Ranking-Preserving Cross-Source Learning for Image Retargeting Quality Assessment

Yong-Jin Liu, *Senior Member, IEEE*, Yiheng Han, Zipeng Ye, Yu-Kun Lai, *Member, IEEE*

Abstract—Image retargeting techniques adjust images into different sizes and have attracted much attention recently. Objective quality assessment (OQA) of image retargeting results is often desired to automatically select the best results. Existing OQA methods train a model using some benchmarks (e.g., RetargetMe), in which subjective scores evaluated by users are provided. Observing that it is challenging even for human subjects to give consistent scores for *retargeting results of different source images* (diff-source-results), in this paper we propose a learning-based OQA method that trains a General Regression Neural Network (GRNN) model based on relative scores — which preserve the ranking — of *retargeting results of the same source image* (same-source-results). In particular, we develop a novel training scheme with provable convergence that learns a common base scalar for same-source-results. With this source specific offset, our computed scores not only preserve the ranking of subjective scores for same-source-results, but also provide a reference to compare the diff-source-results. We train and evaluate our GRNN model using human preference data collected in RetargetMe. We further introduce a subjective benchmark to evaluate the generalizability of different OQA methods. Experimental results demonstrate that our method outperforms ten representative OQA methods in ranking prediction and has better generalizability to different datasets.

Index Terms—Image retargeting, image quality assessment, learning to rank, general regression neural network.

1 Introduction

Image retargeting refers to techniques that adjust a source image into different sizes, which has become an increasingly demanded tool with the diversification of display devices. Although a large number of retargeting methods have been developed, no single method works well on arbitrary input images [9], [26], [27]. Subjective quality assessment involving human judgment is usually time-consuming and laborious, and thus unpractical in many situations. As summarized in Section 2, despite recent progress, existing *objective* quality assessment (OQA) methods are still far from ideal in predicting human preference. Therefore, a good OQA method correlating well with human judgements is essential in automatically selecting the best retargeting results and helpful for developing new retargeting methods.

Existing OQA methods train a model using some benchmarks (e.g., [18], [12]) — in which subjective scores evaluated by users are provided — and the absolute subjective scores of all retargeted results from different source images are used indistinguishably for training. A key observation that motivates the work presented in this paper is that in most cases, the subjective scores of retargeted images are only meaningful with the *same* source image. Even for human subjects, it is often difficult to give consistent scores for *retargeting results of different sources* (diff-source-results). An example is shown in Figure 1, in which the two retargeting results 1 and 2 have lower subjective scores, but appear to be more plausible than the results 3 and 4 that have higher scores. Therefore, instead of training a model using the absolute subjective scores indistinguishably for different source images, in this paper we propose a learning-based OQA method that trains a regression model based on the relative scores of *retargeting results of the same source*

Figure 1. Subjective scores are only comparable for retargeting results of the same source image. In each row, two retargeting results are presented and their scores are shown in parentheses (the first numbers). These subjective scores provided in the RetargetMe benchmark [18] are numbers of votes that people cast when comparing this image against other images with the same source image. Higher scores mean better results. Although the scores of the two retargeting results 3 and 4 are higher than the scores of results 1 and 2, we cannot conclude that the results 3 and 4 are better than the results 1 and 2; instead, the opposite appears to be true. The second numbers in parentheses are objective scores output from the method proposed in this paper. The scores not only preserve the ranking of retargeted images with the same source image, but also provide a reference to compare retargeted images from different sources. As a comparison, the third numbers in parentheses are objective scores predicted by [3], which cannot compare retargeted images from different sources.

- *Y.-J. Liu, Y. Han and Z. Ye are with BNRist, the Department of Computer Science and Technology, MOE-Key Laboratory of Pervasive Computing, Tsinghua University, Beijing, China. E-mail: liuyongjin@tsinghua.edu.cn*
- *Y.-K. Lai is with School of Computer Science and Informatics, Cardiff University, UK.*

image (same-source-results), which preserve the ranking and are easy to obtain reliably.

Our method uses the General Regression Neural Network (GRNN) [22] to model a combination of nine known OQA metrics collected from [9], [27]. We train this GRNN model using the human preference data collected in the elaborate RetargetMe benchmark [18]. The GRNN model is known to work effectively with relatively few training samples, which suits our task well due to the limited availability of subjective data. For a source image I, we denote its retargeted images as a set $\mathbf{R}(I)$. Our method is based on a simple idea that if we add a common scalar to all subjective scores of $\mathbf{R}(I)$, their ranking will not be changed. We develop a novel training scheme with provable convergence that learns a common base scalar c_i for $\mathbf{R}(I_i)$, $i = 1, 2, \cdots$. The final score of a retargeted image $R_{ij} \in \mathbf{R}(I_i)$ is $c_i + f_{ij}$, where f_{ij} is the relative score of R_{ij} in $\mathbf{R}(I_i)$.

In our previous conference paper [3], we propose a method for learning to rank retargeted images, which also uses the GRNN model. In this method, the GRNN model takes the features of a pair of retargeted images as input and predicts their *relative quality difference* (RQD). By computing RQDs of all pairs in each $\mathbf{R}(I_i)$, post-processing is needed to transform RQDs into a global ranking. In this paper, we substantially extend and improve upon [3] in four aspects and make the following contributions:

- The GRNN model in [3] treats symmetry and non-symmetry images separately, and in the test phase, the user needs to specify whether the input pair of images are symmetric or not, which requires extra effort. Our new model removes this requirement;
- Unlike the model $\mathcal{F}'(v(R_{ij_a}), v(R_{ij_b}))$ in [3], which takes the features of a pair of retargeted images as input[1], our new model $\mathcal{F}(v(R_{ij}))$ only uses the features of a single retargeted image as input, where $v(R_{ij})$ is a feature representation of R_{ij};
- We propose a novel training scheme with provable convergence, which directly predicts a global score $\mathcal{F}(v(R_{ij}))$ for the input retargeted image R_{ij}, whereas the model in [3] needs a post-process to transform the relative scores $\mathcal{F}'(v(R_{ij_a}), v(R_{ij_b}))$, $\forall R_{ij_a}, R_{ij_b} \in \mathbf{R}(I_i)$, $j_b \neq j_a$, into a global score $f(R_{ij_a})$, which is only meaningful in a retargeted image set $\mathbf{R}(I_i)$ of the same source image I_i;
- The output global scores $\mathcal{F}(v(R_{ij}))$ not only preserve the ranking of same-source-results, but also provide a reference to compare diff-source-results i.e., $\mathcal{F}(v(R_{ij}))$ and $\mathcal{F}(v(R_{i'j'}))$, $i \neq i'$, can be directly compared; see Figure 1.

Experimental results demonstrate that our OQA method correlates better with human judgements than ten representative OQA methods (including [3]) and has better generalizability. We also conduct a new user study using an approach similar to RetargetMe benchmark [18] with better quality control. The novel dataset obtained in this user study will be made publicly available to provide a useful dataset for evaluating *generalizability* of different OQA methods.

1. E.g., $\mathcal{F}'(v(R_{ij_a}), v(R_{ij_b})) > 0$ indicates that R_{ij_a} is better than R_{ij_b}, where R_{ij_a} and R_{ij_b} must be retargeted images of the same source image I_i.

2 Related Work

Image retargeting has attracted considerable attention and many content-aware methods have been developed [20]. To compare different retargeting algorithms, several quality assessment methods have been proposed, which can be divided into two types: subjective and objective methods.

Subjective quality assessment designs elaborate perceptual studies and systematically analyzes user preferences. RetargetMe [18] is a well-established benchmark that contains a decent number of source images and their retargeting results produced by eight representative methods. A comprehensive, comparative subjective study is also included in RetargetMe. It is the first in-depth perceptual study with a large number of users for image retargeting quality assessment. A different subjective study was proposed in [12], in which the user evaluation was carried out by simultaneous double stimulus for continuous evaluation that scored only one retargeted image each time rather than pairwise comparison. Castillo et al. [2] developed an image retargeting survey using eye tracking technology. All these subjective methods can provide good evaluation, but they are laborious and very time-consuming. Nevertheless, these studies provide valuable benchmarks for developing OQA methods. Our method proposed in this paper mainly depends on the RetargetMe benchmark and we further perform an extended user study for evaluating generalizability.

Objective quality assessment (OQA) defines metrics that can be calculated from pixels of images. Edge Histogram (EH) [14] and Color Layout (CL) [7] are two image content based measures in the MPEG-7 standard. They are low-level metrics that treat images as a whole and define image distances based on similarity of edge or color distribution. Bidirectional Similarity (BDS) [21] treats an image as a collection of patches and calculates a bidirectional mapping of these patches between two images as a measure. Bidirectional Warping (BDW) [19] is similar to BDS, but the mapping in BDW takes an asymmetric dynamic time warping, which simultaneously minimizes the warping cost and preserves the patch order. BDS and BDW are relatively easy to calculate; however, they treat every patch as equally important for the final distance and do not take salient regions or aesthetic perspectives into account. Thus their results are not always consistent with subjective ranking. OQA methods based on SIFT flow (SFlow) [10] and Earth-Mover's Distance (EMD) [17] can capture the structural properties more robustly. Liu et al. [11] proposed a top-down model to define a saliency-based image similarity metric in the CIE Lab color space. Recently, an aspect ratio similarity (ARS) metric [26] was proposed, which characterizes how the source image is resized into the target image by geometric changes and provides an efficient solution based on a Markov random field. Noting that human judgment often involves multiple factors, several state-of-the-art methods combine multiple metrics that characterize different factors of image retargeting quality [12], [13], [9], [27].

Our proposed method is inspired by the works in [9], [27] that both elaborately design several novel metrics and develop an OQA method by combining them. Liang et al. [9] combine seven metrics and make use of a linear combination of these metrics, with the weights learned from

the RetargetMe benchmark. This method provides an all-round characterization of retargeted images. However, the linear combination is over-simplified and does not always produce a consistent prediction to human preference. Zhang et al. [27] use three features covering multiple levels, i.e., aspect ratio similarity feature (low level), edge group similarity feature (mid-level) and face block similarity feature (high level). To fuse these three features and map feature scores into quality indices, the Support Vector Regression (SVR) is used for learning. However, in the training process, Zhang's method considers the absolute subjective scores indistinguishably for different source images. In this paper, rather than using the over-simplified linear combination, we propose to use a machine learning approach to provide the necessary flexibility for feature fusion. We also develop a novel training scheme with provable convergence that can learn effective OQA values from relative scores of same-source-results. Experimental results show that our method has better prediction performance than [9], [27] and can predict quality comparable across different source images.

3 A Learning-based OQA Method

The quality of image retargeting depends on multiple factors and composite metrics are needed. In recent work [9], [27], several elaborately designed metrics were proposed. We briefly summarize nine selected metrics $\{Q_1, \cdots, Q_9\}$ in Section 3.1. Given a source image I and a retargeted image R, each metric $Q_i(I, R)$ computes a scalar in $[0, 1]$ to reflect the retargeting quality in one factor.

To construct an objective function $F(Q_1, \cdots, Q_n)$ from a set of selected metrics $\{Q_i\}_{i=1}^n$, an additive value function

$$F = \sum_{i=1}^{n} w_i Q_i \tag{1}$$

is used in [9]. The value of F is in $[0, 1]$ and a lower value of F means better quality. We argue that the linear form in Eq. (1) is over-simplified and we propose to find a better (possibly nonlinear) form for F by machine learning from human preference.

In our study, we pay attention to artificial neural networks (ANNs), which have been well studied and widely used in image processing. The universal approximation theorem [6] states that simple neural networks can represent a wide range of useful functions when given appropriate parameters. Among many types of ANNs, the RBF network is a universal approximator[2] and is a popular alternative to the multi-layer perceptrons, due to its simpler structure and faster training process. Our work in this paper uses the general regression neural network (GRNN) [22], which is a representative RBF network and can obtain good results even with sparse data in a multidimensional measurement space, particularly suitable for our problem.

Zhang et al. [27] also propose a machine learning method that fuses a selected set of metrics $\{Q_i\}_{i=1}^3$ using SVR. Their method directly maps the consolidation of metric values to the subjective scores for all retargeted images from different source images in the training phase. We argue that it is challenging even for human subjects to give consistent scores for retargeting results of different source images, and therefore, only the *relative* scores among retargeted images $\mathbf{R}(I)$ with the same source I are meaningful. If we add a common scalar to the subjective scores in $\mathbf{R}(I)$, their relative scores and ranking in $\mathbf{R}(I)$ will not be changed.

In Section 3.2, we propose to train a model that learns a common scalar c_i for each retargeting set $\mathbf{R}(I_i)$ with the source image I_i. In particular, we represent each retargeted image $R_{ij} \in \mathbf{R}(I_i)$ as a nine-dimensional vector

$$v(R_{ij}) = (Q_1(I_i, R_{ij}), Q_2(I_i, R_{ij}), \cdots, Q_9(I_i, R_{ij})) \tag{2}$$

and learn an objective function $\mathcal{F}$ which aims to achieve

$$\mathcal{F}(v(R_{ij})) = c_i + f(R_{ij}) \tag{3}$$

where $f(R_{ij})$ is the subjective score of R_{ij} in the benchmark dataset. The objective function $\mathcal{F}$ automatically preserves the ranking of retargeting results $\mathbf{R}(I_i)$ and the scalar c_i provides a reference to compare retargeting results from different sources I_i, $i = 1, 2, \cdots$. Accordingly, we call our method *ranking-preserving cross-source (RPCS) learning*.

Thanks to a property of probability estimator in GRNN [22], in Section 3.2 we propose a simple yet novel GRNN training scheme with provable convergence to obtain the objective function $\mathcal{F}$ in Eq. (3).

3.1 Nine metrics

By carefully analyzing existing retargeting methods and their outcomes, we select nine metrics in four categories of critical factors that determine image quality for a retargeting result. These factors and their related metrics are summarized below.

Preservation of global structure. This factor is measured by three metrics Q_1, Q_2 and Q_3.

Both Q_1 and Q_2 evaluate the global structure similarity by a weighted sum of local similarity windows from every pair of pixel correspondence [9]. Q_1 considers the structural similarity between two images by analyzing the degradation of structural information between corresponding windows in I and R using the SSIM metric [23]:

$$Q_1 = \sum_{i=1}^{n_t} (1 - SSIM(p_i, p_i')), \tag{4}$$

and Q_2 applies a VDP2 model [15] of human perception to predict the overall quality of R, when compared to I:

$$Q_2 = \sum_{i=1}^{n_t} (1 - \frac{VDP2(p_i, p_i')}{100}), \tag{5}$$

where n_t is the number of pixels in I, p_i' is the ith pixel of I and p_i is the corresponding pixel in R.

Since humans can easily perceive structure information from edges or contours of objects, Q_3 uses sparse edge groups [28] to measure structure-related distortion [27]. Let $E_k = \{e_i\}$ and $E_k' = \{e_j'\}$ be the kth pair of edge groups in source and retargeted images, respectively.

$$Q_3 = e^{-\beta \sqrt{\frac{1}{n_e} \sum_{k=1}^{n_e} d_c(E_k, E_k')}}, \tag{6}$$

where $\beta = 0.2$, n_e is the number of edge group pairs and $d_c(E_k, E_k')$ is the Chamfer distance between E_k and E_k' [1].

2. That is, the RBF network is not restricted to any particular form and does not require any prior knowledge of the appropriate form.

Preservation of salient regions. This factor is measured by three metrics Q_4, Q_5 and Q_6: the first two deal with general salient regions [25] and the last one is specially designed for facial regions.

Q_4 considers the area change of general salient regions between the source image I and retargeted image R [9]:

$$Q_4 = |A_I - A_R| / \max(A_I, A_R), \tag{7}$$

where A_I and A_R represent the areas of the salient regions in I and R, respectively.

Q_5 considers variations in content as changes in the color histogram of salient regions [16], [9]:

$$Q_5 = \frac{1}{2}\sqrt{\sum_{i=0}^{255}(h'_I - h'_R)^2}, \tag{8}$$

where h'_I and h'_R represent the normalized color histograms in the source and retargeted salient regions, respectively.

Q_6 detects human faces in the source image using the Face++ toolkit[3] and establishes the retargeted faces using the bounding box based on the estimated pixel correspondence [27]:

$$Q_6 = \begin{cases} \frac{1}{n_f}\sum_{i=1}^{n_f} s_{ar}(i), & n_f > 0 \\ 1 & n_f = 0 \end{cases} \tag{9}$$

where n_f is the number of detected faces and $s_{ar}(i)$ is the aspect ratio change of the ith face block pair, defined as

$$s_{ar}(i) = \left[\frac{2r_w(i)r_h(i) + \hat{c}}{r_w^2(i) + r_h^2(i) + \hat{c}}\right] \cdot e^{-\tilde{c}(r_m(i)-1)^2} \tag{10}$$

where $r_w(i)$ and $r_h(i)$ are the width and height change ratios of bounding boxes in the ith block pair, $r_m(i) = \frac{r_w(i)+r_h(i)}{2}$, $\hat{c}$ and $\tilde{c}$ are small constants [26].

Influence of visual distortion and introduced artifacts. This factor is characterized by two metrics Q_7 and Q_8.

Q_7 is a bidirectional similarity metric that takes into account the influence of saliency [21]:

$$\begin{aligned} Q_7 = \quad & 0.5\frac{\frac{1}{N_I}\sum_{U\subset I} S_U \min_{V\subset R} D(U,V)}{\max_{U\subset I}(S_U \min_{V\subset R} D(U,V))} + \\ & 0.5\frac{\frac{1}{N_I}\sum_{V\subset R} S_V \min_{U\subset I} D(U,V)}{\max_{V\subset R}(S_V \min_{U\subset I} D(U,V))}, \end{aligned} \tag{11}$$

where U and V are 3×3 patches from the source and retargeted images respectively, N_I and N_R are the numbers of patches in the source image I and retargeted image R, D is the distance measure between two patches as defined in [21], and S_U and S_V are saliency weights given by the average of the salience values of all pixels contained in patches U and V.

Q_8 measures pixel-level aspect ratio similarity [26], which partitions the source image into dense regular blocks and maps blocks into the retargeted image based on pixel correspondence. Q_8 uses the bounding box of retargeted blocks to estimate the local block deformation:

$$Q_8 = \sum_{i=1}^{n_b} w_i s_{ar}(i) \tag{12}$$

where n_b is the number of blocks, w_i is the weight measured by visual importance and $s_{ar}(i)$ measures the change of aspect ratio for the ith block, as defined in Eq. (10).

3. Available at https://www.faceplusplus.com

Aesthetics. This factor is measured by two rules in computational aesthetics [4], i.e., the rule of thirds T_{third} and visual balance V_{bal}:

$$Q_9 = 0.5T_{third}(I, R) + 0.5V_{bal}(I, R) \tag{13}$$

See [9] for detailed computation for the rules of Q_9.

3.2 Training GRNN for $\mathcal{F}$ with RPCS Learning

3.2.1 Training dataset

We use all the 37 groups of images in RetargetMe dataset [18] — a well-known benchmark in image retargeting — to train and evaluate our OQA model. In this dataset, each group has one source image I_i and eight retargeted images $R_{ij} \in \mathbf{R}(I_i)$, $i = 1, 2, \cdots, 37$, $j = 1, 2, \cdots, 8$. We partition the 37 groups into two classes: one for training and the other for testing (Section 4). Hereafter, we denote the training set as Ω_T and the groups in it as $(I_i, \mathbf{R}(I_i)) \subseteq \Omega_T$.

In RetargetMe, a comparative user study based on *linked-paired comparison design* [5] was performed to ensure balanced voting. Three complete sets were collected for each retargeted image to guarantee statistical robustness. Each time a participant was shown two retargeted images side by side, and was asked to simply choose the one he/she liked better. Each retargeted image appeared 3 times for a participant and judged by 21 participants, meaning that a retargeted image received a maximum of $21 \times 3 = 63$ votes. The number of votes for a retargeted image shows the subjective quality by human observers. As demonstrated in Figure 1, such subjective scores cannot be used to effectively compare human preference with *different* source images, but work reasonably well for retargeted images with the *same* source image. In Section 3.2.2, we use normalized subjective scores which are the numbers of votes divided by 63.

3.2.2 Ranking-preserving cross-source learning

Unlike the multi-level feature fusion method [27], which uses SVR to train an objective function $F(v(R_{ij})) \approx f(R_{ij})$, we target on training an objective function aiming to satisfy Eq. (3), in which $f(R_{ij})$ is updated to the normalized subjective score of R_{ij}, $R_{ij} \in \Omega_T$.

To achieve this goal, we extract one retargeted image R_{i*} from each image group $(I_i, \mathbf{R}(I_i)) \subseteq \Omega_T$ and denote the remaining retargeted images of I_i as $\widetilde{\mathbf{R}}(I_i) = \mathbf{R}(I_i) \setminus \{R_{i*}\}$. Let $\Omega_{T*} = \bigcup_i (I_i, \widetilde{\mathbf{R}}(I_i))$ and $\mathbf{R}_* = \bigcup_i \{R_{i*}\}$.

Our training process is iterative and each iteration contains two steps. At iteration k ($k > 0$), in the first step, we train the GRNN model using Ω_{T*}, aiming to achieve

$$\mathcal{F}_k(v(R_{ij})) = f_k(R_{ij}) \tag{14}$$

where $R_{ij} \in \Omega_{T*}$ and $f_k(R_{ij})$ is the kth training score of R_{ij}, initialized by $f_1(R_{ij}) = f(R_{ij})$, i.e., the normalized subjective score in the RetargetMe dataset.

We model $\mathcal{F}_k$ using GRNN, due to its approximation capability with relatively few training samples. The input to this model is a feature vector v of a retargeted image R_{ij}, which is a concatenation of nine metric values in Eq. (2). We use the standard configuration for our GRNN model with the output layer being a scalar corresponding to the predicted score $\mathcal{F}_k(v(R_{ij}))$. The spread parameter σ in

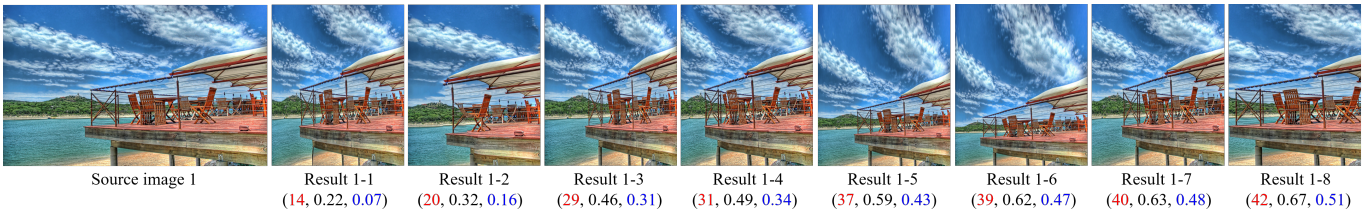

Figure 2. Two image groups in the RetargetMe benchmark [18]: each group has a source image and eight retargeted images. For each retargeted image, the numbers in parentheses are its subjective score (red), normalized subjective score (black) and the objective score computed by our method (blue). For each group, the difference between normalized subjective score and the objective score is a constant, and therefore, the objective scores predicted by our method preserve the ranking of subjective scores. Subjective scores are only comparable for retargeting results of the same source image. For example, although the subjective score of result 1-2 is higher than the subjective score of result 2-2, result 2-2 appears to be better than result 1-2. The objective scores computed by our method reveal this fact.

GRNN controls the influence range of radial basis functions and is set to 1.4 in our experiments.

In the second step, we evaluate the trained GRNN model $\mathcal{F}_k$ using $\mathbf{R}_*$ and update the training scores of $R_{ij} \in \Omega_{T*}$. In more details, for each $R_{i*} \in \mathbf{R}_*$, we compute $\mathcal{F}_k(v(R_{i*}))$ and update the training scores for all $R_{ij} \in \widetilde{\mathbf{R}}(I_i)$:

$$f_{k+1}(R_{ij}) = f(R_{ij}) + \frac{1}{2}(\mathcal{F}_k(v(R_{i*})) - f(R_{i*})) \tag{15}$$

In Section 3.2.3, we prove that this simple two-step iteration scheme converges quickly at the cth iteration, which satisfies

$$\begin{aligned} &f_c(R_{ij}) - f(R_{ij}) = \mathcal{F}_c(v(R_{i*})) - f(R_{i*}), \\ &\forall R_{ij} \in \widetilde{\mathbf{R}}(I_i),\ \widetilde{\mathbf{R}}(I_i) \in \Omega_{T*} \end{aligned} \tag{16}$$

Then $c_i = \mathcal{F}_c(v(R_{i*})) - f(R_{i*})$ is the learned common base scalar for the ith image group in Ω_T, which provides a reference to compare the retargeting results of different source images.

Two examples are illustrated in Figure 2. The pseudo-code is summarized in Algorithm 1.

3.2.3 Proof of convergence

Let n_g be the number of image groups in the training set Ω_t. Without loss of generality, we assume $\forall i,\ R_{i*} = R_{i8}$.

Given the training data $(v(R_{ij}), f_k(R_{ij}))$, $i = 1, 2, \cdots, n_g$, $j = 1, 2, \cdots, 7$, where $v(R_{ij})$ is an instance of an independent variable v and $f_k(R_{ij})$ is the corresponding instance of a dependent variable $\mathcal{F}_k(v)$, the learned GRNN model $\mathcal{F}_k$ can be represented by [22]

$$\mathcal{F}_k(v) = \frac{\sum_{i=1}^{n_g}\sum_{j=1}^{7} f_k(R_{ij}) e^{-\frac{D_{ij}^2}{2\sigma^2}}}{\sum_{i=1}^{n_g}\sum_{j=1}^{7} e^{-\frac{D_{ij}^2}{2\sigma^2}}} \tag{17}$$

where

$$D_{ij}^2 = (v - v(R_{ij}))^T (v - v(R_{ij})) \tag{18}$$

In the second step of the kth iteration, we predict the score $\mathcal{F}_k(v(R_{i*}))$ of each $R_{i*} \in \mathbf{R}_*$ using the learned model

Algorithm 1 Ranking-preserving cross-source learning

Input: A training set Ω_t with n_g image groups in RegargetMe dataset.
Output: A trained GRNN model $\mathcal{F}$ satisfying Eq. (16).

for each image group $(I_i, \mathbf{R}(I_i))$ in Ω_t **do**
  Compute the mean subjective score m of eight retargeted images in $\mathbf{R}(I_i)$ and select the retargeted image whose subjective score is closest to m as R_{i*}.
  Re-index the set $\mathbf{R}(I_i)$ such that $R_{i*} = R_{i8}$.
  Set $\widetilde{\mathbf{R}}(I_i) = \mathbf{R}(I_i) \setminus \{R_{i*}\}$.
end for
Set $\Omega_{T*} = \bigcup_i (I_i, \widetilde{\mathbf{R}}(I_i))$ and $\mathbf{R}_* = \bigcup_i \{R_{i*}\}$.
for each retargeted image R in Ω_t **do**
  Set $f(R) =$ normalized subjective score of R.
end for
Initialize $\varepsilon = 1$
while $\varepsilon > 10^{-3}$ **do**
  $\varepsilon = 0$.
  Train the GRNN model $\mathcal{F}$ using Ω_{T*} (ref. Eq. (14)).
  for each retargeted image R_{i*} in $\mathbf{R}_*$ **do**
    Evaluate the trained GRNN model $\mathcal{F}$ by computing $\varepsilon_i = \mathcal{F}(v(R_{i*})) - f(R_{i*})$.
    Update $\varepsilon = \varepsilon + |\varepsilon_i|$.
    for each retargeted image R_{ij} in $\widetilde{\mathbf{R}}(I_i)$ **do**
      Update $f(R_{ij}) = f(R_{ij}) + \frac{\varepsilon_i}{2}$
    end for
  end for
end while
Output $\mathcal{F}$

in Eq. (17). To express this prediction in a matrix form, we pack all predicted scores of $\mathbf{R}_*$ into an $n_g \times 1$ vector $\mathbf{B}_k$:

$$\mathbf{B}_k = \begin{pmatrix} \mathcal{F}_k(v(R_{1*})) & \cdots & \mathcal{F}_k(v(R_{i*})) & \cdots & \mathcal{F}_k(v(R_{n_g*})) \end{pmatrix}^T \tag{19}$$

and all kth training scores of $\widetilde{\mathbf{R}}(I_i)$ into a $7n_g \times 1$ vector $\mathbf{A}_k$:

$$\mathbf{A}_k = \begin{pmatrix} A_1 & \cdots & A_i & \cdots & A_{n_g} \end{pmatrix}^T \tag{20}$$

where $A_i = \begin{pmatrix} f_k(R_{i1}) & f_k(R_{i2}) & \cdots & f_k(R_{i7}) \end{pmatrix}^T$ is a 7×1

	Lines/edges	Faces/people	Texture	Foreground objects	Geometric structure	Symmetry	All	p-value
BDS [21]	0.040	0.190	0.089	0.167	−0.004	−0.012	0.083	0.017
BDW [19]	0.031	0.048	−0.009	0.060	0.004	0.119	0.046	0.869
EH [14]	0.043	−0.076	−0.063	−0.079	0.103	0.298	0.004	0.641
CL [7]	−0.023	−0.181	−0.089	−0.183	−0.009	0.214	−0.068	0.384
SFlow [10]	0.097	0.252	0.161	0.218	0.085	0.071	0.145	0.031
CSim [11]	0.091	0.271	0.188	0.258	0.063	−0.024	0.151	0.028
Liang's [9]	0.351	0.271	0.304	0.381	0.415	**0.548**	0.399	$5e{-12}$
ARS [26]	0.463	0.519	0.444	0.330	0.505	0.464	0.452	$1e{-11}$
MLF [27]	0.486	0.605	0.384	0.544	0.536	0.536	0.512	$1e{-14}$
L2Rank [3]	0.437	0.505	0.429	0.536	0.438	0.536	0.473	$6e{-13}$
Ours	**0.591**	**0.619**	**0.445**	**0.611**	**0.607**	0.476	**0.575**	$1e{-17}$

Table 1
The mean Kendall correlation coefficients of 37 groups of images in RetargetMe. The last column shows p-value over all image types.

sub-vector. Then the matrix form of Eq. (17) is:

$$\mathbf{B}_k = \mathbf{G}\mathbf{A}_k \tag{21}$$

where $\mathbf{G}$ is an $n_g \times 7n_g$ matrix, whose (p, q) entry is

$$\mathbf{G}(p,q) = \frac{e^{-\frac{\widetilde{D}_{pq}^2}{2\sigma^2}}}{\sum_{l=1}^{7n_g} e^{-\frac{\widetilde{D}_{pl}^2}{2\sigma^2}}} \tag{22}$$

$$\widetilde{D}_{pq}^2 = (v(R_{p*}) - v(R_{xy}))^T (v(R_{p*}) - v(R_{xy})) \tag{23}$$

R_{p*} is the retargeted image corresponding to the pth entry in $\mathbf{B}_k$ and R_{xy} is the retargeted image corresponding to the qth entry in $\mathbf{A}_k$, i.e., $x = \lfloor \frac{q}{7} \rfloor$ and $y = q - 7x$.

Similarly, to express Eq. (14) in a matrix form, we pack the subjective scores of $R_{i*} \in \mathbf{R}_*$ into an $n_g \times 1$ vector $\mathbf{B}_0$:

$$\mathbf{B}_0 = \begin{pmatrix} f(v(R_{1*})) & \cdots & f(v(R_{i*})) & \cdots & f(v(R_{n_g*})) \end{pmatrix}^T \tag{24}$$

and pack the subjective scores of $R_{ij} \in \Omega_{T*}$ into a $7n_g \times 1$ vector $\mathbf{A}_0$:

$$\mathbf{A}_0 = \begin{pmatrix} \overline{A}_1 & \cdots & \overline{A}_i & \cdots & \overline{A}_{n_g} \end{pmatrix}^T \tag{25}$$

where $\overline{A}_i = \begin{pmatrix} f(R_{i1}) & f(R_{i2}) & \cdots & f(R_{i7}) \end{pmatrix}^T$ is a 7×1 sub-vector. Note that we initialize the iteration by setting $\mathbf{A}_1 = \mathbf{A}_0$.

Now Eq. (14) can be re-expressed by

$$\mathbf{A}_{k+1} - \mathbf{A}_0 = \frac{1}{2}\mathbf{Q}(\mathbf{B}_k - \mathbf{B}_0), \quad k = 1, 2, \cdots \tag{26}$$

where $\mathbf{Q}$ is a $7n_g \times n_g$ matrix:

$$\mathbf{Q} = \begin{pmatrix} Q_1 & \cdots & Q_i & \cdots & Q_{n_g} \end{pmatrix}^T \tag{27}$$

where

$$Q_i = \begin{array}{c} \begin{matrix} \text{1st col} & & i\text{th col} & & n_g\text{th col} \end{matrix} \\ \begin{pmatrix} 0 & \cdots & 1 & \cdots & 0 \\ 0 & \cdots & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & \cdots & 0 \end{pmatrix} \end{array} \tag{28}$$

is a $7 \times n_g$ sub-matrix, in which the ith column is filled by 1 and all other entries are 0.

By substituting Eq. (21) into Eq. (26), we have

$$\mathbf{A}_{k+1} = \frac{1}{2}\mathbf{Q}\mathbf{G}\mathbf{A}_k + \mathbf{C}, \quad k = 1, 2, \cdots \tag{29}$$

where $\mathbf{C} = \mathbf{A}_0 - \mathbf{Q}\mathbf{B}_0$ is a constant matrix.

Let $\mathbf{M} = \frac{1}{2}\mathbf{Q}\mathbf{G}$ and $\mathbf{M}_1 = \mathbf{Q}\mathbf{G}$. To prove that the iteration scheme specified in Eq. (29) converges for any $\mathbf{C}$ and $\mathbf{A}_0$, we need to show that the spectral radius of the iteration matrix $\mathbf{M}$ is less than unity, i.e., $\rho(\mathbf{M}) < 1$ (ref. Theorem 4.1 in [24]).

Note that each entry $\mathbf{G}(p, q)$ in the matrix $\mathbf{G}$ is a non-negative real number representing a probability and each row in $\mathbf{G}$ sums to 1, and then $\mathbf{G}$ is a right stochastic matrix. $\mathbf{M}_1 = \mathbf{Q}\mathbf{G}$, meaning that $\mathbf{M}_1$ repeats every row in $\mathbf{G}$ seven times, and then is again a right stochastic matrix. Since the spectral radius of every right stochastic matrix is at most 1 [8], we have $\rho(\mathbf{M}_1) \leq 1$ and $\rho(\mathbf{M}) \leq \frac{1}{2}$. Then the iteration in Eq. (29) converges for any $\mathbf{C}$ and $\mathbf{A}_0$. □

4 Experiments

We implemented the proposed OQA method in MATLAB and the source code is available[4]. We compare our method with ten representative OQA methods: BDS [21], BDW [19], EH [14], CL [7], SFlow [10], CSim [11], Liang's method [9], ARS [26], MLF [27] and learn-to-rank (L2Rank) [3]. The comparison is performed in three experiments. The first is the leave-one-out cross validation on the RetargetMe benchmark [18] (Section 4.1) and the second is a generalizability evaluation on a novel dataset constructed in a new user study (Section 4.2). Since L2Rank uses the same GRNN model and six metrics as ours, finally we make a detailed ablation study and comparison with L2Rank (Section 4.3).

4.1 Leave-one-out cross validation on RetargetMe

RetargetMe has 37 groups of images with subjective preference scores and each group has one source image and eight retargeted images. These 37 groups are classified into six types: lines/edges (25), faces/people (15), texture (6), foreground objects (18), geometric structure (16) and symmetry (6). These classifications are not mutually exclusive, meaning that one image can belong to more than one type.

To verify the performance of our method and compare it with eight representative methods, we apply leave-one-out cross validation (LOOCV) in RetargetMe. In each fold of LOOCV, one group is used as the test set, with the remaining groups as the training set. After 37 folds, each group has been used as a test set once.

To estimate how well the objective ranking agrees with the participants' subjective voting, we follow the method in [18] to use the Kendall correlation coefficient τ. The value of τ is in $[-1, 1]$ and higher value means better agreement. The results are summarized in Table 1, classified according to six image types. We also compute the mean Kendall correlation coefficient using all the images (last column in Table 1). The results show that except for the symmetry type, our method

4. http://cg.cs.tsinghua.edu.cn/people/~Yongjin/Yongjin.htm

(a) Source image (b) Checkpoint image pair 1 (c) Checkpoint image pair 2 (d) Checkpoint image pair 3

Figure 3. Three check point pairs. In (b) and (c), retargeted images on the left are obviously better than those on the right. In (d), the retargeted image on the right is obviously better than the one on the left.

	Lines/edges	Faces/people	Texture	Foreground objects	Geometric structure	Symmetry	All	p-value
Liang's [9]	0.250	0.381	0.214	0.295	0.082	0.232	0.313	$8e{-4}$
ARS [26]	0.351	0.345	0.571	0.371	**0.388**	**0.607**	0.313	$6e{-5}$
MLF [27]	0.240	0.471	0.500	0.352	0.224	**0.500**	0.330	$9e{-4}$
L2Rank [3]	**0.393**	**0.524**	**0.786**	**0.400**	0.347	0.339	**0.407**	$5e{-7}$
Ours	**0.435**	**0.543**	**0.743**	**0.481**	**0.367**	0.429	**0.445**	$2e{-4}$

Table 2

The mean Kendall correlation coefficients of 26 groups of images in our novel dataset. The top two results of each type are shown in bold.

consistently produces significantly better results than all other methods. The degraded performance on the symmetry type is possibly due to the lack of sufficient training data, i.e., only five symmetry images for training in LOOCV. In Table 1, to evaluate the statistical significance, we follow [18] to use p-value in statistical hypothesis testing: $p < 0.01$ indicates significant results.

4.2 Generalizability evaluation on a novel dataset

To evaluate the generalizability of OQA methods to *different* image datasets, we conducted a new user study on 26 new groups selected in RetargetMe that lack subjective scores[5]. These 26 groups are also classified into six types: lines/edges (11), faces/people (5), texture (1), foreground objects (15), geometric structure (7) and symmetry (4).

The original web-based user study in RetargetMe [18] was based on the linked-paired comparison design [5]. In the website of the survey, two retargeted images and the source image were shown simultaneously at each time. Each participant was asked to choose the retargeted image with better quality. To avoid unreliable user input such as random picking, we extend the web-based user study in RetargetMe by adding *checkpoint input* and *time check* for quality control. Any user input failed in either of these two checks is discarded.

Checkpoint input refers to three special pairs of retargeted images with obvious preference (Figure 3). In each user study session, these image pairs were randomly distributed, in which the obviously better images were located on the left in two occasions and on the right in one occasion. According to our preparatory experiments, participants with high concentration can easily choose correct images, while those who just randomly select images are likely to fail in at least one checkpoint input.

Time check is a constraint that the average selection time for an input image pair should not be shorter than 3 seconds. In our preparatory experiments, we found that setting a fixed time limit for each image pair does not provide reliable indication as some cases are genuinely easier to decide than others. However, the average selection time is effective in differentiating reliable and unreliable user input. A participant who randomly selects images may still pass the checkpoint input test by chance, but their average selection time is likely to be much shorter than proper input.

We employed 232 participants who were postgraduate students in research labs from Australia, UK, Canada, China and USA. 168 of them passed all the checks and their subjective scores were collected for 26 groups of images.

To evaluate the generalizability of OQA methods, we use 37 groups of images from RetargetMe with provided subjective scores as the training set. The trained model is then applied to the novel dataset with 26 new groups of images. We compare top four methods (i.e., Liang's method [9], ARS [26], MLF [27] and L2Rank [3]) as indicated in Table 1. Among five methods, three (MLF, L2Rank and ours) train a regression model and their training complexities are $O(n^3)$, $O(n^2)$ and $O(n^2)$, respectively, where n is the number of samples in the training set. On a PC with an Intel i7-8700 CPU and 16 GB RAM, the training times[6] are 7.7 seconds (MLF), 0.44 seconds (L2Rank) and 0.74 seconds (ours). Our method is only slightly slower than L2Rank with the same asymptotic complexity. To evaluate an image or an image pair, nearly all time is used to extract features. Our method (9.0 seconds) is slower as it includes all the features of MLF (7.1 seconds) and L2Rank (1.9 seconds). The results on the novel dataset are summarized in Table 2, showing that 1) In the image types of lines/edges, faces/people and foreground objects, our method outperforms all other methods. 2) In the image types of texture and geometric structure, our method is ranked second and close to the top one. This may be because there is only one group in the texture type and ARS specifically considers geometric changes while our method is much more balanced on all image types. 3) Overall, our method has better performance than all other methods. Our method does not perform well in the symmetry image type. We will improve our model by considering more reliable symmetry features and training on more symmetry images. We put this in the future work.

4.3 Ablation study and comparison with L2Rank

Both L2Rank [3] and our method use GRNN. Meanwhile, L2Rank also takes image features as input, i.e., it considers

5. There are 80 groups in RegargetMe. Only 37 of them have subjective preference scores. From the remaining groups, we chose all the groups without substantial similarity to those in the original 37 groups.

6. Image features are pre-stored in the training data and the training time does not include image feature extraction.

Figure 4. Two examples not in RetargetMe are illustrated, each of which has two retargeting results. Result 4 is obviously better than Result 2 because human observers are more sensitive to the change of human subjects. The scores predicted by our method provide a correct reference ($0.87 > 0.65$), while the scores predicted by L2Rank are in the wrong order due to cross source ($0.0011 < 0.0023$).

six (without Q_3, Q_6 and Q_8) from nine metrics in Section 3.1. Our method can evaluate retargeted images with difference sources (Figure 1), while L2Rank can only evaluate retargeted images with the same source; see the visual comparison of two examples in RetargetMe (Figure 1) and two examples not in RetargetMe (Figure 4). In addition to this significant difference, below we show that even for retargeted images with the same source, our method has higher Kendall correlation coefficient τ (indicating better agreement with subjective voting) than L2Rank.

To evaluate the role of nine metrics and the proposed RPCS learning, we compare L2Rank and our method with the six metrics in [3] (denoted as L2Rank and Ours_6), and with all nine metrics in Section 3.1 (denoted as L2Rank_9 and Ours). We choose GRNN because it works well with relatively few training samples. We compare with alternative regression models: support vector regression (Ours_SVR), random forest (Ours_RF) and extreme learning machine (Ours_ELM). The mean τ values of LOOCV in RetargetMe are: Ours (0.575), Ours_SVR (0.524), Ours_RF (0.488), Ours_ELM (0.521), L2Rank_9 (0.519), L2Rank (0.473) and Ours_6 (0.499), showing that the GRNN model, three additional metrics (Q_3, Q_6 and Q_8) and RPCS learning can effectively improve the OQA performance.

5 Conclusion

In this paper, we propose a simple yet effective learning method for image retargeting quality assessment. After representing a retargeted image in a nine-dimensional vector representation using nine metrics selected from [9], [26], [27], we propose a novel training scheme with provable convergence to train a GRNN model with the subjective preference scores from RetargetMe [18]. Experiments show that our method consistently outperforms ten representative OQA methods.

Acknowledgements

This work was supported by the Natural Science Foundation of China (61725204, 61521002, U1736220) and Royal Society-Newton Advanced Fellowship (NA150431).

References

[1] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *Intl. J. Conf. Artificial Intelligence (IJCAI)*, pages 659–663, 1977.

[2] S. Castillo, T. Judd, and D. Gutierrez. Using eye-tracking to assess different image retargeting methods. In *ACM SIGGRAPH Symp. Applied Perception in Graph. and Vis. (APGV)*, pages 7–14, 2011.

[3] Y. Chen, Y.-J. Liu, and Y.-K. Lai. Learning to rank retargeted images. In *IEEE Conf. Comp. Vis. Patt. Recog. (CVPR)*, pages 4743–4751, 2017.

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *European Conf. Computer Vision (ECCV)*, pages 288–301, 2006.

[5] H. A. David. *The Method of Paired Comparisons*, volume 12. DTIC Document, 1963.

[6] K.-L. Du and M. N. Swamy. *Neural Networks and Statistical Learning*. Springer London, 2013.

[7] E. Kasutani and A. Yamada. The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval. In *IEEE Intl. Conf. Image Processing (ICIP)*, volume 1, pages 674–677, 2001.

[8] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA SIAM, 1st edition, 1999.

[9] Y. Liang, Y.-J. Liu, and D. Gutierrez. Objective quality prediction of image retargeting algorithms. *IEEE Trans. Vis. Comp. Graph.*, 23(2):1099–1110, 2017.

[10] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman. Sift flow: Dense correspondence across different scenes. In *European Conf. Computer Vision (ECCV)*, pages 28–42, 2008.

[11] Y.-J. Liu, X. Luo, Y.-M. Xuan, W.-F. Chen, and X.-L. Fu. Image retargeting quality assessment. *Comp. Graph. Forum*, 30(2):583–592, 2011.

[12] L. Ma, W. Lin, C. Deng, and K. N. Ngan. Image retargeting quality assessment: a study of subjective scores and objective metrics. *IEEE J. Sel. Top. Signal Processing*, 6(6):626–639, 2012.

[13] L. Ma, W. Lin, C. Deng, and K. N. Ngan. Study of subjective and objective quality assessment of retargeted images. In *IEEE Intl. Symp. Circuits and Systems (ISCAS)*, pages 2677–2680, 2012.

[14] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Trans. Circuits and Systems for Video Technology*, 11(6):703–715, 2001.

[15] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph.*, 30(4):40:1–40:14, 2011.

[16] C. L. Novak and S. A. Shafer. Anatomy of a color histogram. In *IEEE Conf. Comp. Vis. Patt. Recog. (CVPR)*, pages 599–605, 1992.

[17] O. Pele and M. Werman. Fast and robust earth mover's distances. In *IEEE Intl. Conf. Computer Vision (ICCV)*, pages 460–467, 2009.

[18] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir. A comparative study of image retargeting. *ACM Trans. Graph.*, 29(6):160, 2010.

[19] M. Rubinstein, A. Shamir, and S. Avidan. Multi-operator media retargeting. *ACM Trans. Graph.*, 28(3):23, 2009.

[20] A. Shamir, O. Sorkine, and A. Hornung. Modern approaches to media retargeting. *SIGGRAPH Asia Courses*, 2012.

[21] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani. Summarizing visual data using bidirectional similarity. In *IEEE Conf. Comp. Vis. Patt. Recog. (CVPR)*, pages 1–8, 2008.

[22] D. F. Specht. A general regression neural network. *IEEE Trans. Neural Networks*, 2(6):568–576, 1991.

[23] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612, 2004.

[24] S. Yousef. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, Philadelphia, 2nd edition, 2003.

[25] L. Zhang and W. Lin. *Selective Visual Attention: Computational Models and Applications*. Wiley-IEEE Press, 1st edition, 2013.

[26] Y. Zhang, Y. Fang, W. Lin, X. Zhang, and L. Li. Backward registration-based aspect ratio similarity for image retargeting quality assessment. *IEEE Trans. Image Processing*, 25(9):4286–4293, 2016.

[27] Y. Zhang, W. Lin, Q. Li, W. Cheng, and X. Zhang. Multiple-level feature-based measure for retargeted image quality. *IEEE Trans. Image Processing*, 27(1):451–463, 2018.

[28] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *European Conf. Computer Vision (ECCV)*, pages 391–405, 2014.