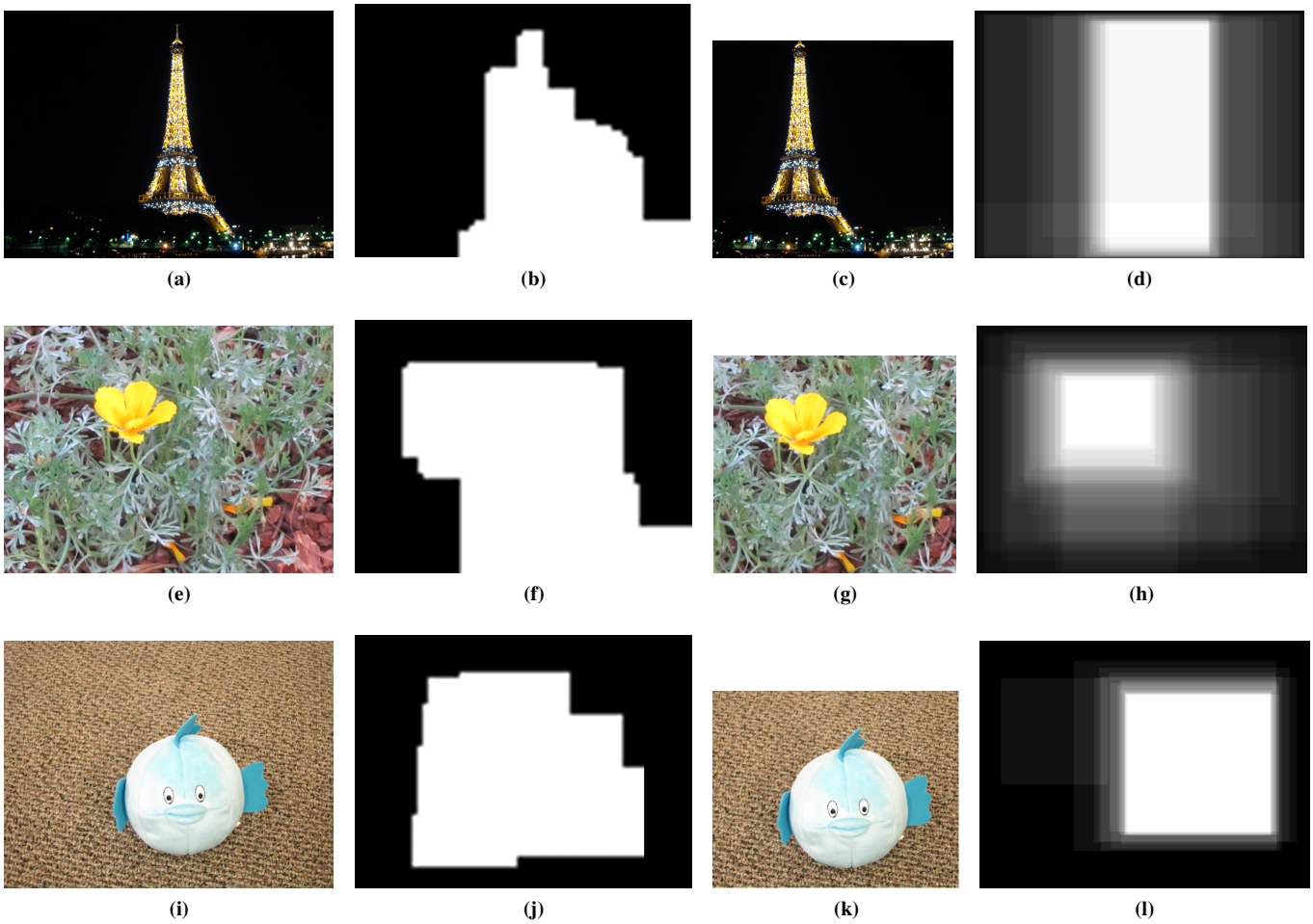

Figure 6: *Results. From left to right: original photograph (a,e,i), our crop mask (b,f,j), final crop (c,g,k), ground truth aggregated crops (d,h,l)*