

DeepIR: A Deep Semantics Driven Framework for Image Retargeting

Jianxin Lin, Tiankuang Zhou, Zhibo Chen

University of Science and Technology of China, Hefei, China
{linjx, zhoutk}@mail.ustc.edu.cn, chenzhibo@ustc.edu.cn

Abstract

We present *Deep Image Retargeting (DeepIR)*, a coarse-to-fine framework for content-aware image retargeting. Our framework first constructs the semantic structure of input image with a deep convolutional neural network. Then a uniform re-sampling that suits for semantic structure preserving is devised to resize feature maps to target aspect ratio at each feature layer. The final retargeting result is generated by coarse-to-fine nearest neighbor field search and step-by-step nearest neighbor field fusion. We empirically demonstrate the effectiveness of our model with both qualitative and quantitative results on widely used RetargetMe dataset.

1 Introduction

The heterogeneity of the display devices have imposed indispensable demand for appropriately adapting image into screens with different resolutions. The image retargeting techniques meanwhile have been proposed to resize images to arbitrary sizes while keeping the important content of original images. These content-aware image retargeting methods, such as seam carving (Avidan and Shamir 2007), multi-operator (Rubinstein, Shamir, and Avidan 2009; Zhu et al. 2016), and streaming video (Krähenbühl et al. 2009), try to resize the image to target resolution by preserving the important content information and shrinking the unimportant regions of the original image.

Traditional content-aware methods usually utilize one saliency map to define the significance of each pixel. However, saliency detection is designed from attention mechanism, one saliency map has its limitations not only on representing the high-level semantic content, but also on integrating both high-level semantic content and low-level details. In recent years, Convolutional Neural Network (CNN) has shown its superior performance in many high-level computer vision problems (He et al. 2016; Long, Shelhamer, and Darrell 2015) and low-level image processing problems (Ledig et al. 2016; Lin, Zhou, and Chen 2018; Mao, Shen, and Yang 2016). One main advantage of CNN is its better ability to extract multi-level semantic information and better representations for semantic structures (Zeiler and Fergus 2014). Therefore, how to utilize CNN to guide image retargeting is interesting and important to explore.

In this paper, we propose a new framework for content-aware image retargeting. Instead of applying image retarget-

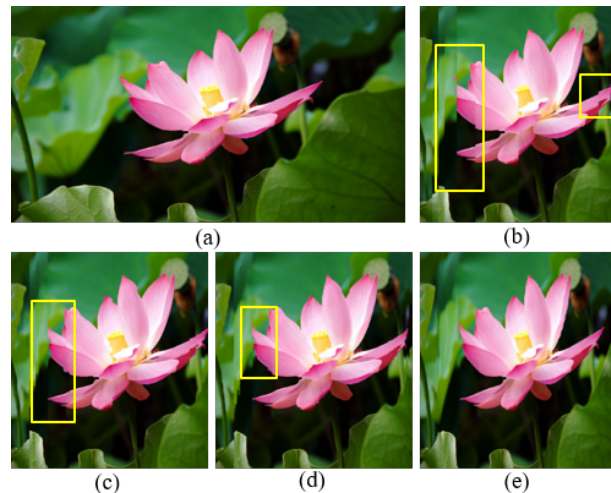


Figure 1: Illustration of DeepIR on image retargeting. (a) Original image. (b)-(e) Coarse-to-fine image retargeting refinement. (e) Final retargeted image. At the coarsest layer, the retargeted image (b) maintains the main semantic structure while suffering unsmoothness. The coarse-to-fine refinement preserves the main semantic structure and content smoothness.

ing techniques on image pixel level directly, our approach uses a pre-trained deep CNN, such as VGG-19 (Simonyan and Zisserman 2014), to construct a feature space in which image retargeting is performed. In order to resize original image to the target aspect ratios in the deep feature space, we devise a uniform re-sampling (UrS) that uniformly removes columns/rows in a cumulative columns/rows obscurity map. Such a UrS ensures the semantic structure completeness of resized feature maps, and content smoothness is also preserved in the final retargeted image as illustrated in Figure 1(e). Then a coarse-to-fine nearest neighbor field (NNF) search (Liao et al. 2017) is used to find spatial correspondence between intermediate feature layers of original image and retargeted image. At each layer, two NNFs obtained from reconstructed features and retargeted features respectively are fused to achieve a combination of high level and low level information, which is called as step-by-step

NNF fusion.

The main contribution of our work includes three aspects:

- We propose a new Deep Image Retargeting (DeepIR) framework that retargets images in the deep feature space. We demonstrate that DeepIR is effective for semantic content and visual quality preservation.
- We propose a UrS that suits for retargeting in the deep feature space.
- We propose a step-by-step NNF fusion that effectively combines the high level semantic content and low level details.

The remaining parts are organized as follows. We introduce related work in Section 2 and present the details of our method in Section 3. Then we report experimental results in Section 4 and conclude in Section 5.

2 Related Works

Numerous works have been carried out for image retargeting in the past decades. Unlike traditional image retargeting methods, such as uniform Scaling (SCL), recent developments in image retargeting usually seek to change the size of the image while maintaining the important content. By using face detectors (Viola and Jones 2001) or the visual saliency detection methods (Itti, Koch, and Niebur 1998) to detect important regions in the image, one simple way to resize image is using Cropping (CR) to eliminate the unimportant region from the image. However, directly eliminating region by CR may result in information loss. Seam Carving (SC) (Avidan and Shamir 2007) is proposed to iteratively remove an 8-connected seam in the image to preserve the visual saliency. To avoid drawbacks of using single retargeting method, multi-operator techniques (MULTIOP) (Rubinstein, Shamir, and Avidan 2009; Zhu et al. 2016) combine SC, SCL and CR to resize the image based on the defined optimal energy cost, such as image retargeting quality assessment metrics. Pritch, Kav-Venaki, and Peleg (2009) described a Shift-Map (SM) technique to remove or add band regions instead of scaling or stretching images. All the above methods resize the image by removing discrete regions. Other approaches also put attempts on resizing the image in continuous and summarization perspectives.

Continuous retargeting methods continuously transform image to the target size and have been realized through image warping or the mapping by constraining deformation and smoothness (Wang et al. 2008; Wolf, Guttmann, and Cohen-Or 2007; Krähenbühl et al. 2009; Guo et al. 2009; Lin et al. 2014). Wang et al. (2008) presented a “scale-and-stretch” (SNS) warping method that operates by iteratively computing optimal local scaling factors for each local region and updating a warped image that matches these scaling factors as close as possible. Wolf, Guttmann, and Cohen-Or (2007) described a nonhomogeneous warping (WARP) for video retargeting, where a transformation that respects the analysis shrinks less important regions more than important ones. Krähenbühl et al. (2009) presented a simple and interactive framework called Streaming Video (SV) that

combines key frame based constraint editing with numerous automatic algorithms for video analysis. The key component of their framework is a non-uniform and pixel accurate warping considering automatic as well as interactively defined features. Guo et al. (2009) constructed a saliency-based mesh representation for an image, which enables preserving image structures during retargeting. To avoid object deformation caused by warping, Lin et al. (2014) utilized the information of matched objects rather than that of matched pixels during warping process, which allows the generation of an object significance map and the consistent preservation of objects. In (Karni, Freedman, and Gotsman 2009), authors proposed a energy minimization based shape deformation method (LG) in which the set of “legal transformations” being not considered to be distorted is expressed.

Summarization based retargeting methods (Simakov et al. 2008; Cho et al. 2008; Barnes et al. 2009; Wu et al. 2010) resize image by eliminating insignificant patches and maintaining the coherence between original and retargeted image. Simakov et al. (2008) measured the bidirectional patch similarity (i.e., completeness and coherence) between two images and iteratively change the original image’s size to retargeted image’s. Cho et al. (2008) broke an image to non-overlapping patches and retargeted image is constructed from the patches with “patch domain” constraints. Barnes et al. (2009) proposed a fast randomized algorithm called PatchMatch to find dense NNF for patches between two images, and retargeted image can be obtained by the similar retargeting method in (Simakov et al. 2008). Wu et al. (2010) analyzed the “translational symmetry” widely existed in the real-world images, and summarize the image content based on symmetric lattices.

Most previous content-aware image retargeting algorithms leverage saliency detection or object detection information to select important regions. However, real-world images usually contain complex and abundant semantic information that should be modeled in a more appropriate manner. Recently, CNN has shown its superior performance in many high-level computer vision problems (He et al. 2016; Long, Shelhamer, and Darrell 2015) and low-level image processing problems (Ledig et al. 2016; Lin, Zhou, and Chen 2018; Mao, Shen, and Yang 2016). Cho et al. (2017) first applied deep CNN to image retargeting. A weakly- and self-supervised deep convolutional neural network (WSSDCNN) is trained for shift map prediction. However, quantitative results of WSSDCNN only indicate its superiority over SCL or SC. Compared with WSSDCNN, our method doesn’t need any training procedure when a pre-trained CNN is given. In addition, our DeepIR shows comparable performance against SOTA methods (i.e., MULTIOP, SV) in Section 4.

3 Approach

Figure 2 shows the overall architecture of the proposed DeepIR framework. Our model mainly includes three parts: deep feature construction, deep feature retargeting and retargeted image reconstruction. The details will be presented in the following sections.

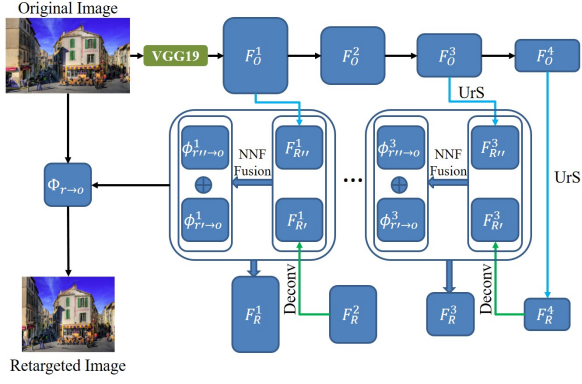


Figure 2: Deep image retargeting (DeepIR) framework.

3.1 Pre-Processing

In order to obtain deep feature space of original image, we utilize the VGG-19 network (Simonyan and Zisserman 2014) that is pre-trained on the ImageNet database (Rusakovsky et al. 2015) as our deep CNN. For original image O , we can obtain a pyramid of its feature maps as $\{F_O^L\}$ ($L = 1 \dots 4$). The reason we choose the first four layers of VGG-19 network as our feature space is that the resolutions of feature maps in the higher layers are too small and it will be too difficult to reconstruct retargeted image based on them. The feature maps of original image are 3D tensors with $h_O^L \times w_O^L \times c_O^L$ for each layer, where h_O^L is height, w_O^L is width and c_O^L is channel number for each F_O^L .

3.2 Uniform Re-Sampling

In order to resize original image to the target aspect ratios in the deep feature space, we devise a uniform re-sampling (UrS) that uniformly removes columns/rows in a cumulative columns/rows obscenity map. As (Zeiler and Fergus 2014) proved that region of higher semantic significance results in stronger activation in feature maps, given feature maps F_O^L , an importance map is first computed as:

$$m_O^L(i, j) = \sum_{c=1}^{c_O^L} F_O^L(i, j, c), \quad (1)$$

where i, j, c are the height, width and channel index respectively. Then without loss of generality, the obscenity map, which gives higher weight to less important (obscure) columns, is computed as:

$$u_O^{L,w}(j) = - \sum_{i=1}^{h_O^L} m_O^L(i, j). \quad (2)$$

Then the obscenity map $u_O^{L,w}$ is normalized by min-max normalization, expressed as $\hat{u}_O^{L,w}$. After that, the cumulative obscenity map is obtained as:

$$s_O^{L,w}(j) = \begin{cases} \hat{u}_O^{L,w}(1), & j = 1, \\ s_O^{L,w}(j-1) + \hat{u}_O^{L,w}(j), & j > 1. \end{cases} \quad (3)$$

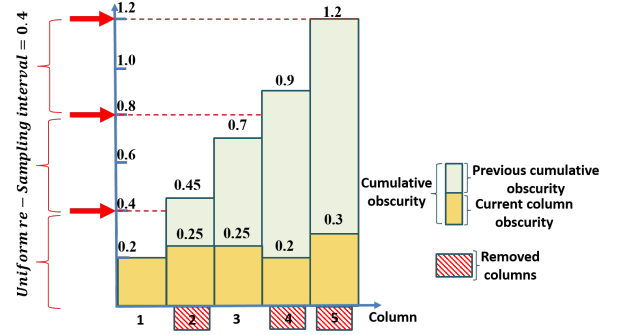


Figure 3: An example of UrS method resizing 5 columns to 2 columns. For each column other than the first one, there are two overlapping bins, in which the lower one represents the normalized obscenity and the higher one represents the cumulative obscenity. Then a uniform sampling is applied on the cumulative obscenity map. Obscenity sum $s(5) = 1.2$. Sampling interval $\tau = s(5)/(5-2) = 0.4$. The bins marked by red tilted lines represent the removed columns.

Given a retargeting aspect ratio ϵ , $(w_O^L - \epsilon w_O^L)$ columns are to be removed. Then a uniform sampling is performed on the $s_O^{L,w}$ with sampling interval $\tau = s_O^{L,w}(w_O^L)/(w_O^L - \epsilon w_O^L)$. And obscure columns \mathcal{R} to be removed are defined as below:

$$\mathcal{R} = \{j | s_O^{L,w}(j-1) \leq r \cdot \tau < s_O^{L,w}(j), \exists r \in \{1, \dots, w_O^L - \epsilon w_O^L\}\}. \quad (4)$$

Denoting all original columns as \mathcal{A} , the preserved columns are $\mathcal{P} = \mathcal{A} - \mathcal{R}$. We call this process as uniform re-sampling that is illustrated in Figure 3. The final retargeted feature maps can be achieved as:

$$F_R^L(i, k, c) = F_O^L(i, p(k), c), \quad k \in \{1, \dots, \epsilon w_O^L\}, \quad (5)$$

where p is a list that contains all the elements in \mathcal{P} and is sorted from small to large.

Since deep CNN feature is extracted by non-linear mappings and cascade convolutions, previous retargeting methods based on minimum importance cost, such as seam carving and column removal, may destroy underlying relationship and distribution of CNN features, which causes structure distortion and/or content loss in the retargeted image (Figure 9). The main advantage of UrS is that it avoids distortions (e.g., content loss, structure distortion) from columns/rows over-removing in regions of low feature importance but high structure importance, and meanwhile preserves the semantic importance of CNN features. We detail the comparison between UrS and other retargeting method when being applied in the DeepIR framework in Section 4.4, which verifies the effectiveness of UrS.

3.3 Reconstruction

After obtaining the retargeted image's feature maps at the highest layer (i.e., F_R^4), the following question is how to propagate the F_R^4 back to the lower layers. As presented in (Liao et al. 2017), this problem can be solved by minimizing the following loss function:

$$\ell_{F_{R'}^{L-1}} = \|\text{CNN}_{L-1}^L(F_{R'}^{L-1}) - F_R^L\|, \quad (6)$$

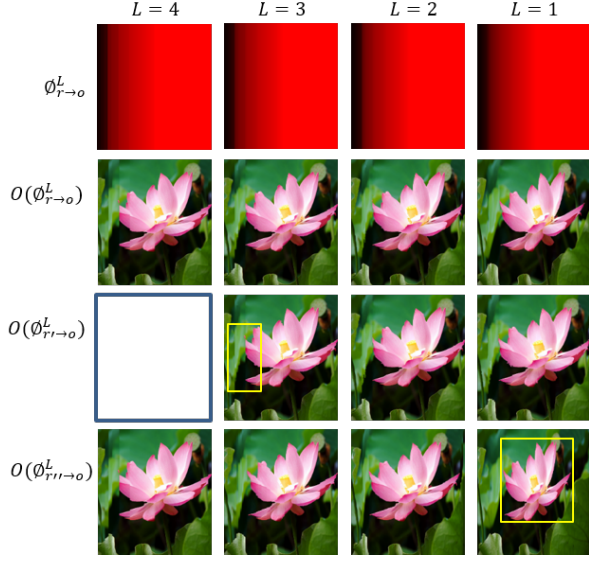


Figure 4: Visualization of coarse-to-fine reconstructed results with step-by-step NNF fusion. First row shows $\phi_{r \rightarrow o}^L$ in each layer. The following three rows show the retargeted images obtained by different NNF mapping functions.

where $F_{R'}^{L-1}$ is the reconstructed feature maps in the level $L-1$, $\text{CNN}_{L-1}^L(\cdot)$ is the L th CNN layer. This optimization problem can be solved by gradient descent (L-BFGS) (Zhu et al. 1997).

In the following three steps, we propose to obtain more appropriate feature maps F_R^{L-1} from consideration of both high-level and low-level information. First, we estimate a NNF mapping function $\phi_{r' \rightarrow o}^{L-1}$ that maps a point in feature map $F_{R'}^{L-1}$ to another in F_O^{L-1} . Second, we obtain the second NNF mapping function $\phi_{r'' \rightarrow o}^{L-1}$ between F_O^{L-1} and $F_{R''}^{L-1}$, where $F_{R''}^{L-1}$ is deep features resized from F_O^{L-1} by UrS. Then we propose to fuse $\phi_{r \rightarrow o}^{L-1}$ and $\phi_{r'' \rightarrow o}^{L-1}$ using a weighted sum:

$$\phi_{r \rightarrow o}^{L-1} = \alpha^{L-1} \phi_{r' \rightarrow o}^{L-1} + (1 - \alpha^{L-1}) \phi_{r'' \rightarrow o}^{L-1}, \quad (7)$$

where α^{L-1} controls the trade-off between current layer's and previous layer's mapping information. Finally, the F_R^{L-1} is obtained by warping F_O^{L-1} with $\phi_{r \rightarrow o}^{L-1}$: $F_R^{L-1} = F_O^{L-1}(\phi_{r \rightarrow o}^{L-1})$. Figure 4 shows that our retargeted results are gradually updated from coarse to fine. We can observe that retargeted image $O(\phi_{r \rightarrow o}^L)$ constructed directly from $F_{R'}^L$ usually lacks the constraint of low-level details and suffers leaf unsmoothness in high layers (e.g., $O(\phi_{r' \rightarrow o}^3)$). Though $O(\phi_{r'' \rightarrow o}^L)$ contains relatively smoother content, it lacks the supervision of high-level semantic information in low layer results (e.g., $O(\phi_{r \rightarrow o}^1)$). So $O(\phi_{r \rightarrow o}^L)$ is the result from achieved balance between high-level semantic content and low-level details.

After we obtain NNF $\phi_{r \rightarrow o}^1$ at the lowest feature layer, we assume that the pixel level mapping function $\Phi_{r \rightarrow o}$ is equal to $\phi_{r \rightarrow o}^1$. Then we reconstruct retargeted image R :

$R = \frac{1}{n} \sum_{x \in N(p)} (O(\Phi_{r \rightarrow o}(x)))$, where $n = 5 \times 5$ is the size of patch $N(p)$.

Algorithm 1 DeepIR Framework

Require: One RGB image O , target aspect ratio ϵ .

Ensure: One pixel-location mapping function $\Phi_{r \rightarrow o}$ and one retargeted image R .

- 1: **Preprocessing:**
 - 2: $\{F_O^L\} (L = 1 \dots 4) \leftarrow$ feeding O to VGG-19.
 - 3: $F_{R''}^4 \leftarrow$ resizing F_O^4 by UrS and ϵ .
 - 4: $F_R^4 \leftarrow F_{R''}^4$.
 - 5: **Reconstruction**
 - 6: **for** $L = 4$ **to** 2 **do**
 - 7: $F_{R'}^{L-1} \leftarrow$ solving loss function Eqn.(6).
 - 8: $F_{R''}^{L-1} \leftarrow$ resizing $F_{R'}^{L-1}$ by UrS and ϵ .
 - 9: $\phi_{r' \rightarrow o}^{L-1} \leftarrow$ mapping $F_{R'}^{L-1}$ to F_O^{L-1} .
 - 10: $\phi_{r'' \rightarrow o}^{L-1} \leftarrow$ mapping $F_{R''}^{L-1}$ to F_O^{L-1} .
 - 11: $\phi_{r \rightarrow o}^{L-1} \leftarrow$ fusing NNFs by Eqn.(7).
 - 12: $F_R^{L-1} \leftarrow F_O^{L-1}(\phi_{r \rightarrow o}^{L-1})$.
 - 13: **end for**
 - 14: $\Phi_{r \rightarrow o} \leftarrow \phi_{r \rightarrow o}^1$.
 - 15: $R \leftarrow \frac{1}{n} \sum_{x \in N(p)} (O(\Phi_{r \rightarrow o}(x)))$.
-

3.4 Algorithm and Performance

The pseudo code of our implementation is shown in Algorithm 1. We have implemented our algorithm and conducted experiments on one NVIDIA K80 GPU. The time of retargeting images in our experiments ranges from 20 seconds to 60 seconds, which depends on the sizes of input images. Specially, the time of PatchMatch ranges from 8 seconds to 16 seconds. The feature deconvolution may require 10 seconds to 35 seconds to converge.

4 Experiment Results

We conduct experiments on the widely used RetargetMe benchmark (Rubinstein et al. 2010). The RetargetMe dataset contains real-world images with various attributes and contents, so it's quite suitable to test the robustness and generality of retargeting method. We provide more experimental results in the supplementary document.

4.1 Qualitative Evaluation

Discrete and continuous image retargeting methods are compared in this section, such as manually chosen Cropping (CR), Streaming Video (SV) (Krähenbühl et al. 2009), multi-operator (MULTIOP) (Rubinstein, Shamir, and Avidan 2009), Seam Carving (SC) (Avidan and Shamir 2007), uniform Scaling (SCL) and Nonhomogeneous warping (WARP) (Wolf, Guttman, and Cohen-Or 2007). We compare our method with these algorithms on RetargetMe images as shown in Figure 5, and Figure 6. We can observe that: (1) Although CR tries to shrink image by removing parts of it, some content of semantic importance is lost, such as streets and buildings in Figure 5(b), and skiers in Figure

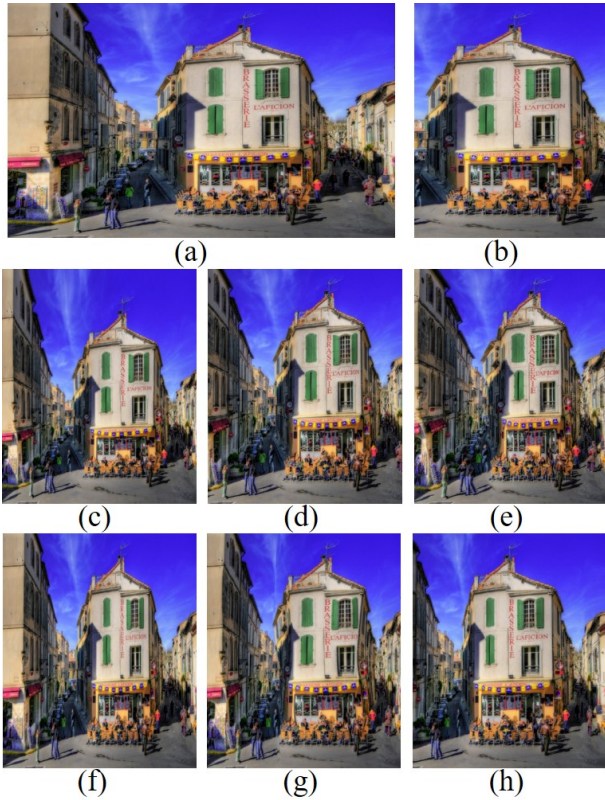


Figure 5: Comparison on “Brasserie_L_Aficion” in RetargeMe. (a) Original image. (b) CR. (c) SV. (d) MULTIOP. (e) SC. (f) SCL. (g) WARP. (h) Ours.

6(b). (2) SV can cause retargeted image to be out of proportion (i.e., some objects in retargeted image are over-shrunk or over-stretched compared to others) since SV requires a fixed global scaling factor for the entire image, the examples are shown in Figure 5(c). (3) SC and WARP tend to create distortions of lines and edges, and deformation of people and objects when there are not enough homogenous contents in the input image as shown in Figure 5(e,g) and Figure 6(e). (4) MULTIOP combines three operators (i.e., CR, SCL and SC) to avoid drawbacks of single operators. Although it outperforms single operators when the scenario is relatively simple (Figure 5(d)), it can not preserve important contents as well as our method due to using low-level saliency detection, as shown in Figure 6(d). (5) By using deep CNN for deep feature extraction, our method can effectively locate semantic contents as shown in Figure 5(h) and Figure 6(h). The results also demonstrate that our method can produce visually preferred results by using UrS and NNF fusion. Note that, although our method utilizes deep CNN for better semantic areas selection, our method does not need any training procedure given a pre-trained deep CNN. So such a framework we propose can be combined with other retargeting algorithms (feasibility is proved in Section 4.4), which bridges the gap between traditional low level retargeting methods and newly-developed deep features.

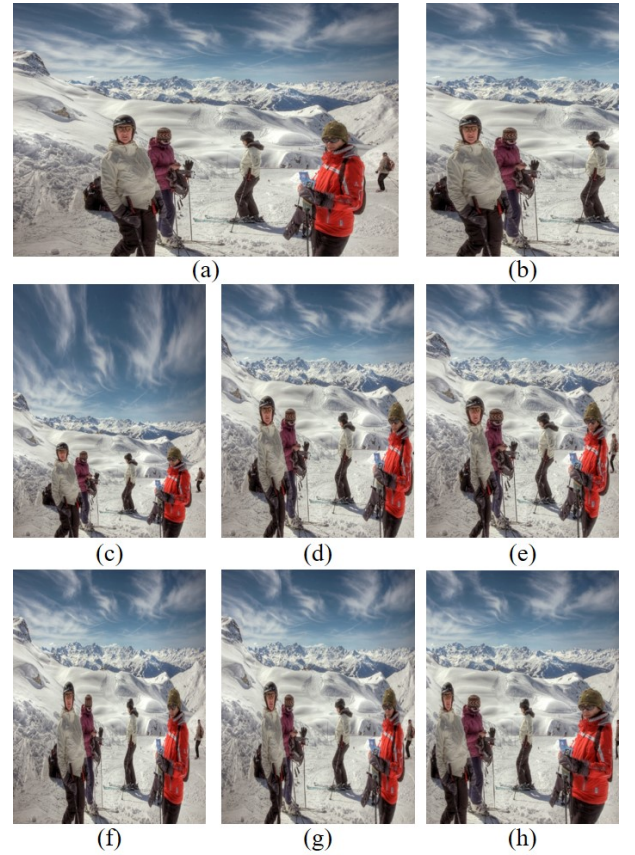


Figure 6: Comparison on “ski” in RetargeMe. (a) Original image. (b) CR. (c) SV. (d) MULTIOP. (e) SC. (f) SCL. (g) WARP. (h) Ours.

4.2 Quantitative Evaluation

	Ours	CR	SV	MULTIOP	SCL
User Study	101	36	98	90	35
FRR	0.5307	0.5239	0.5091	0.5279	0.5215
FD	5.3349	2.8944	6.3555	5.7765	6.338

Table 1: Objective and subjective scores of different methods. Best scores are in bold.

For quantitative evaluation, we compare with SCL and three state-of-the-art retargeting methods, i.e., CR, SV and MULTIOP, as reported in (Rubinstein et al. 2010). We first carry out a user study to assess quality of retargeted images. We ask total 18 subjects (10 males, 8 females, age range 20 – 35) from different backgrounds to make comparison of 20 sets of retargeted images. The retargeted images are selected from RetargetMe of aspect ratio $\epsilon = 0.5$, which is easier for subjects to judge. We show the subjects original image, our result and results from other methods. Then each subject selects the favorite retargeted images over other im-



Figure 7: Influence of aspect ratio ϵ on retargeted results. From left to right and from top to down, the aspect ratio ϵ increases 0.5 for each retargeted image.

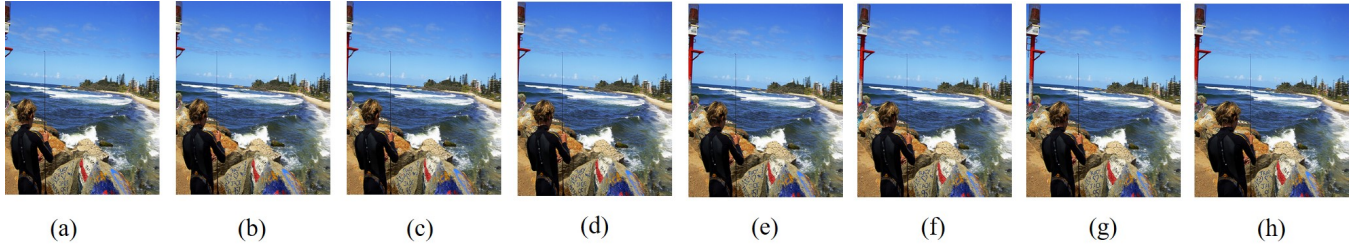


Figure 8: Influence of $\{\alpha^L\}_{L=1}^3$ on retargeted results. (a) $\{\alpha^L\}_{L=1}^3 = \{0.1, 0.2, 0.3\}$. (b) $\{\alpha^L\}_{L=1}^3 = \{0.2, 0.3, 0.4\}$. (c) $\{\alpha^L\}_{L=1}^3 = \{0.3, 0.4, 0.5\}$. (d) $\{\alpha^L\}_{L=1}^3 = \{0.4, 0.5, 0.6\}$. (e) $\{\alpha^L\}_{L=1}^3 = \{0.5, 0.6, 0.7\}$. (f) $\{\alpha^L\}_{L=1}^3 = \{0.6, 0.7, 0.8\}$. (g) $\{\alpha^L\}_{L=1}^3 = \{0.7, 0.8, 0.9\}$. (h) $\{\alpha^L\}_{L=1}^3 = \{0.8, 0.9, 1.0\}$.

ages. Then we also calculate two empirical objective scores: feature remain ratio (FRR), i.e.,

$$\text{FRR} = \frac{1}{4} \sum_{L=1}^4 \frac{\sum_{i,j,c} F_R^L(i,j,c)}{\sum_{i,j,c} F_O^L(i,j,c)}, \quad (8)$$

which measures the proportion of deep features remaining in the retargeted images, and feature dissimilarity (FD), i.e.,

$$\text{FD} = \frac{1}{4} \sum_{L=1}^4 \sum_{i,j,c} \|F_O^L(\phi_{r \rightarrow o}^L)(i,j,c) - F_R^L(i,j,c)\|^2, \quad (9)$$

which calculates the square difference between original and retargeted images in the feature space. $\{F_R^L\}$ is obtained by feeding retargeted images to the VGG-19. The larger FRR and lower FD score is, the better image quality is. As shown in the Table 1, our method achieves best performance in terms of user study and FRR score, and second best performance in FD score. CR achieves the best FD score because it

maintains original regions in the original images, which results in highly similar deep features as original ones'. Compared with SV which relies on both human labeled and automatic features, the quantitative results suggest the effectiveness of our method on semantic structure preserving.

4.3 Robustness of DeepIR

In order to investigate the influence of different aspect ratios ϵ and balancing parameters $\{\alpha^L\}_{L=1}^3$, we show the results of progressively increasing ϵ for retargeted image in Figure 7 and results of varying $\{\alpha^L\}$ in Figure 8. From Figure 7, we can observe that our method can effectively resize an image continuously while preserving important objects with various aspect ratios ϵ . From Figure 8, we can observe that the larger $\{\alpha^L\}$ tends to preserve more important contents, while too large $\{\alpha^L\}$ may lead to information loss, such as incomplete railing in the left of Figure 8(h). The smaller $\{\alpha^L\}$ also has its drawback at preserving important

content in image, such as the over-squeezed boy in Figure 8(a). Therefore, $\{\alpha^L\} = \{0.7, 0.8, 0.9\}$ is chose as our configuration to achieve a trade-off between high-level semantic content and low-level details.

4.4 Combining Retargeting Methods with DeepIR



Figure 9: One example of using CR, SCL, SC, column removal and our UrS for deep feature retargeting in DeepIR framework. (a) Original image. (b) DeepIR with CR. (c) DeepIR with SCL. (d) DeepIR with SC suffers structure distortion. (e) DeepIR with column removal suffers content discontinuity. (f) DeepIR with UrS produces visually preferred result.

In order to verify the effectiveness of UrS on feature retargeting and flexibility of our DeepIR, we combine other retargeting methods with DeeIR framework, including CR, SCL, SC and column removal that removes columns/rows based on minimum importance cost. In each layer, the feature maps are resized by these retargeting methods instead of UrS and propagated to the lowest layer to generate final retargeted images. The results are illustrated in Figure 9. However, we can observe that most of these methods cannot utilize the advantages of deep features and even not able to surpass the corresponding low-level retargeting method. Specifically, CR inevitably removes semantically important content in the original image (Figure 9(b)). In Figure 9(c), SCL simply stretches the object to the target aspect ratio as its behavior in Figure 5(f). Also, SC removes the least importance seams in the feature space and may destroy the semantic structures in the deep features, which results in structure non-homogeneity in the retargeted image (Figure 9(d)). Similar problem is also raised by column removal in Figure 9(e), where the content of retargeted image is discontinuous. To summarize, the proposed UrS method that resizes feature maps based on cumulative obscurity map sampling can best maintain the semantic structure of original image and eliminate the discontinuity in retargeted image so far. Although previous retargeting methods based on saliency map are not suitable for deep feature retargeting, we could find that DeepIR is flexible to incorporate other image retargeting methods into the framework.

4.5 Failure Cases

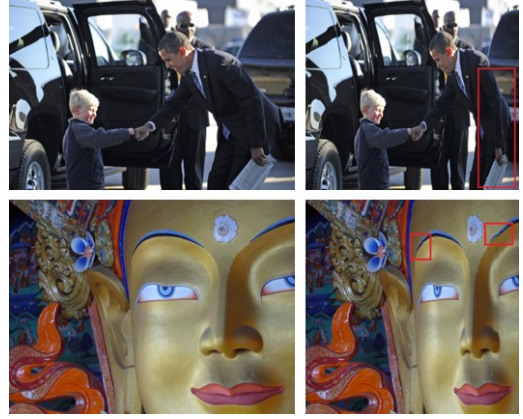


Figure 10: Failure cases. First row left: “obama” image. First row right: retargeted “obama” image, $\epsilon = 0.75$. Second row left: “buddha” image. Second row right: retargeted “buddha” image, $\epsilon = 0.75$.

The deep image retargeting technique relies on two previous works: the pre-trained VGG-19 network (Simonyan and Zisserman 2014) and PatchMatch (Barnes et al. 2009). Therefore, due to the limited capability of representing all objects and visual structures, we find that the DeepIR may over squeeze the important areas or over maintain the less important areas in some cases. We show an example in the first row of Figure 10, where the body of Obama is over squeezed and the car is over preserved. We expect there will be a network with stronger representation ability in the future. For PatchMatch in the deep feature space, there still exists mismatching in some cases as shown in the second row of Figure 10, where the eyebrow of Buddha suffers some discontinuous mismatching. One possible solution is to improve PatchMatch for scale-invariant matching in the future.

5 Conclusions

In this paper, we propose a new technique called *Deep Image Retargeting (DeepIR)*. Our method utilizes deep CNN for the content-aware image retargeting. We first use a pre-trained deep convolutional neural network to extract deep features of original image. Then we propose a uniform re-sampling (UrS) image retargeting method to resize feature maps at each feature layer. The UrS can effectively preserve semantic structure as well as maintain content smoothness in the retargeted image. Finally, we reconstruct the deep features of retargeted image in each layer through a coarse-to-fine NNF search and a step-by-step NNF fusion. The retargeted image is obtained by using spatial correspondence at the lowest layer. We have carried out sufficient experiments to validate the effectiveness of our proposed technique on RetargetMe dataset. In general, our method solves the problem of applying pre-trained deep neural network for content-aware image retargeting. Our model is flexible and one can easily combine other retargeting algorithms with proposed framework for future works.

References

- [Avidan and Shamir 2007] Avidan, S., and Shamir, A. 2007. Seam carving for content-aware image resizing. In *ACM Transactions on graphics (TOG)*, volume 26, 10. ACM.
- [Barnes et al. 2009] Barnes, C.; Shechtman, E.; Finkelstein, A.; and Goldman, D. B. 2009. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics-TOG* 28(3):24.
- [Cho et al. 2008] Cho, T. S.; Butman, M.; Avidan, S.; and Freeman, W. T. 2008. The patch transform and its applications to image editing. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8. IEEE.
- [Cho et al. 2017] Cho, D.; Park, J.; Oh, T.-H.; Tai, Y.-W.; and Kweon, I. S. 2017. Weakly-and self-supervised learning for content-aware deep image retargeting. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 4568–4577. IEEE.
- [Guo et al. 2009] Guo, Y.; Liu, F.; Shi, J.; Zhou, Z.-H.; and Gleicher, M. 2009. Image retargeting using mesh parametrization. *IEEE Transactions on Multimedia* 11(5):856–867.
- [He et al. 2016] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- [Itti, Koch, and Niebur 1998] Itti, L.; Koch, C.; and Niebur, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence* 20(11):1254–1259.
- [Karni, Freedman, and Gotsman 2009] Karni, Z.; Freedman, D.; and Gotsman, C. 2009. Energy-based image deformation. In *Computer Graphics Forum*, volume 28, 1257–1268. Wiley Online Library.
- [Krähenbühl et al. 2009] Krähenbühl, P.; Lang, M.; Hornung, A.; and Gross, M. 2009. A system for retargeting of streaming video. In *ACM Transactions on Graphics (TOG)*, volume 28, 126. ACM.
- [Ledig et al. 2016] Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; and Shi, W. 2016. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. *ArXiv e-prints*.
- [Liao et al. 2017] Liao, J.; Yao, Y.; Yuan, L.; Hua, G.; and Kang, S. B. 2017. Visual attribute transfer through deep image analogy. *ACM Trans. Graph.* 36(4):120:1–120:15.
- [Lin et al. 2014] Lin, S.-S.; Lin, C.-H.; Chang, S.-H.; and Lee, T.-Y. 2014. Object-coherence warping for stereoscopic image retargeting. *IEEE Transactions on Circuits and Systems for Video Technology* 24(5):759–768.
- [Lin, Zhou, and Chen 2018] Lin, J.; Zhou, T.; and Chen, Z. 2018. Multi-Scale Face Restoration with Sequential Gating Ensemble Network. *ArXiv e-prints*.
- [Long, Shelhamer, and Darrell 2015] Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440.
- [Mao, Shen, and Yang 2016] Mao, X.-J.; Shen, C.; and Yang, Y.-B. 2016. Image Restoration Using Very Deep Convolutional Encoder-Decoder Networks with Symmetric Skip Connections. *ArXiv e-prints*.
- [Pritch, Kav-Venaki, and Peleg 2009] Pritch, Y.; Kav-Venaki, E.; and Peleg, S. 2009. Shift-map image editing. In *Computer Vision, 2009 IEEE 12th International Conference on*, 151–158. IEEE.
- [Rubinstein et al. 2010] Rubinstein, M.; Gutierrez, D.; Sorkine, O.; and Shamir, A. 2010. A comparative study of image retargeting. In *ACM transactions on graphics (TOG)*, volume 29, 160. ACM.
- [Rubinstein, Shamir, and Avidan 2009] Rubinstein, M.; Shamir, A.; and Avidan, S. 2009. Multi-operator media retargeting. In *ACM Transactions on graphics (TOG)*, volume 28, 23. ACM.
- [Russakovsky et al. 2015] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.
- [Simakov et al. 2008] Simakov, D.; Caspi, Y.; Shechtman, E.; and Irani, M. 2008. Summarizing visual data using bidirectional similarity. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8. IEEE.
- [Simonyan and Zisserman 2014] Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- [Viola and Jones 2001] Viola, P., and Jones, M. 2001. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, I–I. IEEE.
- [Wang et al. 2008] Wang, Y.-S.; Tai, C.-L.; Sorkine, O.; and Lee, T.-Y. 2008. Optimized scale-and-stretch for image resizing. In *ACM Transactions on Graphics (TOG)*, volume 27, 118. ACM.
- [Wolf, Guttmann, and Cohen-Or 2007] Wolf, L.; Guttmann, M.; and Cohen-Or, D. 2007. Non-homogeneous content-driven video-retargeting. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 1–6. IEEE.
- [Wu et al. 2010] Wu, H.; Wang, Y.-S.; Feng, K.-C.; Wong, T.-T.; Lee, T.-Y.; and Heng, P.-A. 2010. Resizing by symmetry-summarization. *ACM Transactions on Graphics (TOG)* 29(6):159.
- [Zeiler and Fergus 2014] Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833. Springer.
- [Zhu et al. 1997] Zhu, C.; Byrd, R. H.; Lu, P.; and Nocedal, J. 1997. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)* 23(4):550–560.

[Zhu et al. 2016] Zhu, L.; Chen, Z.; Chen, X.; and Liao, N. 2016. Saliency & structure preserving multi-operator image retargeting. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 1706–1710. IEEE.