Adaptive Energy Selection for Content-Aware Image Resizing

Kazuma Sasaki* Yuya Nagahama* Zheng Ze Satoshi Iizuka
Edgar Simo-Serra Yoshihiko Mochizuki Hiroshi Ishikawa
Department of Computer Science and Engineering
Waseda University, Tokyo, Japan

Abstract

Content-aware image resizing aims to reduce the size of an image without touching important objects and regions. In seam carving, this is done by assessing the importance of each pixel by an energy function and repeatedly removing a string of pixels avoiding pixels with high energy. However, there is no single energy function that is best for all images: the optimal energy function is itself a function of the image. In this paper, we present a method for predicting the quality of the results of resizing an image with different energy functions, so as to select the energy best suited for that particular image. We formulate the selection as a classification problem; i.e., we 'classify' the input into the class of images for which one of the energies works best. The standard approach would be to use a CNN for the classification. However, the existence of a fully connected layer forces us to resize the input to a fixed size, which obliterates useful information, especially lower-level features that more closely relate to the energies used for seam carving. Instead, we extract a feature from internal convolutional layers, which results in a fixed-length vector regardless of the input size, making it amenable to classification with a Support Vector Machine. This formulation of the algorithm selection as a classification problem can be used whenever there are multiple approaches for a specific image processing task. We validate our approach with a user study, where our method outperforms recent seam carving approaches.

1. Introduction

Due to the advent of devices with diverse sizes of screens, resizing images and videos to matching aspect ratio has become increasingly necessary. Beyond simple scaling and cropping, various techniques for reshaping images without changing their feel and content have been proposed. Such techniques, which are called the content-aware image resizing, try to resize the image by cutting out unimportant regions, without reducing the size of important objects and regions. For instance, a picture of a person with a large background scene might be reduced in size by only cutting the background, without changing the size of the foreground.

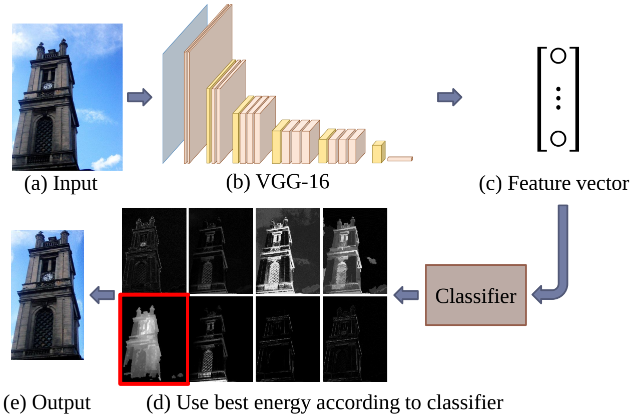
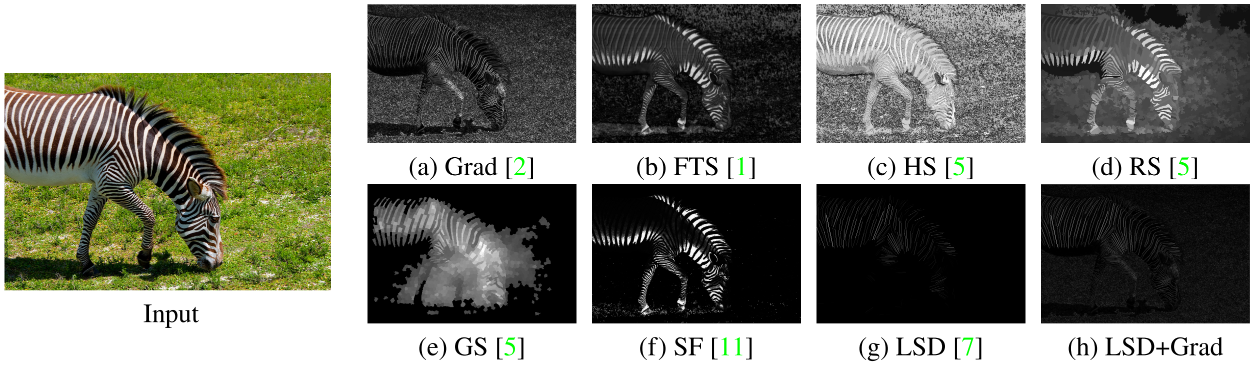


Figure 1. A flowchart of our method. (a) Input image. (b) Feature extraction using a trained CNN. (c) Extracted low-dimensional feature vector. (d) Classification of the feature vector to determine the best energy. (e) The chosen energy is used for seam carving.

Among the proposed techniques, seam carving [2] is a fast and effective method that has been a focus of research and improvement. It first computes an *energy*, which estimates the importance of each pixel, and use it to avoid removing those pixels with high energy. In the original paper [2], the intensity gradient is used as the energy. However, this can lead to warping important regions with complex background, such as a forest. In such a case, using the visual saliency as the energy can obtain good results [4]. Along these lines, various energy functions have been proposed for seam carving. However, each has its strength; an energy works for certain kind of images, another for others. Though the results can be very different, it is not simple to choose the right energy to use for each image.

In this paper, we propose adaptively selecting the energy to use for seam carving according to the input image. While

*equal contribution



Input

(a) Grad [2]

(b) FTS [1]

(c) HS [5]

(d) RS [5]

(e) GS [5]

(f) SF [11]

(g) LSD [7]

(h) LSD+Grad

Figure 2. The eight different energy maps we use, computed on an example input image.

we focus on the seam carving task, our approach is amenable to any application in which a set of approaches exists for an image processing task. We formulate the problem as a classification problem using a pre-trained Convolutional Neural Network (CNN) to extract image features. In contrast to the standard approach, where the use of a fully connected layer fixes the size of the input image, we use the mean and the covariance of the internal convolutional layers. This allows using input images of arbitrary size, as well as using lower level features, that more closely relate to the energies used for seam carving. We then train a classifier to estimate the quality of the result of the different seam carving approaches using these features, which allows us to improve the overall quality. An overview of the approach can be seen in Fig. 1.

We create a dataset of seam carving results and their quality score, and evaluate considering eight different energy functions for the seam carving problem. In total, we use 600 images for training our model. By a user study, we also compare the quality of the results of using the original single energy function with that of using the energy function chosen by our approach.

2. Related Work

Content-aware image resizing aims to reduce the size of an image without touching important objects and regions. Early examples of such methods include [15], which detects human faces in pictures and crops them to create thumbnails, as well as [4], where images are resized by automatically detecting important regions in the image according to a visual attention model that includes the notion of regions of interest (ROI) and attention value (AV). There is also a method [13] that crops an image preserving important regions that are detected by tracking the movement of the eyes of a human observer, obtaining good resizing results.

Seam carving [2] is one of the most representative of the methods that define an energy function assessing the importance of each pixel and then resize images according to the energy. Seam carving uses the intensity gradient as the energy map in cutting out a “seam”, which is defined as a string of pixels that connects the left and the right, or the top and the bottom, of the image such that there is exactly one pixel in each column or row. This has been improved in various ways, including an extension to videos that cuts out a 2D seam using graph cuts [12], and a combination with scaling [6]. Matthias et al. [8] took advantage of a chronologically and spatially discontinuous seams to resize videos robustly. In [10], for esthetically pleasing resizing of images with multiple objects, depth maps are used to take the depth of the scene and the distribution of objects into consideration. Cao et al. [3] divided the image into a set of strips and from within each strip chose a single seam leading to improved results.

Thus, many different variations of energies for seam carving have been proposed. Yet there is no method to select the energy best suited for each particular image. In this paper, we propose a method to do that by learning the correspondence between images and resizing results using various energies, and apply it for overall better automatic resizing results.

3. Proposed Method

We propose a method to select the energy function most suited to resize a given input image using seam carving in terms of the resizing quality. We train a classifier to estimate the quality of the result of the different seam carving energy functions. For a given input image, we use the classifier to select the energy function that gives the highest predicted score and then use it for seam carving.

3.1. Seam Carving

A horizontal seam is an eight-neighbor-connected string of pixels that connects the left and the right side of the image such that there is exactly one pixel in each column. Similarly, a vertical seam connects the top and the bottom and there is exactly one pixel in each row. For example, in an $n \times m$

image, a vertical seam $\boldsymbol{s}$ is given by

$$\boldsymbol{s} = \{s_i\}_{i=1}^{n} = \{(x(i), i)\}_{i=1}^{n}, \quad (1)$$

for some function x that gives an x-coordinate for each y-coordinate $i = 1, \ldots, n$ such that $|x(i) - x(i-1)| \leq 1$ for all $i = 2, \ldots, n$.

In seam carving [2], the seam that connects together the least important pixels, according to an energy map that gives the importance of each pixel, is found by dynamic programming. By removing this seam, the image is reduced by one pixel in an direction (vertical or horizontal) without losing important regions. This is repeated as necessary in both directions.

Let $e(\boldsymbol{I}, p)$ be the energy function that gives the importance of the pixel p in image $\boldsymbol{I}$ and define the cost $E(\boldsymbol{I}, \boldsymbol{s})$ of a seam $\boldsymbol{s}$ on image $\boldsymbol{I}$ by

$$E(\boldsymbol{I}, \boldsymbol{s}) = \sum_{i=1}^{n} e(\boldsymbol{I}, s_i). \quad (2)$$

Then the optimal seam s^* is given by minimizing $E(\boldsymbol{I}, \boldsymbol{s})$:

$$s^* = \operatorname*{argmin}_{\boldsymbol{s}} E(\boldsymbol{I}, \boldsymbol{s}) = \operatorname*{argmin}_{\boldsymbol{s}} \sum_{i=1}^{n} e(\boldsymbol{I}, s_i). \quad (3)$$

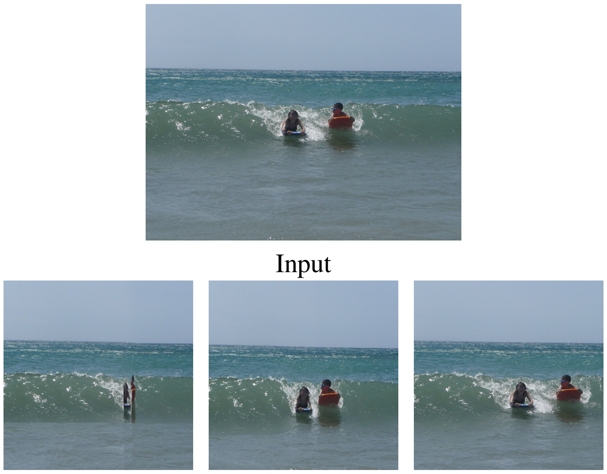
In this paper, we use multiple energy functions $e(\boldsymbol{I}, p)$ and choose one for a given input image so that the result of the resizing using it is the best in quality.

3.2. Energy

The energy function gives each pixel a value that represents its importance as a number in the interval $[0, 1]$, 1 being the most important. Since the best energy function for each image can be different, here we use eight to choose from. In addition to the gradient of intensity (Grad) used in the original seam carving paper [2], we use five saliency-based energies: the Frequency-tuned saliency [1](FTS), the Histogram-based saliency [5](HS), the Region-based saliency [5](RS), the Geodesic saliency [17](GS), and the Saliency filter [11](SF). In addition, we also use the Line Segment Detector [7](LSD), which is based on line segment detection, and its (normalized) sum with the gradient (LSD+Grad). Fig. 2 visualizes the eight energy functions on an example image.

3.3. Seam Carving Dataset

We use the images from the MS COCO dataset [9] in our dataset. We randomly chose 1000 images and resized them by seam carving, using the eight energy functions above and subjectively scored the naturalness of each result. Each image was resized to 60% of the original width and 80% of the original height. The scoring was done based on the following criteria:


Input

(a) score 0 (b) score 1 (c) score 2

Figure 3. Examples of scoring the result of resizing. (a) Most of the persons are cut out. (b) Some of the persons are cut out, also their distance reduced. (c) The persons are intact. (a), (b), and (c) are the results of using LSD, HS, and GS as energies, respectively.

Score 0: The object of interest is clearly distorted.

Score 1: The object of interest is incomplete to a small degree, or there is an obviously unnatural part in the image.

Score 2: The object of interest is almost intact and there is no unnatural part in the image.

An example of resizing results and their scores is shown in Fig. 3. In (a), most of the persons are cut out, while in (b) some of the persons are cut out, still seeming unnatural. In contrast, in (c) the persons are intact and the whole image seems natural. The examples (a), (b), and (c) are the results of using LSD, HS, and GS as energies, respectively.

3.4. Feature Vector

As seam carving is used for image resizing, it is important to be able to compute feature vectors from images of arbitrary size. The standard approach of using the output of the fully-connected layers of pre-trained CNNs as features is thus not applicable, as it limits the input images to a fixed size. Uniformly resizing input images to a fixed size obliterates useful information, especially lower-level features that more closely relate to the energies used for seam carving. We instead propose using the outputs of the convolutional layers, and converting them into fixed-length feature vectors, even though they are computed on images of arbitrary sizes.

Our method here is inspired by the Region Covariance [16], extending it to the activation map with higher-dimensional "pixels". It has the important feature that the output vector is always of the same dimension irrespective of the size of

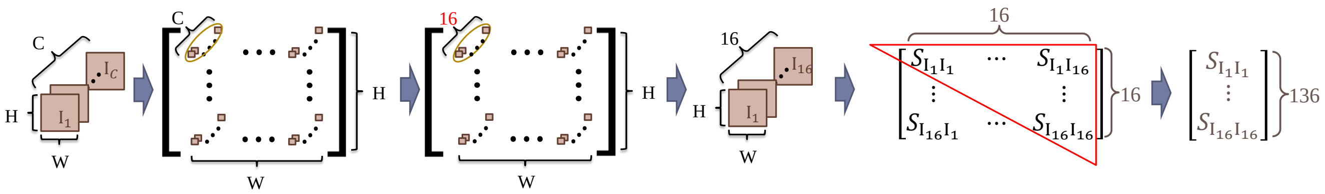


Figure 4. Overview of the process to extract feature vectors from the convolutional layers of a CNN. We illustrate the example assuming that the activation map is of the size $W \times H$-"pixels" with C channels. The size of the map depends on the input size. First, the C channels are reduced to 16 dimensions for each "pixel", using PCA. Then, the "pixel"-wise mean and covariance is computed, resulting in a 16-dim mean vector and a 16×16 covariance matrix. Since the covariance matrix is redundant, we only take its upper triangle, which gives 136 numbers. Adding the mean vector, the result is a 152-dim feature vector. Note the size of the feature vector is independent of the input image size.

the input image. Our approach consists of computing the pixel-wise mean and covariance of the activation map of a convolutional layer, and using them as a fixed-length feature vector. However, as the number of features can be very large, thus leading to enormous feature vectors, we initially reduce the number of channels by applying Principal Component Analysis (PCA) pixel-wise. Finally, as the covariance matrix is symmetric, the values on one side of the diagonal become redundant, which is why we only use the upper triangle. As a feature vector, we use the concatenation of the mean vector and the upper triangle of the covariance matrix, which is flattened into a vector. An overview of the feature extraction approach is illustrated in Fig. 4.

3.5. Classifier

We train a support vector machine by stochastic gradient descent with the feature vector and the score of the images in our training dataset. For an input image, the classifier predicts the score (0-2) each of the eight energy functions would obtain.

4. Results

We train using 600 of the images in the dataset and validate using the remaining 400 images. As a feature extractor, we consider the VGG-16 pre-trained CNN [14], in which we use the *conv4* layer as the source of the feature vector. We reduce the activation output of the layer to a 152-dimensional vector using our feature extraction approach. Among the classifiers we tried, the best predicted the score with an accuracy of 60.25% on the validation set.

4.1. Feature Evaluation

We evaluate different layers of the VGG-16 network as possible features. In particular, we consider the *conv2*, *conv3*, *conv4*, and *conv5* convolutional layers and the fully-connected *fc6* layer. For convolutional layers we reduce the channels to 16 using PCA, which allows us to obtain a 152-dimensional feature vector when concatenating the mean vector and the covariance matrix of the compressed activation maps (see Fig. 4). The fully-connected layer, which extracts 4096-dimensional features, is one of the popular approaches for extracting features with pre-trained CNNs. The results of predicting the score of the different energies is shown in Table 1, in which we also compare the accuracy of resizing the images to 224×224 pixels and using the full resolution of the images. We can see that our proposed feature extraction approach, which allows using full resolution images, gives a significant improvement in classification performance. For all further results, we use the features extracted from the *conv4* layer for our energy function classifier.

Table 1. Classification results for features extracted for different layers for predicting which seam carving energy to use. We also compare with using 224×224 pixel fixed sizes (Resized), and using the original size (Full Res.). Best result is highlighted in bold.

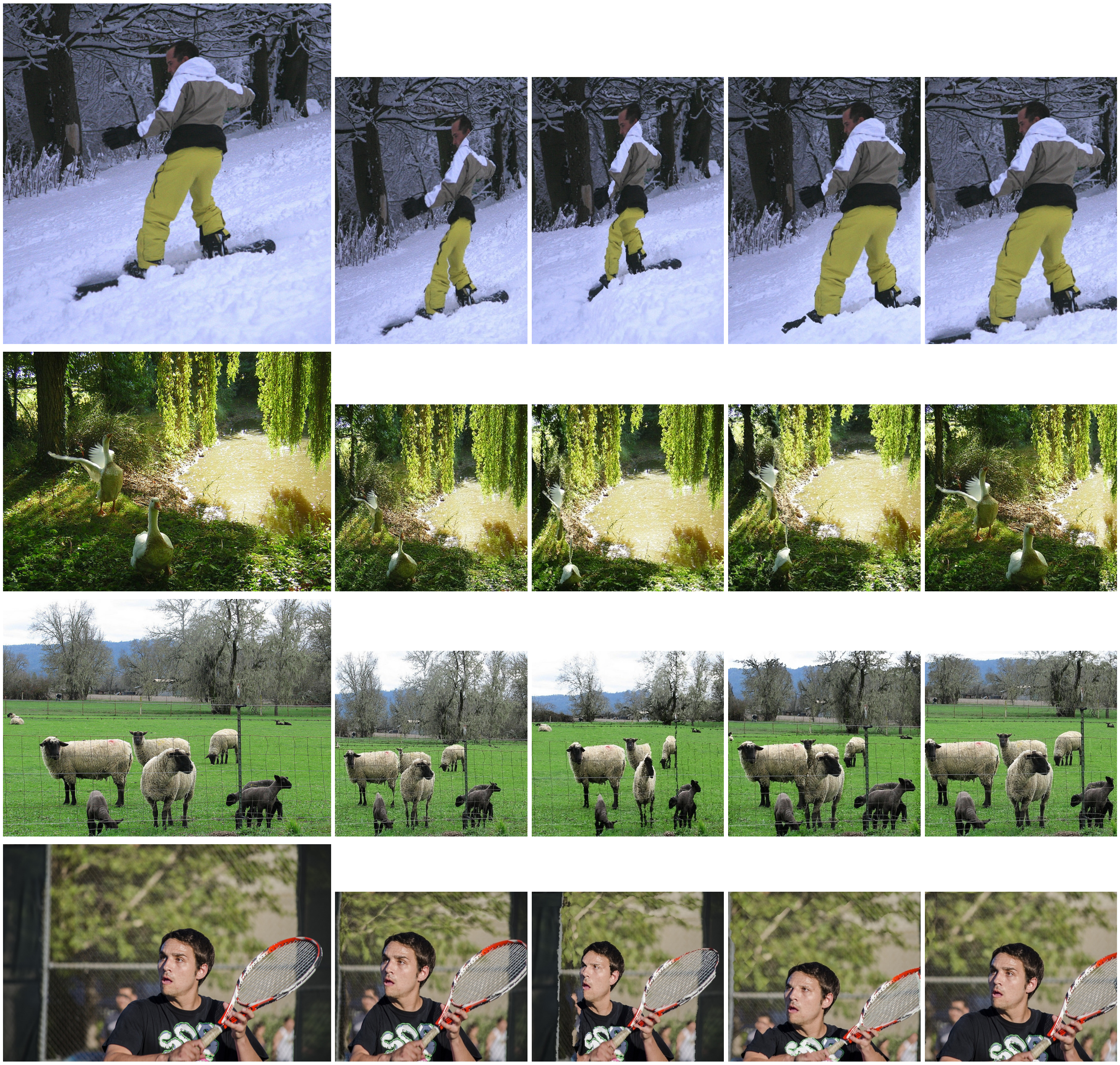
	conv2	*conv3*	*conv4*	*conv5*	*fc6*
Resized	43.19	32.78	52.34	47.00	39.84
Full Res.	56.16	57.88	**60.25**	54.38	N/A

4.2. User Study

We also evaluate our approach in a user study using 400 images from the MS COCO dataset. We compared our approach against Grad [2], HS [5], and SF [11]. A total of 8 people participated in the user study and were asked to choose the better one out of each pair of results shown. Each user was shown 100 random pairs of images consisting of the result of our approach and the result of one of the other approaches, chosen randomly. Results are shown in Table 2, where we see our proposed approach significantly outperforms all the other approaches.

Table 2. Results of the user study in which we compare our approach against three recent seam carving approaches.

	vs Grad [2]	vs HS [5]	vs SF [11]
Ours (% preferred)	62.25	67.00	57.38



(a) Input (b) Grad [2] (c) HS [5] (d) SF [11] (e) Ours

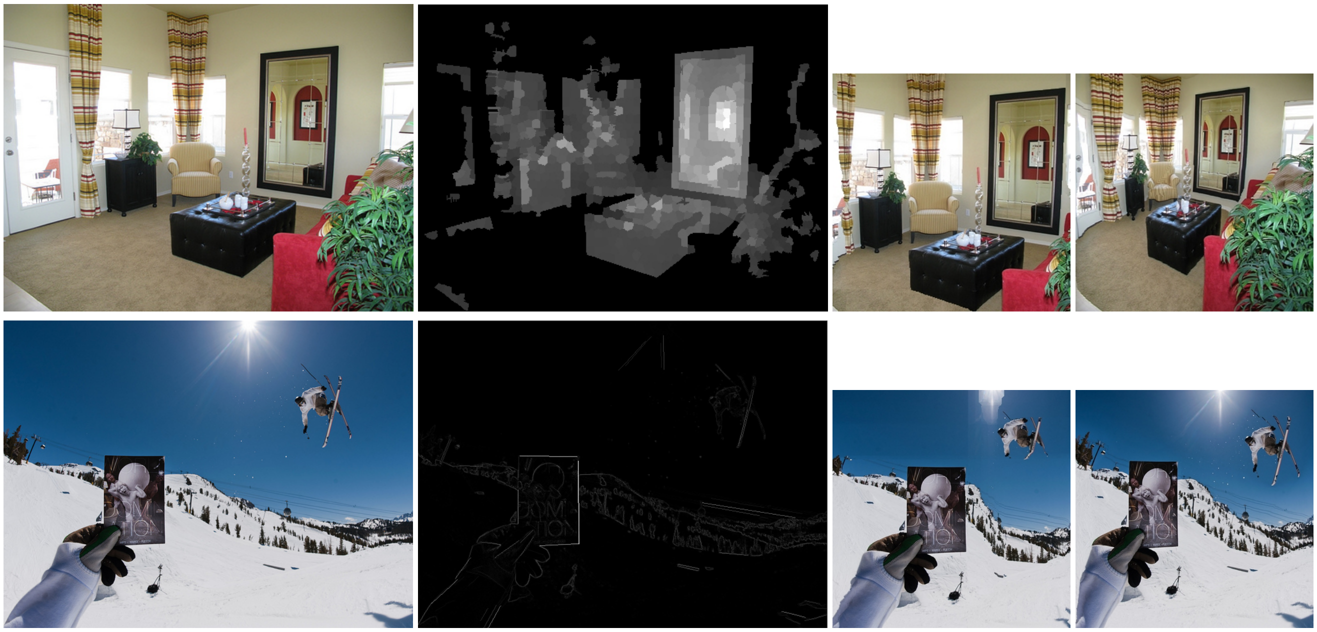
Figure 5. Qualitative comparison of resizing results.

4.3. Qualitative Comparison

We show qualitative results in Fig. 5. Our approach adaptively chooses the energy function for each case. Given that the optimal energy function depends on the image at hand, this allows obtaining more convincing results in a larger variety of cases than always using the same energy. While we have focused on chosing energy for the original seam carving [2], it is also possible to use our approach with other varieties, such as the improved seam carving [12]. Some examples of using both approaches are shown in Fig. 6.

5. Conclusion

In this paper, we have presented a method for adaptively selecting the best energy function for seam carving. We have shown the effectiveness of formulating the selection as classification using CNN, using a size-independent feature so that small low-level feature is preserved. Although we have focused on a specific problem, our approach is applicable to a wide variety of problems, and can be easily extended to new approaches.



(a) Input (b) Chosen energy map (c) [2] (d) [12]

Figure 6. Using the energy function chosen by our approach with both seam carving [2], and improved seam carving [12].

Acknowledgements This work was partially supported by JST ACT-I Grant Number JPMJPR16U3, JST ACT-I Grant Number JPMJPR16UD, and JST CREST Grant No. JPMJCR14D1.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *Computer Vision and Pattern Recognition*, pages 1597–1604, 2009.

[2] S. Avidan and A. Shamir. Seam carving for content-aware image resizing. *ACM Transactions on graphics (TOG)*, 26(3):10, 2007.

[3] L. Cao, L. Wu, and J. Wang. Fast seam carving with strip constraints. In *Proceedings of the 4th International Conference on Internet Multimedia Computing and Service*, pages 148–152, 2012.

[4] L.-Q. Chen, X. Xie, X. Fan, W.-Y. Ma, H.-J. Zhang, and H.-Q. Zhou. A visual attention model for adapting images on small displays. *Multimedia systems*, 9(4):353–364, 2003.

[5] M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S. Hu. Global contrast based salient region detection. *Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015.

[6] W. Dong, N. Zhou, J.-C. Paul, and X. Zhang. Optimized image resizing using seam carving and scaling. *ACM Transactions on Graphics (TOG)*, 28(5):125, 2009.

[7] R. Grompone von Gioi, J. Jakubowicz, J. Morel, and G. Randall. LSD: a line segment detector. *Image Processing On Line*, 2:35–55, 2012.

[8] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Discontinuous seam-carving for video retargeting. In *Computer Vision and Pattern Recognition*, pages 569–576, 2010.

[9] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[10] A. Mansfield, P. Gehler, L. Van Gool, and C. Rother. Scene carving: Scene consistent image retargeting. In *European Conference on Computer Vision*, pages 143–156, 2010.

[11] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *Computer Vision and Pattern Recognition*, pages 733–740, 2012.

[12] M. Rubinstein, A. Shamir, and S. Avidan. Improved seam carving for video retargeting. *ACM transactions on graphics (TOG)*, 27(3):16, 2008.

[13] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen. Gaze-based interaction for semi-automatic photo cropping. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 771–780, 2006.

[14] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[15] B. Suh, H. Ling, B. B. Bederson, and D. W. Jacobs. Automatic thumbnail cropping and its effectiveness. In *Proceedings of the 16th annual ACM symposium on User interface software and technology*, pages 95–104, 2003.

[16] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *European Conference on Computer Vision*, pages 589–600. 2006.

[17] Y. Wei, F. Wen, W. Zhu, and J. Sun. Geodesic saliency using background priors. In *European Conference on Computer Vision*, pages 29–42. 2012.