Quality assessment of retargeted images using hand-crafted and deep-learned features

Zhenqi Fu, Feng Shao, *Member*, *IEEE*, Qiuping Jiang, *Student Member*, *IEEE*, Randi Fu, Yo-Sung Ho, *Fellow*, *IEEE*

Abstract—Since the goal of image retargeting is to adapt source images to target displays with different sizes and aspect ratios, how to objectively evaluate the quality of retargeted images is particularly important to optimize the retargeting operations. In this paper, we proposed a new image retargeting quality assessment (IRQA) metric, which constructs the metric using both hand-crafted features and deep-learned features. To enhance the reliability and accuracy of the proposed method, 1) we use similarity transformation as local descriptor to extract hand-craft features, and measure structure distortion and content loss from the hand-craft features; 2) we use deep learning architecture to construct encoders and extract deep-learned features, and measure texture similarity and semantic similarity from the deep-learned features. We conduct experiments on two databases: RetargetMe and CUHK. Experimental results show that our method can achieve superior performance to the state-of-the-art metrics.

***Index Terms*—Image retargeting quality assessment; hand-crafted feature; deep-learned feature; structure distortion; content loss.**

I. INTRODUCTION

WITH the rapid development of mobile devices, such as smart phones and tablets, image retargeting has received much attention in recent years, which aims to adapt source images to target displays with different sizes and aspect ratios [1-3]. Even the traditional manual cropping (CR) and linear scaling (SCL) methods can achieve the goal, the limitation of these methods is that they do not take the diversity of image content in account, leading to poor visual quality. Recently, content-aware image retargeting has received considerable attention due to the advantages in balancing the structure distortion and content loss.

The existing content-aware image retargeting algorithms can be broadly classified into two categories: discrete and continuous approaches. Discrete methods resize an image by iteratively removing or inserting pixels in the less important regions. Seam-Carving (SC) [4] and Shift-Maps (SM) [5] are two representative discrete approaches. Even intuitive, the disadvantage of discrete approaches is that it may lead to noticeably jagged edges and artifacts of image objects. In contrast, Warping (WARP) [6], Streaming Video (SV) [7], Scale-and-Stretch (SNS) [8] and Multi-Operators (MULTIOP) [9] provide continuous solutions to keep important areas. Each continuous approach has its advantage and disadvantage in addressing the structure distortion and content loss for different retargeting operators [10]. Particularly, MULTIOP approach provides better result than other single operator by using the best combination of operators. Therefore, it is significant to develop an effective objective image retargeting quality assessment (IRQA) method to improve the retargeting techniques and select the best retargeting operator [11-12].

Different from traditional image quality assessment (IQA) that mainly evaluates fidelity, structure distortion and content loss are two crucial factors for quality degradation in IRQA. Moreover, due to the retargeting operations, the resolutions of the source and retargeted images are quite inconsistent, thus traditional IQA methods, such as Peak Signal to Noise Ratio (PSNR), structural similarity index (SSIM) [13], and feature similarity (FSIM) [14], cannot be directly applied to IRQA. Since the purpose of IRQA is to establish the mapping between the source and retargeted images and measure their distance as similarity index, most IRQA methods use hand-crafted features to calculate and measure the distance between the matched pixels. Although these features can represent the distortion levels of the retargeted images to a certain extent, the reliability of hand-crafted features is largely dependent on the accuracy of correspondence matching. In addition, hand-crafted features are usually low-level, lacking semantics and discriminative capacity.

Inspired by recently popular deep convolution neural networks (CNNs) that extract the high-level information by developing deep architectures in image based tasks [15-17], we use CNN model proposed in [16] as feature extractor, and propose a new framework that uses hand-crafted and deep-learned features for IRQA. As discussed, different retargeting operators may cause different degrees of structure distortion and content loss, which will affect the semantic for understanding, but the hand-crafted features cannot capture such information. Therefore, this paper takes a different view

This work was supported in part by the Natural Science Foundation of China (grant 61622109), and the Zhejiang Natural Science Foundation of China (grant R18F010008). It was also sponsored by K.C.Wong Magna Fund in Ningbo University.

Zhenqi Fu, Feng Shao, Qiuping Jiang and Randi Fu are with the Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China (e-mail: shaofeng@ nbu.edu.cn).

Yo-Sung Ho is with the School of Information and Communications, Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Korea (e-mail: hoyo@gist.ac.kr).

of this problem and supply the hand-crafted features with deep-learned representations. The overall target of this work is to provide a solution for IRQA by combining both hand-crafted and deep-learned features. The main contributions of our work are two-fold: 1) We use similarity transformation as descriptor to establish the relationship between the source and retargeted images and devise a pyramidal model to obtain multi-scale structure distortion, and measure the area change in each grid of retargeted images to capture the image content loss; 2) We construct CNN architecture as encoder to extract deep-learned features, including texture feature and semantics feature from bottom layer and top layer of CNN respectively, and measure texture similarity and semantic similarity to improve the prediction accuracy of IRQA.

The rest of this paper is organized as follows. In Section II, we present related works and motivations for this work. We detail the proposed method in Section III, and finally present experimental results in Section IV and conclusions in Section V.

II. Background

A. Related Works

The early IRQA methods mostly utilize intensity or color distance to measure the quality of retargeted images, but the evaluation results are commonly unsatisfactory. Edge Histogram (EH) [18] and Color Layout (CL) [19] are two IRQA methods in MPEG-7 standard, which use histogram distance and color distribution as representations respectively. Bi-Directional Similarity (BDS) [20] and Bi-Directional Warping (BDW) [9] calculate bidirectional mapping distances between the patches in the source and retargeted images as measurements. However, these two methods ignore the importance of image contents and have low correlation with subjective ranking. SIFT-flow [21] uses SIFT descriptors to establish the matching relationship between the source and retargeted images, and the cost function based on the displacements of adjacent pixels is applied to measure the dissimilarity of two images. Earth mover's distance (EMD) [22] solves a transportation issue instead of solving a matching issue, and the minimal cost in transforming the source signature to match a target signature is measured as similarity index. Relatively, SIFT-flow and EMD metrics can capture the structural properties of an image, and have high consistency with subjective rankings compared with EH, CL, BDS and BDW. However, SIFT-flow and EMD still ignore the influence of critical content loss in IRQA.

In recent years, IRQA has received extensive attentions with the development of image retargeting techniques. Fang *et al*. [23] devised a multi-scale SSIM (IR-SSIM) metric to measure how the structural information in source image is preserved in the retargeted images. Hsu *et al*. [24] defined a weighted combination of geometric distortion and content loss as the similarity of two images. In this method, geometric distortion is measured by the local variance of SIFT-flow and content loss is calculated from the saliency area loss. Zhang *et al*. [12] elaborated the geometric change by a backward registering in Markov random field and measured the aspect ratio similarity (ARS) in local blocks to reveal the geometric change, but the ARS metric does not take global distortion of the retargeted images into consideration. Liang *et al*. [25] incorporated the factors of similarity, aesthetics and symmetry to predict the quality of retargeted images. Karimi *et al*. [11] define three groups of features including shape features, area features and aspect ratio to reveal the geometric distortion and content loss based on dense correspondence between the scaled and retargeted images. However, the above methods relied on accurate correspondence matching and effective feature descriptors. Recently, Jiang *et al*. [26] evaluated the quality of retargeted images by learning two over-complete dictionaries. This method does not need correspondence matching between the source and retargeted images and can capture high-level information of retargeted images by the dictionaries. Other relevant works can be found in [27-29].

Deep learning techniques have been achieved astonishing advances in recent years, and there have been a number of attempts to develop deep CNN architectures for image recognition [30-31], image repairing [32], image segmentation [33], image denoising [34] and super resolution [35-36]. Cho *et al*. [37] employed deep CNN for image retargeting, which can be classified into two steps: learning an attention map via end-to-end training and generating a content-aware shift map for image retargeting. Thus, content and structure loss of the retargeted images can be computed from image-level annotations. Zhang *et al*. [38] used a CNN model to learn the local structures of stereoscopic image for no-reference quality assessment. In the method, using image patches as inputs, high-level representations are summarized as final quality scores, which are learned with multiple layers architecture of CNN. Kim *et al*. [39] adopted a full reference IQA metric to obtain the quality score of each patch as label information, and trained a local patch-based CNN model for no-reference IQA. Kao *et al*. [17] employed a multi-task deep CNN model to automatically assess the quality of aesthetic images, and meanwhile capture the important semantic information of each aesthetic image. Although deep learning techniques have been adopted in image retargeting and IQA applications, how to utilize the technique in IRQA is still challenging due to the limitations in lacking of enough samples for training.

B. Motivations

Traditional hand-crafted features require local descriptors, such as SIFT [40], Histogram of Gradient (HOG) [41], Histogram of Optical Flow (HOF) [42], to extract key points and establish their mapping relationship. However, the limitation of these local descriptors in image retargeting is that they cannot capture the intrinsic formation mechanism of structure distortion and content loss which are caused by different retargeting operators. Due to the high performance of warping operator for image retargeting, we are motivated to use warping operator (described by similarity transformation) as local descriptor to simulate different retargeting operators and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2808322, IEEE Access

extract hand-crafted features. Thus, the deviations in the similarity transformations for different retargeting operations can reflect the deformations imposed on the local grids. As shown in Fig. 1, using similarity transformation as local descriptor for different discrete methods (CR and SC) and continuous methods (SCL and WARP), the errors between the retargeted image and its reconstrued one are comparatively small, indicating high similarity.

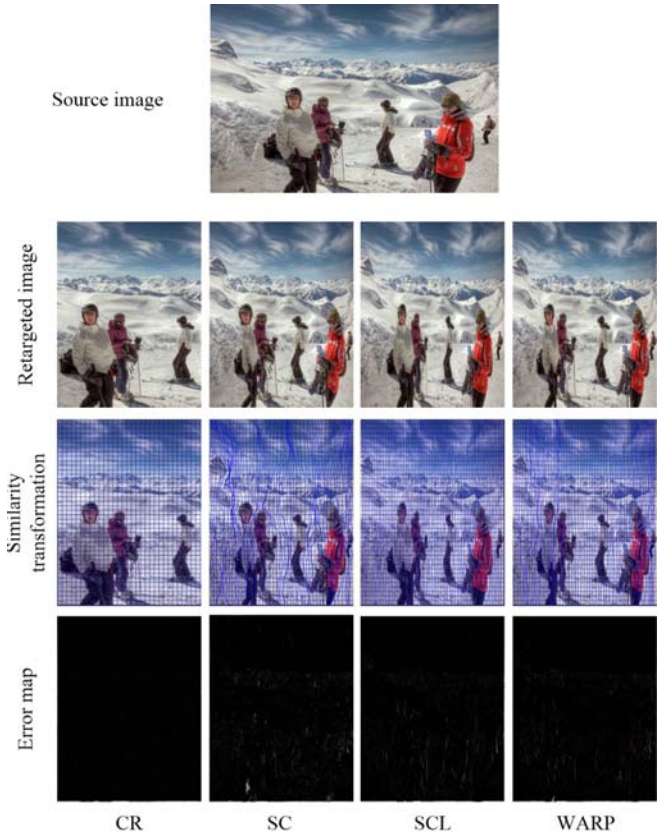


Fig. 1. Similarity transformations for four typical retargeting operators.

The warping descriptor aims to find an optimal transformation based on limited control points. Let $P=\{p_i\}$ be a handle with m distinct control points, its deformed position be $P'=\{p_i'\}$, the distortion energy is defined as [1]:

$$\varepsilon(P',P)=\min_{s\in S}\sum_{i=1}^{m}|\, s(\mathbf{p}_i)-\mathbf{p}_i'\,|^2 \qquad (1)$$

where S is the set of similarity transformations. Obviously, similarity transformation serves as an intermediary between the source and retargeted images to control the deformation of each grid. For example, a large bias of the similarity transforms from the benchmark ones usually leads to serious geometric distortion and content loss.

From another aspect, CNN has been successfully used to generate image feature maps such as semantics map [33] and structure map [16]. At the lower layers of the baseline CNN, features such as gradients and edges are learned, while at the higher layers, the learned features contain meaningful information that can describe the semantic characteristics [15]. The most important information in the content-aware image retargeting is semantic information, IRQA is a high-level task that should understand images in global perspective. Based on these observations, to utilize semantic information in IRQA, we use CNN architecture to construct different encoders for the source and retargeted images, and extract low-level and high-level features (including texture and semantic information) to generate the deep-learned feature representation.

III. Proposed Method

In this paper, we propose a IRQA using hand-crafted and deep-learned features, as shown in Fig.2. The key of IRQA motivated in this method is to dig the low-level and high-level features as feature representation. Thus, the overall framework of the proposed method is composed of two channels: the first channel constructs similarity transformation as local descriptor to extract hand-crafted features, and calculates structure distortion and content loss for measurement. The second channel uses CNN architecture to construct an encoder, and extracts texture feature and semantics feature from the bottom layer and top layer of the CNN. Different with state-of-the-art IRQA methods that rely on low-level features and accurate correspondences, our method simultaneously uses hand-crafted features and deep-learned features to estimate the perceptual quality degradation. In the following sections, we will elaborate on each channel of our method.

A. Hand-crafted feature representation

1). SIFT-flow estimation

As discussed, we use similarity transformation as local descriptor to extract hand-crafted features. The primary task is to establish pixel-to-pixel correspondences between the source and the retargeted images, and convert into a field of transformation. In this work, we use SIFT-flow algorithm [21] to extract dense SIFT descriptor for each pixel in both source and retargeted images to build up the correspondence.

2). Multi-scale structure distortion measurement

As the purpose of this step is to establish the similarity transformation, pixel-wise correspondence cannot obtain such mapping relationship. Refer to warping operator, a set of regular grids are first extracted in the source image. Then, based on the pixel-wise correspondences for four vertexes in a grid, a similarity transformation for the grid can be established. In our case, transformations are a set of 2D affine matrices to represent scaling, rotation and translation, which can be obtained from four vertexes of a grid by M-estimator algorithm [43]. Let $\{\mathbf{v}_k, k=1,\cdots,N\}$ be a set of grids in the source image (N is the number of grids in the source image), and $\{\tilde{\mathbf{v}}_k, k=1,\cdots,N\}$ be the corresponding grids in the retargeted image established by correspondence matching, similarity transformation can be formulated as:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2808322, IEEE Access

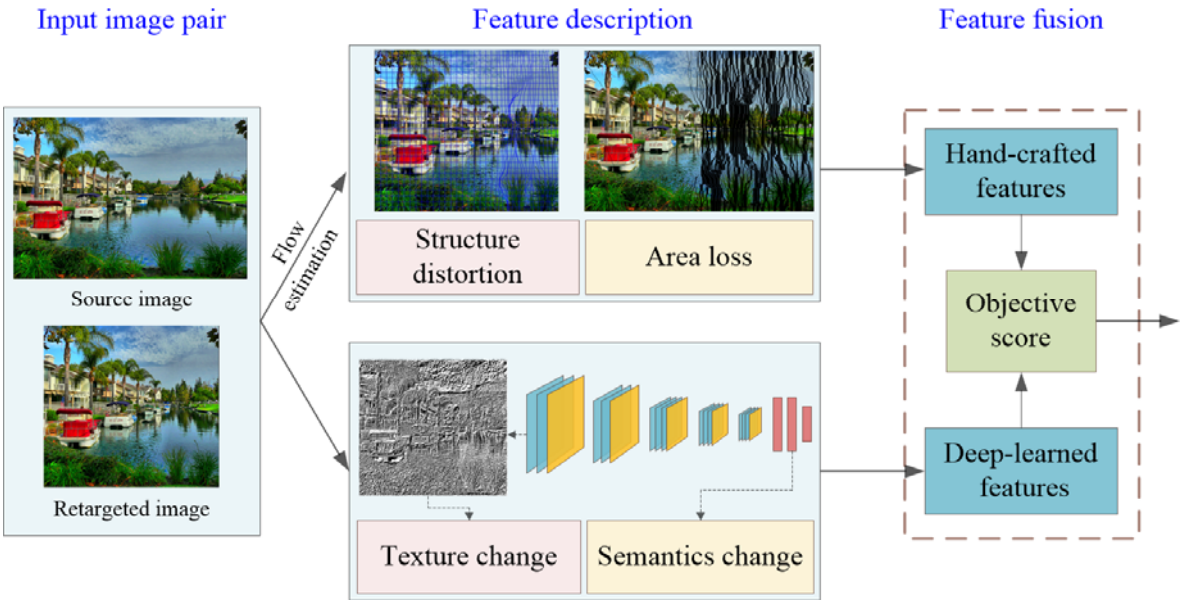


Fig. 2. Overall architecture of the proposed method for image retargeting quality assessment.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a & b & m \\ c & d & n \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \mathbf{P}_k \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2}$$

where $\mathbf{P}_k$ is the similarity transformation matrix. Considering that scaling and rotation will affect structure distortion while translation itself will not cause structure distortion, we define a benchmark transformation matrix $\mathbf{P}_B$ as follows:

$$\mathbf{P}_B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{3}$$

Since only two scaling parameters (a and d) and two rotation parameters (b and c) will have a negative effect on structure distortion, we calculate the distance between a similarity transformation matrix and the benchmark transformation matrix to measure the degree of structure distortion. The distance defined in this work is composed of two components: the absolute distance (AD) and aspect ratio (AR) change, defined as follows:

$$\eta = \underbrace{\left\| \mathbf{Z}_k - \mathbf{P}_B \right\|_2^2}_{\text{AD}} + \underbrace{(a-d)^2 + (b-c)^2}_{\text{AR}} \tag{4}$$

where $\mathbf{Z}_k = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. In the equation, the translation parameters (m and n) are not included, because translation itself will not produce any structure distortion.

Since the motivation of warping-based retargeting is to preserve important shapes and avoid over-deformation for the high-significance objects, the structure distortion defined in this work is calculated with the significant values as weighting:

$$f_1 = \sum_{\mathbf{v}_k} S_{\mathbf{v}_k} \cdot \mathrm{e}^{-\eta_{\mathbf{v}_k}} \Big/ \sum_{\mathbf{v}_k} S_{\mathbf{v}_k} \tag{5}$$

where $S_{\mathbf{v}_k}$ is the average saliency value of a grid in the source image. In this paper, we use Hierarchical saliency detection algorithm (HS) [44] to grade the importance of different regions. Thus, large f_1 value means small structure distortion for the retargeted image. In addition, to indicate the influence of grid's size, we provide a multi-scale solution that uses three different sizes (32×32, 16×16 and 8×8) to measure the structure distortion, denoted as f_1, f_2 and f_3 respectively.

3). Content loss measurement

Besides the structure distortion, content loss is another important factor in image retargeting, which may destroy the completeness of image content and further affect human's understanding of the scene. The above structure distortion measurement can reflect a certain content loss for most retargeting operators, however, for the cropping operation, if a grid in the source image cannot find the corresponding grid in the retargeted image (the information is discarded in the retargeted image), the above structure distortion measurement cannot capture such information. Thus, we make an assumption that if a grid has the same area in both source and retargeted images, the information in this grid is well preserved. Towards this end, we design a simple yet effective metric that calculates the change of grid's area to obtain the content loss:

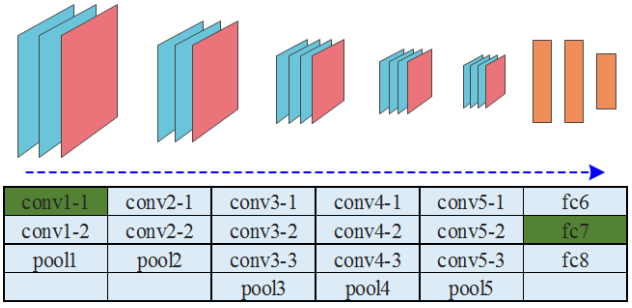
$$f_4 = \sum_{\mathbf{v}_k} S_{\mathbf{v}_k} \cdot \frac{A(\mathbf{v}_k)}{A(\tilde{\mathbf{v}}_k)} \Big/ \sum_{\mathbf{v}_k} S_{\mathbf{v}_k} \tag{6}$$

where $A(\mathbf{v}_k)$ and $A(\tilde{\mathbf{v}}_k)$ are the areas of the grids in the source and retargeted images respectively. Thus, large value of $\frac{A(\mathbf{v}_k)}{A(\tilde{\mathbf{v}}_k)}$ means more salient information is preserved in the retargeted image. Particularly, $\frac{A(\mathbf{v}_k)}{A(\tilde{\mathbf{v}}_k)} = 0$ reflects the grid is discarded in the retargeted image.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2808322, IEEE Access

B. Deep-learned feature representation

Although the above multi-scale structure distortion and content loss measurement can effectively capture the low-level deformations, but it ignores the texture change in each grid, and the content loss defined above only measures the local information loss. Besides, these two features are highly dependent on the accuracy of correspondence matching and the reliability of image saliency information. Thus, in order to compensate the limitations of the hand-crafted feature representation, we explore high-level information of retargeted images. In this paper, we adopt pre-trained VGG16 network [16] to construct encoders for the source and retargeted images, and then measure texture similarity and semantic similarity from the deep-learned features. The structure of the VGG16 network is shown in Fig. 3.

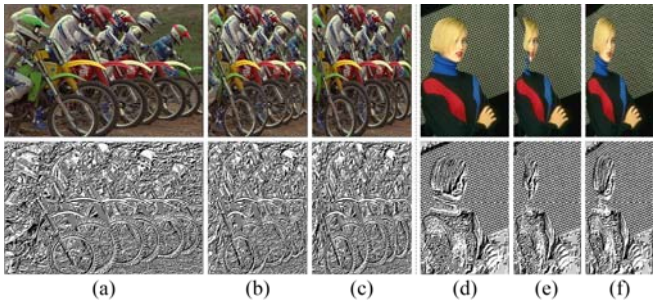


conv1-1	conv2-1	conv3-1	conv4-1	conv5-1	fc6
conv1-2	conv2-2	conv3-2	conv4-2	conv5-2	fc7
pool1	pool2	conv3-3	conv4-3	conv5-3	fc8
		pool3	pool4	pool5	

Fig. 3. The structure of the VGG16 Network

1). Texture similarity measurement

As discussed in [15], activations from bottom layers of CNN architecture provide low-level structural information. Therefore, we utilize activations from conv1-1 of VGG16 for the computation of texture similarity to overcome the limitations of the multi-scale structure distortion. There are 64 feature maps in conv1-1, and we only choose the first feature map for IRQA task. The feature map extracted from conv1-1 of VGG16 is shown in Fig. 4.



(a) (b) (c) (d) (e) (f)

Fig. 4. Image texture information from conv1-1: (a) and (d) Source images; (b) and (e) Retargeted images by SC; (d) and (f) Retargeted images by WARP.

Then, we detect texture change by calculating the homogeneity value of Gray Level Co-occurrence Matrix (GLCM) [45]. GLCM is one of the stable feature extraction methods. In this paper, two typical texture features (energy and entropy) are extracted from each GLCM. The energy is used to describe the gray-level distribution and the entropy can capture the size of texture. The higher values of the energy and entropy, the more texture information would be contained in the image. The functions of energy and entropy are defined as:

$$E_\theta = \sum_{i=1}^{K}\sum_{j=1}^{K} p_{d,\theta}^2(i,j) \tag{7}$$

$$H_\theta = -\sum_{i=1}^{K}\sum_{j=1}^{K} p_{d,\theta}(i,j)\log_2 p_{d,\theta}(i,j) \tag{8}$$

where $p_{d,\theta}(i,j)$ is the GLCM, d is the distance between two pixels (d=1 in this paper), θ is the orientation, and K is the quantized gray level (K=8 in this paper).

After computing the GLCM of four different orientations (0°, 45°, 90° and 135°), the energy similarity and entropy similarity between source and retargeted images are defined as:

$$\gamma_e = \exp\left(-\sum_\theta \left(E_\theta^R - E_\theta^S\right)^2\right) \tag{9}$$

$$\gamma_h = \exp\left(-\sum_\theta \left(H_\theta^R - H_\theta^S\right)^2\right) \tag{10}$$

where E_θ^S and E_θ^R are the energies for the source and retargeted images, respectively, and H_θ^S and H_θ^R are the entropies for the source and retargeted images, respectively.

By integrating the energy similarity and entropies similarity using multiplication combination, the texture similarity between the source and retargeted images is defined as:

$$f_5 = \gamma_e \cdot \gamma_h \tag{11}$$

2). Semantics similarity measurement

Different from the content loss that calculates the changes of local areas, in this section, we design a global measurement to obtain the high-level semantics information. We capture the high-level semantics information from the top layer of VGG16 to detect the ambiguous and inconsistent understanding between the source and retargeted images. In our experiments, the last classification layer of VGG16 is removed, and the output from the second full connection layer is recognized as the semantics feature. The semantics similarity is defined as:

$$f_6 = \frac{1}{M}\sum_{i=1}^{M} p_i \tag{12}$$

where M is the number of fc7 output (shown in in Fig. 3), $\mathbf{f}^S = [f_1^S, \cdots, f_i^S, \cdots, f_M^S]$ and $\mathbf{f}^R = [f_1^R, \cdots, f_i^R, \cdots, f_M^R]$ are the outputs of fc7 layer for the source and retargeted images respectively. In particular, we normalize f_6 through a square threshold to reduce the influence of deformation on object detection:

$$p_i = \begin{cases} 1, & \text{if } \left(f_i^R - f_i^S\right)^2 < 1.5 \\ 0, & \text{else} \end{cases} \tag{13}$$

Here, we set threshold as 1.5 experimentally with the best performance. As an example, we illustrate the semantics similarity score for different retargeted images in Fig. 5. We can find that, CR has the worst semantics similarity score, while SCL has the best result. The reason is intuitive that CR will cause serious content loss, while the objects can be well

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2808322, IEEE Access

preserved in the image retargeted by SCL, although stretching or squeezing may distort the entire image.

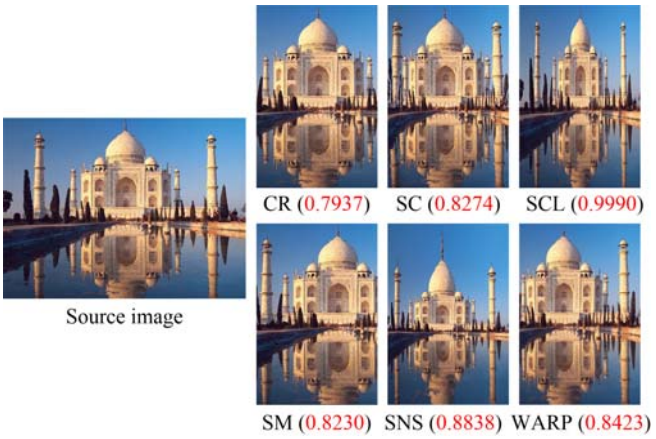


Fig. 5. Example of semantics similarity scores for different retargeting operators.

C. Quality evaluation

With the estimated scores f_1, f_2, f_3, f_4, f_5 and f_6, we train a regression model that maps the 6-dimensional feature vectors to the associated quality scores. In our implementation, a support vector regression (SVR) is adopted to train the function. In addition, we utilize the widely used radial basis function (RBF) kernel to nonlinearly combine the quality scores.

IV. Experimental Results and Analyses

A. Databases

We In this paper, two widely-used databases including RetargetMe [46] and CUHK [47] are used for performance evaluation in our experiment. The brief introductions about these databases are as follows:

RetargetMe database consists of 37 source images which are classified into 6 major attributes including lines/edges, face/people, texture, foreground objects, geometric structures and symmetry. Note that one source image may have one or more attributes. Each source image has been retargeted by eight typical retargeting operators, including SC [4], WARP [6], SV [7], SM [5], SNS [8], MULTIOP [9], CR and SCL. Thus, total 296 retargeted images are included in the database. The ground-truth subjective score of each image in the database is recorded as the number of times that the image is preferred over other retargeted images.

The CUHK database contains 171 retargeted images obtained from 57 source images. For each source image, three different retargeting operators are applied which are randomly selected from ten representative operators, including the eight operators used in RetargetMe database and other two operators namely optimized seam carving and scale (SCSC) [10] and energy-based deformation (ENEN) [48]. Similar to traditional IQA study, the mean opinion score (MOS) is generated for each retargeted image as the subjective quality score via five-grade ranking.

In the CUHK dataset, we randomly divide all images into two groups, 20% are chosen for testing and 80% are chosen for training. We repeat such train-test procedure 1000 times and obtain the average performance as the final quality score. To compare objective and subjective scores, we first fit the objective scores using the following five-parameter mapping function:

$$f(x)=\beta_1\left(\frac{1}{2}-\frac{1}{1+e^{\beta_2(x-\beta_3)}}\right)+\beta_4 x+\beta_5 \tag{14}$$

where β_1, β_2, β_3, β_4 and β_5 are parameters to be fitted. Thus, we apply Pearson linear correlation coefficient (PLCC), Spearman rank order correlation coefficient (SROCC), root mean square error (RMSE) and outlier ratio (OR) to evaluate the performance. Larger SROCC and PLCC indicate better performance of the objective quality measure, while a smaller RMSE and OR indicate higher correlation between objective prediction values and subjective scores

However, subjective scores in RetargetMe database only indicates the relative ranking scores against other images, which are not suitable for training a model. Therefore, we use the SVR model trained on whole CUHK database to predict the quality scores of all images in RetargetMe database. Refer to [46], we use Kendall Rank Correlation Coefficient (KRCC) to evaluate the model performance. Here, KRCC = 1 denotes the best performance while KRCC=-1 means the worst result. The KRCC is defined as:

$$KRCC=\frac{n_c-n_d}{0.5n(n-1)} \tag{15}$$

where n is the length of the ranking and $n = 8$ in the RetargetMe database, n_c is the number of concordant pairs, n_d is the number of dis-concordant pairs.

B. Performance comparisons with other methods

To objectively evaluate the performance of our method, we compare it with state-of-the-art IRQA methods, including BDS [20], EH [18], SIFT-flow [21], EMD [22], and IR-SSIM [23], Liang's method [25], PGDIL [24], and ARS [12]. Table I gives comparisons of mean and standard deviation values of KRCC values as well as p-value on the RetargetMe database. Table II gives the comparison results of PLCC, SROCC, KRCC and RMSE values on the CUHK database. From the tables, we can make the following observations: 1) In the RetargetMe database, our method always performs well on all attributes except Faces People, but it is especially prominent on other attributes. The reason may be that our method is more effective to detect shape and structure distortions by involving deep-learned texture and semantic features, while other methods do not have such consideration. The overall performance of our method is also obviously better than other methods. 2) In the CUHK database, our method outperforms than other methods as expected, because our method takes low-level hand-crafted features and high-level deep-learned features into account. In Fig. 6, we compare the individual KRCC values of SIFT-flow, ARS and our method for the 37 image sets from MIT database. Compared with other two

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2808322, IEEE Access

Table I: Performance of different methods on RetargetMe database.

Metric	Attribute						Total		
	Line Edge	Faces People	Foreground Objects	Texture	Geometric Structure	Symmetry	Mean KRCC	Std KRCC	p-val
BDS[20]	0.040	0.190	0.067	0.060	-0.004	-0.012	0.083	0.268	0.107
EH[18]	0.043	-0.076	-0.079	-0.060	0.103	0.298	0.004	0.334	0.641
SIFT-flow[21]	0.097	0.252	0.218	0.161	0.085	0.071	0.145	0.262	0.031
EMD[22]	0.220	0.262	0.226	0.107	0.237	0.500	0.251	0.272	1e-5
IR-SSIM[23]	0.309	0.452	0.377	0.321	0.313	0.333	0.363	0.271	1e-3
Liang[25]	0.351	0.271	0.381	0.304	0.415	0.548	0.399	-	-
PGDIL[24]	0.431	0.390	0.389	0.286	0.438	0.523	0.415	0.296	6e-10
ARS[12]	0.463	**0.519**	0.444	0.330	0.505	0.464	0.452	0.283	1e-11
Our	**0.497**	0.472	**0.468**	**0.393**	**0.545**	**0.631**	**0.494**	**0.243**	**1.6e-14**

Table II: Performance of different methods on CUHK database.

Metric	PLCC	SROCC	RMSE	OR
BDS[20]	0.2896	0.2887	12.922	0.2164
EH[18]	0.3422	0.3288	12.686	0.2047
SIFT-flow[21]	0.3141	0.2899	12.817	0.1462
EMD[22]	0.2760	0.2904	12.977	0.1696
PGDIL[24]	0.5403	0.5409	11.361	0.1520
ARS[12]	0.6835	0.6693	9.855	0.0702
Our	**0.7170**	**0.6847**	**9.135**	**0.0215**

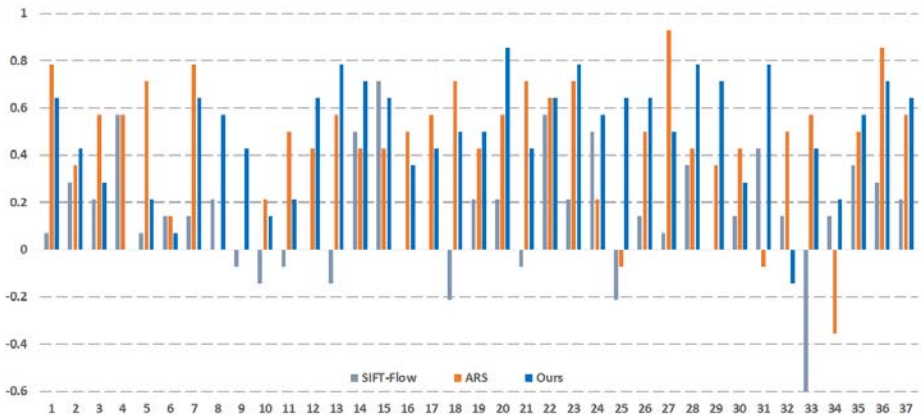


Fig. 6. The individual KRCC results of SIFT-Flow [21], ARS [12] and our method the 37 image sets on RetargetMe database.

methods, our method shows comparatively reliable and stable quality prediction without obvious fluctuations. Overall, our method can achieve good performance on predicting the quality of retargeted images.

C. Performance of hand-craft features and deep-learned features

To further analyze the effectiveness and influence of the hand-craft and deep-learned features in evaluating the perceptual quality, we design the following schemes for comparisons: only using the hand-craft features, the deep-learned features and the integrated features (due to the limitation of space in the table, we use 'HC', 'DL' and 'All' to represent three types of features, respectively) with three typical quality combination methods. The used average and multiply combination methods do not need training process. Table III shows the comparison results of these measurements. We can make the following observations from the table: 1) Compared with results obtained using the hand-craft features and deep-learned features, the hand-craft features are more important to evaluate the quality whose performance is better than that of the deep-learned features; 2) The best performance

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2808322, IEEE Access

Table III. Performance comparison of different quality pooling methods on RetargetMe and CUHK databases.

Dataset	Criteria	Average			Multiply			SVR		
		HC	DL	All	HC	DL	All	HC	DL	All
CUHK	PLCC	0.6961	0.3595	0.6933	0.6760	0.3572	0.6888	0.7034	0.4823	**0.7170**
	SRCC	0.6480	0.3365	0.6527	0.6428	0.3317	0.6502	0.6640	0.4237	**0.6847**
	RMSE	9.6932	12.737	9.7312	9.9497	12.619	9.7916	9.6387	11.6428	**9.1352**
RetargetMe	KRCC	0.4093	-0.029	0.4151	0.3958	-0.031	0.4054	0.4537	0.1718	**0.4942**

Table IV. Performance of each quality component on RetargetMe and CUHK databases.

Dataset	Criteria	f_1	f_2	f_3	f_4	f_5	f_6	Our
CUHK	PLCC	0.6854	0.6976	0.6817	0.6187	0.2740	0.2838	**0.7170**
	SRCC	0.6515	0.6605	0.6419	0.5842	0.4052	0.2699	**0.6847**
	RMSE	9.9540	9.8340	10.0586	10.6463	14.0774	13.2626	**9.1352**
RetargetMe	KRCC	0.4306	0.4344	0.3938	0.2799	0.2124	-0.0676	**0.4942**

is obtained by fusing the two types of features; 3) Since the average and multiply combinations do not need a training process, their performance are worse than the non-linear model trained by SVR.

D. Performance of each quality component

We further analyze the individual contribution of each quality component (f_1, f_2, f_3, f_4, f_5 or f_6) in our method: structure distortions (f_1, f_2 and f_3), area loss (f_4), texture similarity (f_5), and semantics similarity (f_6). Table IV shows the comparison results of these measurements. We can make the following observations from the table: 1) Each measurement has its respective role in characterizing the structure distortion and content loss, and independently applying the quality component cannot obtain the best results on two databases; 2) Among these measurements, structure distortions have the great influence on the overall performance. The reason is that it can effectively capture the geometric distortion through the similarity transformation; 3) Although the performance of deep-learned features (f_5 and f_6) are not good enough, they can capture high-level information to compensate the limitations of the hand-crafted features. So that, the integrated results are obviously better than those of independent ones.

To better explain the effect of each quality component, we test a set of retargeted images generated by different retargeting operators shown in Fig. 7 and reports the corresponding quality values in Table V. From Table V, we can find the influence of each component intuitively: 1) Discrete methods (CR, SC and SM) directly remove some background regions, leading to semantics information change and lower semantics similarity (f_6); 2) For the CR operator, since the geometric distortion is not significant, the measured structure distortions and texture similarity are better than other r quality values; 3) Due to little texture changes in these retargeted images, the values of f_5 are larger than other quality values; 4) Since the car is the most significant information in the scene, the image retargeted by SNS will seriously reduce the size of the car, leading to lower content loss (f_4).

E. Discussions

In this paper, we attempt to combine hand-craft features and deep-learned features to promote the performance of IRQA. Although the experiment results have demonstrated the effectiveness of our method in comparison with state-of-the-art IRQA methods, it still has some limitations: 1) The hand-craft features in our method are highly dependent on accurate correspondences between the original and retargeted images to construct the similarity matching relationship. However, SIFT-flow has its limitation in capturing structure features, especially in smooth areas; 2) Due to the limitations of the existing image retargeting databases, we use the pre-trained encoder to extract the deep-learned features. The issue should be solved by establishing the deep learning network between the source input and the retargeted output. Thus, structure loss and content loss can be derived directly from the network.

V. CONCLUSIONS

In this paper, we propose a new image retargeting quality assessment (IRQA) method using hand-crafted and deep-learned features. Using similarity transformation as descriptor to extract hand-craft features, we measure structure distortion and content loss from the hand-craft features. Using deep learning architecture to construct an encoder and extract deep-learned features, we measure texture similarity and semantic similarity from the deep-learned features. Compared with other metrics, our method achieved the best performance on both RetargetMe and CUHK databases. For future work, we will focus on designing more effective high-level features for IRQA.

REFERENCES

[1] G. Zhang, M. Cheng, S. Hu, RR. Martin, "A shape-preserving approach to image resizing," *Computer Graphics Forum*, vol. 28, no. 7 pp. 1897-1906, 2010.

[2] B. Li, L. Y. Duan, C. W. Lin, T. Huang and W. Gao, "Depth-preserving warping for stereo image retargeting," *IEEE Trans. Image Processing*, vol. 24, no. 9, pp. 2811-2826, Sep. 2015.

Fig.7. Original image and its corresponding retargeted images generated by different retargeting operators.

Table V. The predicted quality values of different metrics for retargeted images in Fig. 7

Operator	f_1	f_2	f_3	f_4	f_5	f_6	All	Subjective score
CR	0.9325	0.9216	0.9069	0.8169	0.9997	0.8818	68.5055	46
SV	0.8768	0.8900	0.8930	0.6792	0.9857	0.9692	64.1089	46
MULTIOP	0.8646	0.8725	0.8785	0.6844	0.9949	0.9890	63.1854	29
SC	0.8016	0.8094	0.8247	0.6766	0.9722	0.9070	57.5777	8
SCL	0.8623	0.8743	0.8770	0.6794	0.9966	0.9998	63.0901	39
SM	0.8866	0.8747	0.8416	0.7851	0.9833	0.8560	64.1302	51
SNS	0.8195	0.8356	0.8379	0.6096	0.9896	0.9124	58.6905	12
WARP	0.8573	0.8714	0.8718	0.6846	0.9600	0.9807	62.5496	21

[3] F. Shao, W. Lin, W. Lin, Q. Jiang, and G. Jiang, "QoE-guided warping for stereoscopic image retargeting," *IEEE Trans. Image Processing*, vol. 26, no. 10, pp. 4790-4805, Oct. 2017.

[4] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Trans. Graphics*, vol. 26, no. 3, article. no. 10, 2007.

[5] Y. Pritch, E. Kav-Venaki, and S. Peleg, "Shift-map image editing," in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, pp. 151-158, 2009.

[6] L. Wolf, M. Guttmann and D. Cohen-Or, "Non-homogeneous content-driven video-retargeting," *in Proc. of IEEE International Conference on Computer Vision (ICCV)*, pp. 1-6, 2007.

[7] P. Krähenbühl, M. Lang, A. Hornung, and M. H. Gross, "A system for retargeting of streaming video," *ACM Trans. Graphics*, vol. 28, no. 5, article no. 126, 2009.

[8] Y. Wang, C. Tai, O. Sorkine, and T. Lee, "Optimized scale-and-stretch for image resizing," *ACM Trans. Graphics*, vol. 27, no. 5, article no.118, 2008.

[9] M. Rubinstein, A. Shamir, S. Avidan, "Multi-operator media retargeting," *ACM Trans. Graphics*., vol. 28, no. 3, article no. 23, 2009.

[10] W. Dong, N. Zhou, J. C. Paul, and X. Zhang, "Optimized image resizing using seam carving and scaling," *ACM Trans. Graphics*, vol. 28, no. 5, article no. 125, 2009.

[11] M. Karimi, S. Samavi, N. Karimi, S. M. R. Soroushmehr, W. Lin, K. Najarian, "Quality assessment of retargeted images by salient region deformity analysis," *Journal of Visual Communication and Image Representation*, vol. 43, pp. 108-118, Feb, 2017.

[12] Y. Zhang, Y. Fang, W. Lin, X. Zhang, L. Li, "Backward registration based aspect ratio similarity for image retargeting quality assessment," *IEEE Trans. Image Processing*, vol. 25, no. 9, pp. 4286-4297, Sep, 2016.

[13] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600-612, Apr, 2004.

[14] L. Zhang, L. Zhang, X. Mou and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Processing*, vol. 20, no. 8, pp. 2378-2386, Aug, 2011.

[15] A. Krizhevsky, I. Sutskever, G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *in Proc. of International Conference on Neural Information Processing Systems*, vol. 2, pp. 1097-1105, 2012.

[16] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[17] Y. Kao, R. He, K. Huang, "Deep aesthetic quality assessment with semantic information," *IEEE Trans. Image Processing*, vol. 26, no. 3, pp. 1482-1495, Mar, 2017.

[18] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, A. Yamada, "Color and texture descriptors," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 11, no.6, pp. 703-715, Jun, 2001.

[19] E. Kasutani and A. Yamada, "The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval," *in Proc. of IEEE International Conference on Image Processing (ICIP)*, pp. 674-677, 2001.

[20] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani, "Summarizing visual data using bidirectional similarity," *in Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-8, 2008

[21] C. Liu, J. Yuen, and A. Torralba, "SIFT Flow: Dense correspondence across scenes and its applications," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 978-994, May, 2011.

[22] O. Pele and M. Werman, "Fast and robust earth mover's distances," in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, pp. 460-467, 2009.

[23] Y. Fang, K. Zeng, Z. Wang, W. Lin, Z. Fang, and C. Lin, "Objective quality assessment for image retargeting based on structural similarity," *IEEE Journal of Emerging and Selected Topics in Circuits and Systems*, vol. 4, no. 1, pp. 95-105, Mar, 2014.

[24] C. Hsu, C. Lin, Y. Fang, and W. Lin, "Objective quality assessment for image retargeting based on perceptual geometric distortion and information loss," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 3, pp. 377-389, Jun, 2014.

[25] Y. Liang, Y. Liu, and D. Gutierrez, "Objective quality prediction of image retargeting algorithms," *IEEE Trans. Visualization and Computer Graphics*, vol. 23, no. 2, pp. 1099-1110, Feb. 2017.

[26] Q. Jiang, F. Shao, W. Lin and G. Jiang, "Learning sparse representation

for objective image retargeting quality assessment," *IEEE Trans. Cybernetics*, DOI: 10.1109/TCYB.2017.2690452, 2017.

[27] Z. Chen, J. Lin, N. Liao and C. W. Chen, "Full reference quality assessment for image retargeting based on natural scene statistics modeling and bi-directional saliency similarity," *IEEE Trans. Image Processing*, vol. 26, no. 11, pp. 5138-5148, Nov, 2017.

[28] Y. Zhang, W. Lin, Q. Li, W. Cheng and X. Zhang, "Multiple-level feature-based measure for retargeted image quality," *IEEE Trans. Image Processing*, vol. 27, no. 1, pp. 451-463, Jan, 2018.

[29] Y. Zhang, K. N. Ngan, L. Ma and H. Li, "Objective quality assessment of image retargeting by incorporating fidelity measures and inconsistency detection," *IEEE Trans. Image Processing*, vol. 26, no. 12, pp. 5980-5993, Dec, 2017.

[30] J. Fu, H. Zheng and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," *in Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4476-4484, 2017.

[31] B. Shi, X. Bai and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298-2304, Nov, 1 2017.

[32] C. Hsu, F. Chen and G. Wang, "High-resolution image inpainting through multiple deep networks," *in Proc. of International Conference on Vision, Image and Signal Processing (ICVISP)*, pp. 76-81, 2017.

[33] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation," *in Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431-3440, 2015.

[34] K. Zhang, W. Zuo, Y. Chen, D. Meng and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Processing*, vol. 26, no. 7, pp. 3142-3155, Jul, 2017.

[35] A. Kappeler, S. Yoo, Q. Dai and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," *IEEE Trans. Computational Imaging*, vol. 2, no. 2, pp. 109-122, Jun, 2016.

[36] R. Liao, X. Tao, R. Li, Z. Ma and J. Jia, "Video super-resolution via deep draft-ensemble learning," *in Proc. of IEEE International Conference on Computer Vision (ICCV)*, pp. 531-539, 2015.

[37] D. Cho, J. Park, T. H. Oh, Y. W. Tai and I. S. Kweon, "Weakly-and self-supervised learning for content-aware deep Image retargeting," *in Proc. of IEEE International Conference on Computer Vision (ICCV)*, pp. 4568-4577, 2017.

[38] W. Zhang, L. C. Qu, L. Ma L, J. Guang, R. Huang, "Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network," *Pattern Recognition*, vol. 59, pp, 176-187, Nov, 2016.

[39] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 206-220, Feb, 2017.

[40] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, Nov, 2004.

[41] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *in Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 886-893, 2005.

[42] N. Dalal, B. Triggs, C. Schmid, "Human detection using oriented histograms of flow and appearance," *in Proc. of European Conference on Computer Vision (ECCV)*, pp. 428-441, 2006.

[43] P. H. S. Torr, A. Zisserman, "MLESAC: A new robust estimator with application to estimating image geometry," *Computer Vision and Image Understanding*, vol. 78, no. 1, pp. 138-156, Apr, 2000.

[44] Q. Yan, L. Xu, J. Shi and J. Jia, "Hierarchical saliency detection," *in Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1155-1162, 2013.

[45] R. M. Haralick, K. Shanmugam and I. Dinstein, "Textural features for image classification," *IEEE Trans. Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610-621, Nov, 1973.

[46] M. Rubinstein, D. Gutierrez, O. Sorkine and A. Shamir, "A comparative study of image retargeting," *ACM Trans. Graphics*, vol. 29, no. 6, article no. 160, 2010.

[47] L. Ma, W. Lin, C. Deng, K.N. Ngan, "Image retargeting quality assessment: a study of subjective scores and objective metrics," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 626-639, Oct, 2012.

[48] Z. Karni, D. Freedman, and C. Gotsman, "Energy-based image deformation," *Computer Graphics Forum*, vol. 28, no. 5, pp. 1257-1268, 2009.