

Weakly- and Self-Supervised Learning for Content-Aware Deep Image Retargeting

Donghyeon Cho
KAIST

Jinsun Park
KAIST

Tae-Hyun Oh
KAIST

Yu-Wing Tai
Tencent YouTu

In So Kweon
KAIST

{cdh12242, zzangjinsun, thoh.kaist.ac.kr, yuwing}@gmail.com, iskweon@kaist.ac.kr

Abstract

This paper proposes a weakly- and self-supervised deep convolutional neural network (WSSDCNN) for content-aware image retargeting. Our network takes a source image and a target aspect ratio, and then directly outputs a retargeted image. Retargeting is performed through a shift map, which is a pixel-wise mapping from the source to the target grid. Our method implicitly learns an attention map, which leads to a content-aware shift map for image retargeting. As a result, discriminative parts in an image are preserved, while background regions are adjusted seamlessly. In the training phase, pairs of an image and its image-level annotation are used to compute content and structure losses. We demonstrate the effectiveness of our proposed method for a retargeting application with insightful analyses.

1. Introduction

Cameras and display devices are designed depending on diverse targets of customers' needs, and hence the resolution and aspect ratio of each module are different. Considering a full screen display scenario of an image, the original image may not perfectly fit the display in full screen, due to the different aspect ratios between the display device and the image. It may rather introduce clipping, stretching or shrinking, as shown in Fig. 1-(b).

Image retargeting research has been actively carried out in computer vision and image processing. Image retargeting techniques adjust the aspect ratio (or size) of an image to fit the target aspect ratio, while not discarding important content in an image. *Content-aware image retargeting* aims to preserve important content as much as possible, and various methods have been suggested [25, 42, 41, 49, 15, 18, 1, 32, 35, 16, 10, 6].

In traditional content-aware approaches, salient region information is widely used. The saliency is conventionally generated from edge map or hand-crafted features, which are restricted by fixed design principles and limit

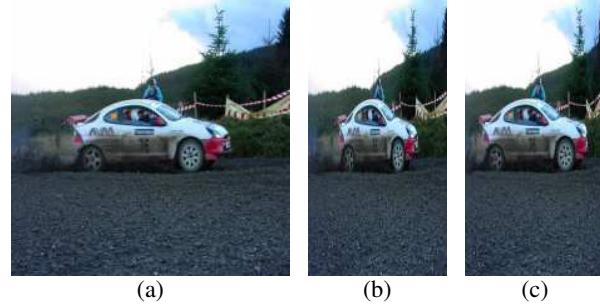


Figure 1: Content-aware retargeting examples. (a) An input image. (b) Linear scaling. (c) Content-aware retargeting (OURS). Our method tends to preserve important parts of the given image well.

its generality. In recent years, deep learning studies have been actively conducted to extract high-level semantic features [34, 38, 40, 14, 13] of an image. Furthermore, research on deep learning has shown excellent performance in low-level image processing tasks [43, 44, 7, 30, 5, 45, 27, 24].

Motivated by this, in this paper, we propose a content-aware image retargeting method using high-level semantic information through deep learning. In order to resize input images within a network, we introduce a shift layer that maps each pixel from the source to the target grid. We demonstrate that end-to-end training is possible through the shift layer, and retargeted images are directly produced using input images with target aspect ratios. The spatial semantic information is learned from image-level annotations, and passed to the shift layer. Image-level annotations are used to compute content loss on the retargeted images. Moreover, input images are used for structure loss to suppress unwanted visual effects after retargeting. In summary, this paper provides the following contributions:

1. We introduce a weakly- and self-supervised learning for content-aware deep image retargeting. We utilize images and its corresponding image-level annotations for structure and content loss computations, respectively. They

- do not require much human effort for output labels to supervise network.
2. Our proposed network takes a source image and a target aspect ratio as input, and then directly produces a retargeted image in a shot. Therefore, end-to-end training is possible and its test time is also fast. To our knowledge, this work is the first attempt to apply deep learning to the image retargeting application.
 3. We design a shift layer that maps each pixel from the source to the target grid. Our method implicitly learns semantic attention information, and passes it to the shift map.

2. Related Works

In this section, we review previous studies related to this paper. In particular, we review recent image processing research based on deep CNN models, and then review image retargeting approaches.

Deep CNN Model for Low-Level Vision Tasks Recently, there have been many works related to CNN for computer vision problems. We introduce a few insightful works that are relevant to image processing problems, and all of these pioneer works have demonstrated successful results: image denoising [17, 43], image artifact removal [8, 4], super resolution [5, 21, 20], deblurring [44, 39], colorization [48, 50], image inpainting [43, 28, 46], and image matting [36, 2]. Interestingly, researchers have shed little light on image retargeting based on a deep CNN approach. We sketch the advances of image retargeting works in a categorical manner as follows.

Image Retargeting Approaches for image retargeting can be categorized in a variety of ways, but we roughly classify them into seam carving based and warping based methods. The *seam carving based* methods [1, 32, 35, 16, 10] change the aspect ratio of an image by repeatedly removing or inserting seams at unimportant areas. Avida *et al.* [1] introduce the concept of seam carving and solve it using dynamic programming. Rubinstein *et al.* [32] later apply seam carving in 3D volume for video retargeting. A representation of multisize media is defined by Shamir *et al.* [35] to have continuous resizing ability in real time. Han *et al.* [16] find multiple seams simultaneously with region smoothness and seam shape prior. Frankovich and Wong [10] propose an enhanced seam carving method by incorporating energy gradient information into optimization framework.

The *warping based* approaches [25, 42, 41, 49, 15, 18] continuously transform an input image into an image of a target size. In Liu and Gleicher [25], a non-linear image

warping is used to emphasize important parts of an image. Wolf *et al.* [42] try to reduce distortion by shrinking less important pixels and preserving important regions. A scale-and-stretch warping method is proposed by Wang *et al.* [41]. Their method iteratively computes optimal scaling factors for local regions and updates a warped image assisted by edge and saliency map. Guo *et al.* [15] suggest a mesh representation based on image structures to preserve the shape of an input image. An interactive content-aware retargeting is proposed by Jin *et al.* [18]. They formulate retargeting using a sparse linear system based on a triangular mesh, and then solve it via quadratic optimization.

Studies that have been conducted thus far have used salient region information of an image for content-aware image retargeting. However, in most seam carving or warping based methods, previous studies have used hand-crafted features such as edge maps to find prominent areas of an image. In this paper, we apply deep CNN for image retargeting to reflect more semantic information by utilizing many images for training. Our method directly produces a retargeted image from an input image with a target size, and it does not require iterative processes. As far as we are aware, this is the first work that seriously addresses the content-aware image retargeting through the weakly- and self-supervised deep CNN model.

3. Architecture

In this section, we introduce our WSSDCNN model for image retargeting. The overall architecture consists mainly of two parts: an encoder-decoder (Sec. 3.1) and a shift layer (Sec. 3.2) as illustrated in Fig. 2. The aspect ratio can be adjusted by changing the width or height of an image. For simplicity, we describe only the case of reducing the width of images, but one can understand the other cases analogously with simple modification, *e.g.*, height adjustment can be done by rotating an image by 90 degrees.

3.1. Encoder-Decoder Model

The most important information for the content-aware image retargeting is semantic information for each object in the scene. Recently, some of previous works [23, 47] have discovered that activations from high-level layers tend to encode semantic information (*i.e.* high-level information). Based on this observation, in order to utilize semantic information, the encoder of our network takes an input image and extract high-level features containing semantic information. The decoder generates an attention map using high-level features from the encoder. We adopt pre-trained VGG16 [38] and inversely symmetric VGG16 architectures for the encoder and decoder, respectively. In addition, we remove fully connected layers of VGG16 and replace ReLU layers in the decoder with ELU [3] layers.

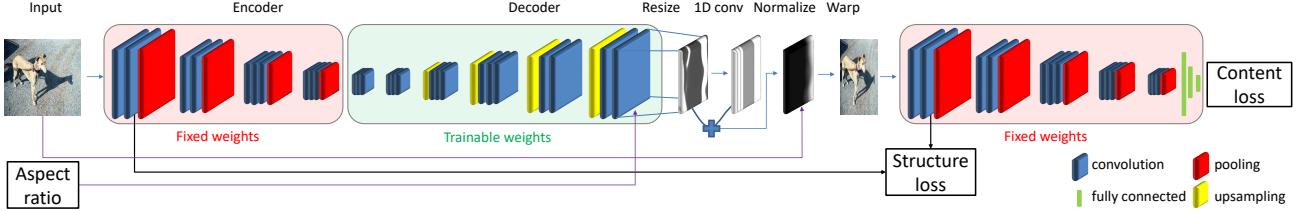


Figure 2: Overall architecture of the proposed network for image retargeting.

3.2. Shift Layer

The shift map defines to what extent a pixel should be shifted from the input of size $W \times H$ to specified output grids of size $W' \times H$ for each pixel. The relationship between the input and output images via shift map is

$$O(x, y) = I(x + S(x, y), y), \quad (1)$$

where I and O denote input and output images and (x, y) are spatial coordinates of the output grid, respectively. $S \in [0, W - W']^{W' \times H}$ is the shift map for the output grid.

We utilize an attention map from the decoder, whereby the shift map acts in a semantic aware manner. First, we resize the attention map to target size.

$$A_r = \text{resize}(A_d, R), \quad (2)$$

where A_r , A_d , and R denote a resized attention map, an output from decoder, and a target aspect ratio, respectively. We then feed it into 1D duplicate convolution and cumulative normalization layers, which are described as follows.

1D Duplicate Convolution An attention map has rough localization information about discriminative parts in an image. By transforming semantic driven location information through the shift map, semantically important parts in an image are preserved while the scales of background regions are adjusted. However, in order to maintain the overall shape of an image, pixels in similar columns should have similar shift values. Therefore, we constrain the shape of an attention response to be uniform along the column axis using 1D duplicate convolution layer:

$$A_{1D} = \text{duplication}(\text{conv}_{1D}(A_r, H), H), \quad (3)$$

where $\text{conv}_{1D}(\cdot, H)$ is a convolution with a H dimensional column vector without padding, and $\text{duplication}(\cdot, H)$ repeats one dimensional vector H times as shown in Fig. 3. All weights of $\text{conv}_{1D}(\cdot, H)$ are learned by training.

Since A_{1D} is restricted to be just column-wise map, to handle residual cases, we use the combination of A_r and A_{1D} as a final attention map.

$$A = \lambda A_r + A_{1D}, \quad (4)$$

where λ is a balancing parameter for A_r and A_{1D} . The rationale behind this composition is rather emblematic. By

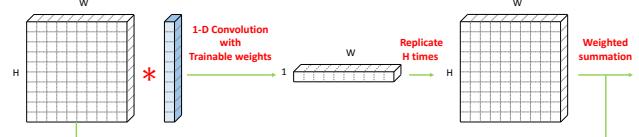


Figure 3: Illustration of 1D duplicate convolution. It reduces visual artifacts after retargeting effectively as shown in Fig. 4.

restricting the structure of attention map A_{1D} , it captures the majority of the attention while making the final map vertically structured, whereas the role of A_r is reduced to draw residual attention, which might be necessary. If we directly use an attention map A_r from the decoder, background areas are twisted around discriminative objects, as shown in Fig. 4-(b).

Cumulative Normalization In order to recover the final mapping from the input grid to the target grid, we convert A into shift map S . Since S should be monotonically increased along the spatial axes, we perform cumulative normalization to constrain it.

$$S(x, y) = \alpha \cdot \frac{\sum_{x' \leq x} A(x', y)}{\sum_x A(x, y)}, \quad (5)$$

where α is $|W - W'|$. From the shift map $S(x, y)$, we retarget an input image to the target grid using Eq. (1).

Image Warping Finally, by Eq. (1), an input image is warped into an image of a target size. Because the shift map has subpixel precision, linear interpolation is performed using four neighboring pixels. As mentioned in [11], applying a differentiable loss to a linearly warped image is also differentiable. To further proceed, when a converted image is passed to a pre-trained network to compute content and structure losses, which are described in the following section, zero padding is performed to fit the input size of the pretrained model.

4. Losses

In order to train the shift map, we utilize two types of supervisions. Image-level annotations and input images themselves are used for content and structure loss calculation, respectively.

4.1. Content Loss

If the main content in an image is well preserved after retargeting, its retargeted image should have a similar outcome of classification using the original image. Taking this point into consideration, we define content loss. Unlike content loss used in [12, 19] to represent reconstruction errors that measure activation differences of the target and the prediction, we use a class score loss for the classification task as content loss in this paper. The content loss is calculated on the retargeted image, and defined as per-class cross entropy loss on sigmoid outputs from a classifier.

$$E_c = -\frac{1}{CN} \sum_{k=1}^C \sum_{i=1}^N \left[l_{ki} \log \hat{l}_{ki} + (1 - l_{ki}) \log (1 - \hat{l}_{ki}) \right], \quad (6)$$

where C and N are the number of classes and samples. l_{ki} and \hat{l}_{ki} are the ground truth label and sigmoid output, respectively. We use VGG16 [38] as a classifier model, and all weights are initialized by pre-trained values, and fixed during both training and test phases.

In order for classification to work well after retargeting, discriminative parts of an image have to be well preserved. Therefore, by content loss, the shift map is trained so as to implicitly contain information about prominent areas of an image.

4.2. Structure Loss

While preserving the most important parts of the given image, we want to suppress unnatural visual artifacts such as distortion. In order to make the retargeted image have a similar structure to the original image, we utilize the input image as a supervision for structural similarity. We design structure loss to make the neighborhood of each pixel in the retargeted image to be as similar as possible to the neighborhood of each corresponding pixel in the input image. Since we can infer correspondences between the input image and the retargeted image from the shift map, we can measure the corresponding patch similarity of the input and output. One of the simplest approaches for the similarity measure is pixel-wise 4-neighbor comparison between input and output pixels. However, pixel-wise comparison easily fails with image noise, contrast change and patch misalignment.

As reported in [23, 47], activations from the first few convolutional layers of CNN provide low-level structural information. These activations are robust to image noise, contrast change, and misalignment due to convolution operations. Therefore, we do not utilize pixel-wise difference, but utilize activations from $conv1$ of VGG16 for the com-

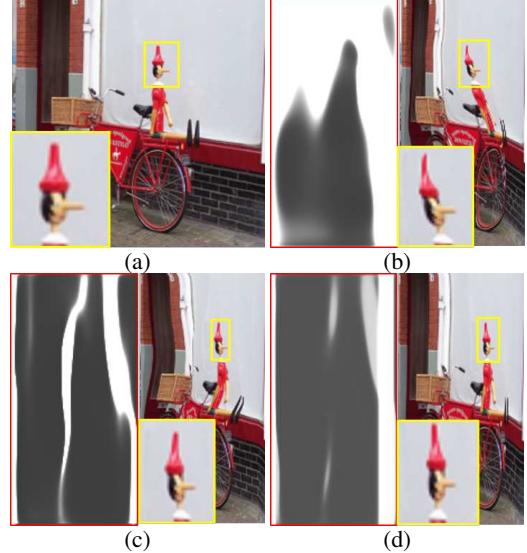


Figure 4: (a) An input image. (b-d) A attention map and a regtargeted image. (b) Content loss. (c) Content loss + structure loss. (d) Content loss + structure loss + 1D convolution.

putation of structural similarity as follows:

$$E_s = \sum_{j=1}^2 \sum_{x,y} [\mathbf{F}_j(O(x,y)) - \mathbf{F}_j(I(x+S(x,y),y))], \quad (7)$$

where \mathbf{F}_j are function of $conv1_j$ in VGG16. Through $conv1_1$ and $conv1_2$, the receptive field of each pixel covers a 5×5 neighborhood.

The network trained with structure loss makes pixels in similar columns of an attention map to have similar values, as shown in Fig. 4-(c). This is desirable in the 1D duplicate convolution layer, and, as a result, artifacts of a retargeted image are significantly reduced compared to the case using only content loss. Also, combining the structure loss with the 1D duplicate convolution layer produces better results, as in Fig. 4-(d). In short, content loss plays a role of preserving salient objects, and the structure loss and the 1D duplicate convolution layer reduce the artifact of retargeted images.

5. Experiments

In this section, we first describe implementation details. Qualitative comparisons of deep content-aware image retargeting to previous works are also shown. After that, we analyze the role and effect of the shift map of our algorithm.

5.1. Training

We train our proposed network using Pascal VOC 2007 dataset [9] with only image-level annotations. First of all, we train the VGG16 model using VOC data, and use the

trained weights for initialization of the encoder and classifier parts. They are fixed during the training of the other parts. The other parts of the network are trained with content and structure losses. Input images are resized to 224×224 , and input aspect ratios are randomly generated for each batch within $\frac{224}{4} \sim \frac{224}{2}$. It takes around $1 \sim 2$ days for 300 epochs on a machine with a GTX 1080 GPU and an Intel i7 3.4GHz CPU. The learning rate, momentum and batch size are set to 10^{-5} , 0.9, and 16, respectively. A forward pass takes around $0.5 \sim 1$ seconds to process an image with a resolution of 224×224 pixels. Except for the classifier used in training, the proposed network is fully convolutional, and thus an image of arbitrary size can be retargeted at the test time.

5.2. Results

VOC Dataset Fig. 5 shows the width retargeting results on VOC 2007 test images. All results are reduced to half size. We compare our results with linear scaling, manual cropping, and seam carving [1]. In addition, results with only content loss, and results with content loss+structure loss are compared to verify the effects of structure loss and 1D duplicate convolution layer. In linear scaling (Fig. 5-(b)), the size of the object can be excessively reduced while cropping (Fig. 5-(c)) takes away important regions of an image. Since seam carving (Fig. 5-(d)) subtracts the discretized seam, if the reducing factor is large or objects occupy a large portion of an image, important parts are also removed and distorted. Also, as mentioned above, with only content loss (Fig. 5-(e)), objects are preserved but distorted. Structure loss (Fig. 5-(f)) suppresses much of the distortion, but there are still wobbles. Through the 1D duplicate convolution layer, however, wobbles are significantly restored, as in Fig. 5-(g). In addition, height retargeting results are shown in Fig. 6. Our results are also better than the results of linear scaling, manual cropping and seam carving. Results of progressively reducing the size of an image are shown in Fig. 7. Even though the size of an image can be changed only in discrete pixels, we can make a continuous changing effect if the aspect ratio is densely sampled.

Interestingly, semantic objects often disappear after seam carving. This is because seam carving depends on low-level features such as edge maps, as well as other existing content-aware image retargeting methods. For instance, in the first and second columns of Fig. 6, the results of seam carving show that a bird and a bicyclist are severely damaged. This is because seam carving, which relies on low-level features, cannot reasonably infer semantic information in complex background images. This problem can be solved by using a high-level feature using deep learning as shown in results of our proposed method.

Benchmark Dataset In order to compare our method with more previous works, we directly quote published results from a benchmark [31]. Fig. 8-(a-h) shows example results of previous methods [22, 33, 32, 29, 41, 42] including cropping and linear scaling. Our results are shown in Fig. 8-(i). Although images from [31] do not have exactly the same labels as the VOC dataset, object regions are well found, and our results are competitive with previous methods.

Image Enlarging Although the proposed network is trained to adjust the aspect ratio of an image by reducing the size of an image, it is also possible to appropriately enlarge an image by using an attention map. In the case of image enlarging, since the main objects should be fixed as much as possible while textures and background regions are expanded, an attention map is reversed to map an input to target grid as follow:

$$A^{-1} = \exp\left(-\frac{A}{\gamma}\right), \quad (8)$$

where γ is a parameter of inversion. After cumulative normalization, the final shift map is obtained. In order to expand an image by k , we first obtain an attention map for the task of reducing an input by k , and then resize it to the input size. For image enlarging, mapping between an input of size $W \times H$ and an output of size $W' \times H$ images is determined as follows:

$$O'(x' + S(x', y'), y') = I(x', y'), \quad (9)$$

where (x', y') are spatial coordinates of the input grid. In this case, S ranges from 0 to $W' - W$. Fig. 9 shows examples of image enlarging. Compared to linear scaling, the main objects are much better preserved while background regions are expanded.

User Study In order to evaluate our proposed method in terms of pleasantness to the human eye, we conduct a user study with 32 people from different backgrounds (age range 24-55) using 30 sets of pascal (20) and non-pascal (10) images. For comparison with the proposed method, linear scaling, center crop, edge crop, seam carving, BDS [37], and AAD [26] methods are used. Each person is guided to select two preferred images over the seven choices. As shown in Fig. 10, our proposed method receives the highest vote. Also, considering the linear scale method is a moderate baseline, it is qualitatively meaningful that our proposed method receives the most votes.

5.3. Analysis

Classification on Retargeted Images In order to quantitatively confirm that the main objects are well preserved

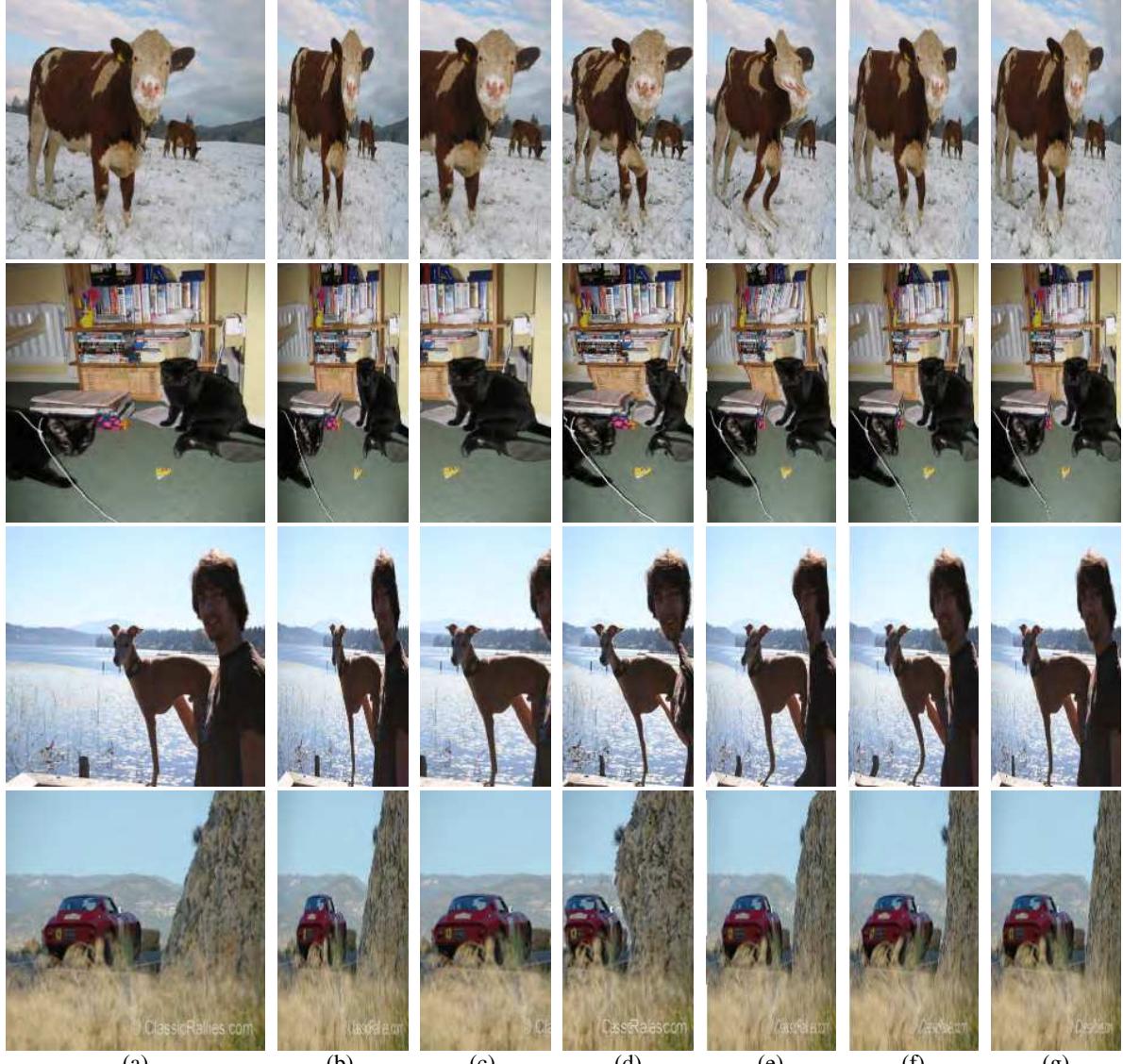


Figure 5: Width adjustment. (a) Input images. (B) Linear scaling. (c) Manual Cropping. (d) Seam carving [1]. (e) Results with content loss. (f) Results with content loss + structure loss. (g) Results with content loss + structure loss + 1D duplicate convolution.

after retargeting, we experiment on how much the classification accuracy differs before and after retargeting. Experiments are conducted on the VOC 2007 test set. For the purpose of fair comparison, we use pre-trained Alexnet and VGG19 instead of VGG16, so that we can evaluate the generality while avoiding any bias induced by VGG16, which was used in our model. Since the VOC dataset has multiple classes in an image, we measure mAP instead of accuracy, and report the ratio with classification results against original images, *i.e.*, $\frac{\text{mAP}(I_{\text{retargeted}})}{\text{mAP}(I_{\text{original}})}$. Images with reduced size are linearly interpolated to their original size before classification. Note that we use zero padding when retargeted images

are passed to the classifier at training time. We compare the mAP ratio of the proposed method with the retargeting methods using center cropping, edge-based cropping, and seam carving [1]. Edge-based cropping is a way to crop a window containing the largest amount of gradients within the given window size. As shown in Fig. 11, the mAP ratio is measured while reducing image sizes from 0.7 to 0.3 times at intervals of 0.1. Our method shows significantly slower decay of performance than the other methods. Through these quantitative experiments, we verify that the proposed method can adjust the size of an image while preserving the main content of an image well.

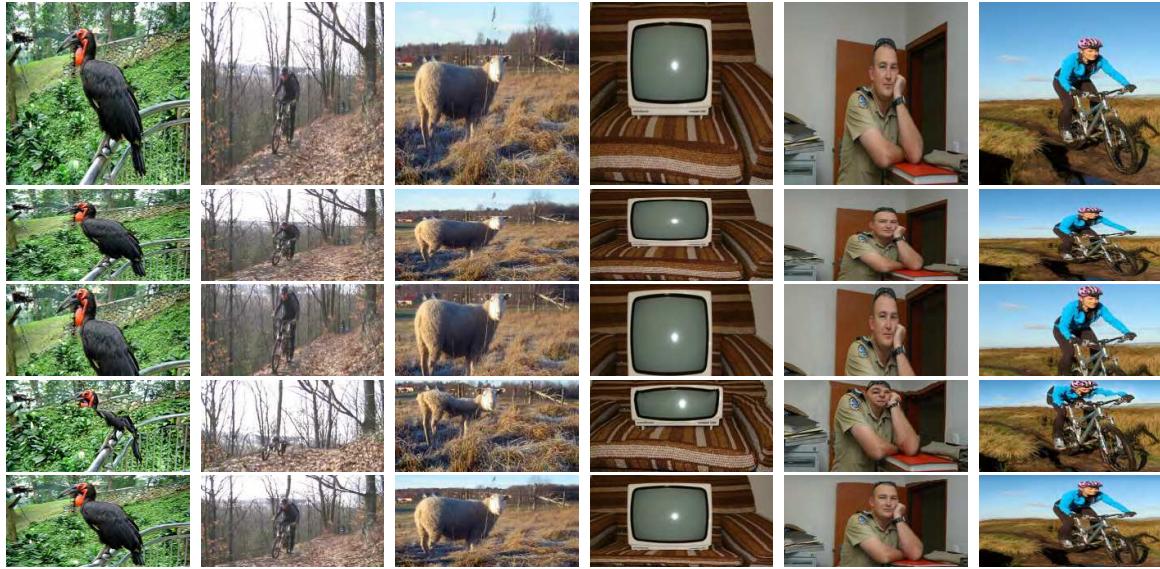


Figure 6: Height adjustment. From the top to the bottom: input images, linear scaling, manual cropping, seam carving [1], and our results (content loss + structure loss + 1D duplicate convolution).



Figure 7: Retargeted images according to varying aspect ratio from 0.9 to 0.3 times width. (Top) Linear scaling. (Bottom) Our results (content loss + structure loss + 1D duplicate convolution).

Discussion Alternative supervision approaches can be conceived. As a specific potential alternative, one may think of directly supervising pixel-level attention map (or shift map) to provide more precise location information. However, it is totally anecdotal how to annotate while incorporating semantic information. Also, if we create directly retargeted images using pixel-level annotations and use them as output, we have to make images with all aspect ratios. Also, it is not desirable to simply set the results of image retargeting as a unique result. In this regard, we decide to use only image-level annotations as supervision, and try to resolve artifacts such as wobbles through the 1D duplicate convolution layer and structure loss.

When retargeting is performed, regions with a high value

in the attention map are reduced more than areas with a low value. As in Fig. 4-(c,d), when we look at the attention map, high values are formed in the shape of curved lines along the background, which is similar to seams to be removed in seam carving. That is, our proposed method shrinks less important parts, similar to seam carving, but uses high-level information induced by semantic understanding.

An image without dominant objects (e.g. landscape) can be regarded as a limitation because the prominent object in an image is either too small or does not appear. However, the network can still perform reasonably well because it has higher responses in textural areas in comparisons with textureless areas.

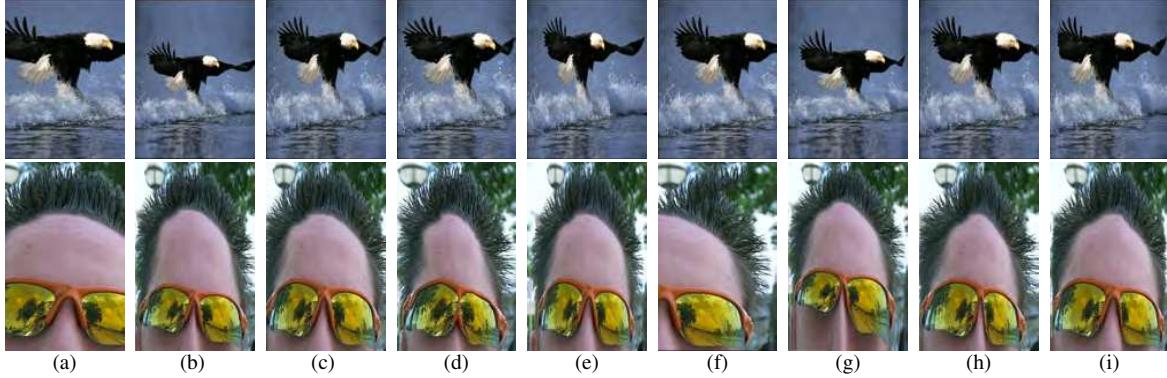


Figure 8: Comparisons on benchmark in [31] (a) Manual Cropping. (b) Streaming Video [22]. (c) Multi-operator [33]. (d) Seam carving [32]. (e) Linear scaling. (f) Shift map [29]. (g) Scale-and-Stretch [41]. (h) Nonhomogeneous warping [42]. (i) Our results (content loss + structure loss + 1D duplicate convolution).

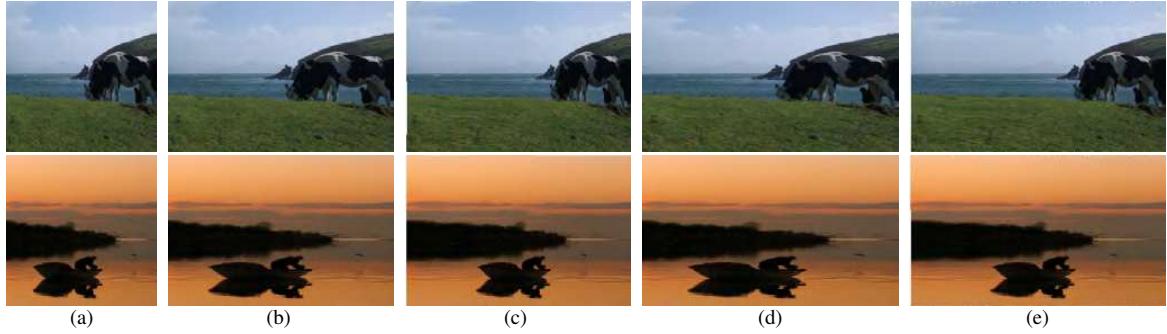


Figure 9: Enlarging examples. (a) Input images. (b-e) 1.5 and 1.7 times enlarged images using linear scaling and our method. (b) Linear scaling (1.5 \times). (c) Our method (1.5 \times). (d) Linear scaling (1.7 \times). (e) Our method (1.7 \times).

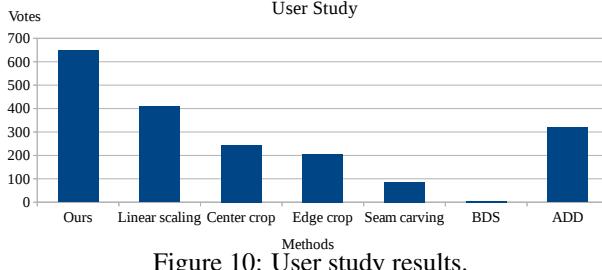


Figure 10: User study results.

6. Conclusion

In this paper, we have proposed a weakly- and self-supervised deep convolutional neural network (WSSD-CNN) for content-aware image retargeting. Our network produces a retargeted image directly, given an input image and a target aspect ratio. The network architecture consists of an encoder-decoder structure for the attention map generation, and the shift map is generated from 1D duplicate convolution and cumulative normalization layers. Because our network is trained in an end-to-end manner using only image-level annotations, we do not have to collect a large amount of finely annotated image data. The proposed content and structure losses ensure a retargeted image pre-

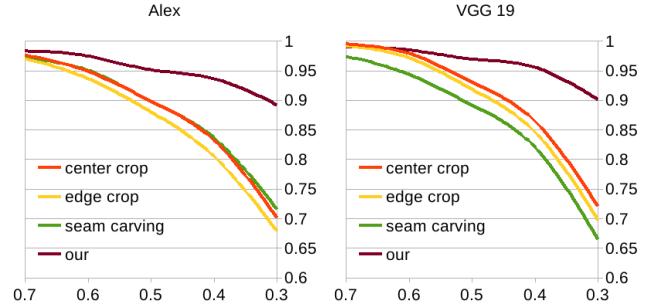


Figure 11: Transition graph of mAP ratio according to the resizing scale. X-axis: scale ratio of image retargeting, Y-axis: mAP ratio against classification results of original images.

serves important objects, and also reduce visual artifacts such as structure distortion. Experimental results demonstrate that our algorithm is superior to previous works, and shows more visually pleasing qualitative results.

Acknowledgements We thank the anonymous reviewers for their valuable comments, and all the participants in our user study.

References

- [1] S. Avidan and A. Shamir. Seam carving for content-aware image resizing. *ACM Trans. on Graph. (ToG)*, 26(3):10, 2007. 1, 2, 5, 6, 7
- [2] D. Cho, Y.-W. Tai, and I. S. Kweon. Natural image matting using deep convolutional neural networks. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2016. 2
- [3] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015. 2
- [4] C. Dong, Y. Deng, C. C. Loy, and X. Tang. Compression artifacts reduction by a deep convolutional network. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2015. 2
- [5] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2014. 1, 2
- [6] W. Dong, F. Wu, Y. Kong, X. Mei, T.-Y. Lee, and X. Zhang. Image retargeting by texture-aware synthesis. *IEEE Trans. on Visualization and Computer Graphics (TVCVG)*, 22(2):1088–1101, 2016. 1
- [7] D. Eigen, D. Krishnan, and R. Fergus. Restoring an image taken through a window covered with dirt or rain. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2013. 1
- [8] D. Eigen, D. Krishnan, and R. Fergus. Restoring an image taken through a window covered with dirt or rain. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2013. 2
- [9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>. 4
- [10] M. Frankovich and M. Alexander Wong. Enhanced seam carving via integration of energy gradient functionals. *IEEE Signal Processing Letters*, 18(6):375 – 378, 2011. 1, 2
- [11] R. Garg, V. K. B. G, and I. D. Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2016. 3
- [12] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2016. 4
- [13] R. Girshick. Fast r-cnn. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2015. 1
- [14] R. B. Girshick, J. Donahue, Trevor Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2014. 1
- [15] Y. Guo, F. Liu, J. Shi, Z.-H. Zhou, and M. Gleicher. Image retargeting using mesh parametrization. *Trans. Multi.*, 11(5):856–867, 2009. 1, 2
- [16] D. Han, M. Sonka, J. Bayouth, and X. Wu. Optimal multiple-seams search for image resizing with smoothness and shape prior. *The Visual Computer*, 26(6):749–759, 2010. 1, 2
- [17] V. Jain and S. Seung. Natural image denoising with convolutional networks. In *Proc. of Neural Information Processing Systems (NIPS)*, 2008. 2
- [18] Y. Jin, L. Liu, and Q. Wu. Nonhomogeneous scaling optimization for realtime image resizing. *The Visual Computer*, 26(6):769–778, 2010. 1, 2
- [19] J. Johnson, A. Alahi, and F. Li. Perceptual losses for real-time style transfer and super-resolution. *CoRR*, abs/1603.08155, 2016. 4
- [20] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. *CoRR*, abs/1511.04587, 2015. 2
- [21] J. Kim, J. K. Lee, and K. M. Lee. Deeply-recursive convolutional network for image super-resolution. *CoRR*, abs/1511.04491, 2015. 2
- [22] P. Krähenbühl, M. Lang, A. Hornung, and M. Gross. A system for retargeting of streaming video. *ACM Trans. on Graph. (ToG)*, 28(5):126:1–126:10, 2009. 5, 8
- [23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. of Neural Information Processing Systems (NIPS)*, 2012. 2, 4
- [24] G. Lee, Y.-W. Tai, and J. Kim. Deep saliency with encoded low level distance map and high level features. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2016. 1
- [25] F. Liu and M. Gleicher. Automatic image retargeting with fisheye-view warping. In *Proc. of User Interface Software Technologies (UIST)*, 2005. 1, 2
- [26] D. Panozzo, O. Weber, and O. Sorkine. Robust image retargeting via axis-aligned deformation. *Comput. Graph. Forum*, 31(2):229–236, 2012. 5
- [27] J. Park, Y.-W. Tai, D. Cho, and I. S. Kweon. A unified approach of multi-scale deep and hand-crafted features for defocus estimation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2017. 1
- [28] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros. Context encoders: Feature learning by inpainting. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2016. 2
- [29] Y. Pritch, E. Kav-Venaki, and S. Peleg. Shift-map image editing. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, pages 151–158, 2009. 5, 8
- [30] J. S. Ren, L. Xu, Q. Yan, and W. Sun. Shepard convolutional neural networks. In *Proc. of Neural Information Processing Systems (NIPS)*, 2015. 1
- [31] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir. A comparative study of image retargeting. *ACM Trans. on Graph. (ToG)*, 29(6):160:1–160:10, 2010. 5, 8
- [32] M. Rubinstein, A. Shamir, and S. Avidan. Improved seam carving for video retargeting. *ACM Trans. on Graph. (ToG)*, 27(3):16:1–16:9, 2008. 1, 2, 5, 8
- [33] M. Rubinstein, A. Shamir, and S. Avidan. Multi-operator media retargeting. *ACM Trans. on Graph. (ToG)*, 28(3):1–11, 2009. 5, 8
- [34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *Int'l Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1

- [35] A. Shamir and S. Avidan. Seam carving for media retargeting. *Commun. ACM*, 52(1):77–85, 2009. [1](#), [2](#)
- [36] X. Shen, X. Tao, H. Gao, C. Zhou, and J. Jia. Deep automatic portrait matting. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2016. [2](#)
- [37] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani. Summarizing visual data using bidirectional similarity. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2008. [5](#)
- [38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. [1](#), [2](#), [4](#)
- [39] J. Sun, W. Cao, Z. Xu, and J. Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2015. [2](#)
- [40] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2015. [1](#)
- [41] Y.-S. Wang, C.-L. Tai, O. Sorkine, and T.-Y. Lee. Optimized scale-and-stretch for image resizing. *ACM Trans. on Graph. (ToG)*, 27(5):118:1–118:8, 2008. [1](#), [2](#), [5](#), [8](#)
- [42] L. Wolf, M. Guttmann, and D. Cohen-Or. Non-homogeneous content-driven video-retargeting. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2007. [1](#), [2](#), [5](#), [8](#)
- [43] J. Xie, L. Xu, and E. Chen. Image denoising and inpainting with deep neural networks. In *Proc. of Neural Information Processing Systems (NIPS)*, 2012. [1](#), [2](#)
- [44] L. Xu, J. S. Ren, C. Liu, and J. Jia. Deep convolutional neural network for image deconvolution. In *Proc. of Neural Information Processing Systems (NIPS)*, 2014. [1](#), [2](#)
- [45] L. Xu, J. S. Ren, Q. Yan, R. Liao, and J. Jia. Deep edge-aware filters. In *Proc. of Int'l Conf. on Machine Learning (ICML)*, 2015. [1](#)
- [46] R. Yeh, C. Chen, T. Lim, M. Hasegawa-Johnson, and M. N. Do. Semantic image inpainting with perceptual and contextual losses. *CoRR*, abs/1607.07539, 2016. [2](#)
- [47] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus. Deconvolutional networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2010. [2](#), [4](#)
- [48] B. S. Zezhou Cheng, Qingxiong Yang. Deep colorization. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2015. [2](#)
- [49] G.-X. Zhang, M.-M. Cheng, S.-M. Hu, and R. R. Martin. A shape-preserving approach to image resizing. *Computer Graphics Forum*, 28(7):1897–1906, 2009. [1](#), [2](#)
- [50] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. *arXiv preprint arXiv:1603.08511*, 2016. [2](#)