10-2017

# FastShrinkage: Perceptually-aware retargeting toward mobile platforms

Zhenguang LIU

Zepeng WANG

Luming ZHANG

Rajiv Ratn SHAH
*Singapore Management University*, rajivshah@smu.edu.sg

Yingjie XIA

***See next page for additional authors***

**DOI:** https://doi.org/10.1145/3123266.3123377

**Author**
Zhenguang LIU, Zepeng WANG, Luming ZHANG, Rajiv Ratn SHAH, Yingjie XIA, Yi YANG, and Wei LIU

# FastShrinkage: Perceptually-aware Retargeting Toward Mobile Platforms

Zhenguang Liu
Zhejiang University
liuzhenguang2008@gmail.com

Zepeng Wang
Hefei University of Technology
zepengw@gmail.com

Luming Zhang
Hefei University of Technology
zglumg@gmail.com

Rajiv Ratn Shah
Singapore Management University
rajivshah@smu.edu.sg

Yingjie Xia
Zhejiang University
xiayingjie@zju.edu.cn

Yi Yang
University of Technology Sydney
yee.i.yang@gmail.com

Xuelong Li
Chinese Academy of Sciences
xuelong_li@opt.ac.cn

## ABSTRACT

Retargeting aims at adapting an original high-resolution photo/video to a low-resolution screen with an arbitrary aspect ratio. Conventional approaches are generally based on desktop PCs, since the computation might be intolerable for mobile platforms (especially when retargeting videos). Besides, only low-level visual features are exploited typically, whereas human visual perception is not well encoded. In this paper, we propose a novel retargeting framework which fast shrinks photo/video by leveraging human gaze behavior. Specifically, we first derive a geometry-preserved graph ranking algorithm, which efficiently selects a few salient object patches to mimic human gaze shifting path (GSP) when viewing each scenery. Afterward, an aggregation-based CNN is developed to hierarchically learn the deep representation for each GSP. Based on this, a probabilistic model is developed to learn the priors of the training photos which are marked as aesthetically-pleasing by professional photographers. We utilize the learned priors to efficiently shrink the corresponding GSP of a retargeted photo/video to be maximally similar to those from the training photos. Extensive experiments have demonstrated that: 1) our method consumes less than 35ms to retarget a $1024 \times 768$ photo (or a $1280 \times 720$ video frame) on popular iOS/Android devices, which is orders of magnitude faster than the conventional retargeting algorithms; 2) the retargeted photos/videos produced by our method outperform its competitors significantly based on the paired-comparison-based user study; and 3) the learned GSPs are highly indicative of human visual attention according to the human eye tracking experiments.

## CCS CONCEPTS

**•Human-centered computing → Ubiquitous and mobile computing systems and tools;**


*MM'17, October 23–27, 2017, Mountain View, CA, USA.*

DOI: https://doi.org/10.1145/3123266.3123377

## KEYWORDS

Mobile platform; Retarget; Perceptual; Deep feature; Probabilistic model

## 1 INTRODUCTION

With the widespread usage of mobile devices, retargeting has becoming an indispensable technique which optimally displays the original high-resolution photo/video on a low-resolution screens with an arbitrary aspect ratio. For example, users usually want to set their iPhone wallpaper as their favorite pictures. So how to effectively adapt an $3264 \times 2448$ photo taken by a DSLR to a $750 \times 1334$ iPhone screen? Non-uniform scaling may lead to visual distortion if the photo contains multiple semantic objects, *e.g.*, human/animal faces and vehicle wheels. Meanwhile, simple photo cropping does not work when the aesthetically-pleasing visual contents are scattered inside a photo. To achieve a semantically-reasonable and well-aesthetic retargeting result, content-aware photo retargeting is developed, which maximally preserves the visually salient regions while keeping the non-salient ones to a minimum scale. Nevertheless, the existing content-aware photo retargeting algorithms are still frustrated by the following drawbacks:

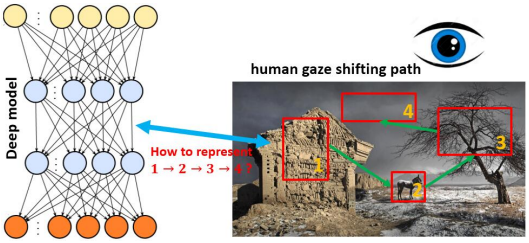


**Figure 1: Encoding human gaze shifting path using an ordered patch sequence $1 \rightarrow 2 \rightarrow 3 \rightarrow 4$, whereas the existing deep networks can only represent a single patch.**

- They may not work efficiently on mobile platforms, although a large quantity of photo/video retargeting tasks are carried out based on iOS/Android devices. For example, it will take a few seconds to process each $3264 \times 2448$

photo using the well-known seam carving [1] on a desktop PC, let alone for mobile platforms. With the assistance of Nvidia CUDA GPU, retargeting can be greatly accelerated on desktop platforms [13]. But how to design a real-time retargeting system on mobile platforms remains a tough challenge;

- It is generally acknowledged that shallow features are less descriptive than the deep features. However, existing retargeting algorithms are generally based on shallow features. Retargeting using deep features might be intolerably time-consuming because of the relatively low performance of mobile processors. Moreover, high-level semantic clues cannot be discovered effectively and efficiently. Even for desktop computers, it may takes seconds to extract the region-level semantic feature from each image/video, such as the object bank [19] and weakly-supervised region semantic encoding [38];
- It is essential to incorporate human visual perception into the retargeting process (as shown in Fig. 1), since viewers generally expect a perceptually well-aesthetic retargeting result. However, current retargeting models can hardly reflect human visual perception, *i.e.*, the human gaze allocations when viewing each image or video clip. Furthermore, current deep models are typically based on images or image patches, they cannot explicitly represent an ordered set of image regions that are sequentially perceived by human eye.

To solve the above problems, we propose a perceptually-aware model which efficiently shrinks the original photo/video by deeply encoding human gaze shifting sequences. Our approach involves three key modules. By extracting a succinct set of object patches from each photo or video frame, a fast graph ranking algorithm is developed to sequentially recognize highly salient object patches for constructing gaze shifting paths (GSPs), wherein the geometrical clue of photo/video are optimally encoded. Since the GSPs are 2D features which may not be explicitly utilized by the existing probabilistic models, we propose an aggregation deep network which sequentially concatenates the object patches along each GSP into its deep representation. Based on the deep representation, we learn the GSP distribution from a large quantity of aesthetically-pleasing photos crawled from Flickr. The learned priors well reflects how human perceives well-aesthetic sceneries, which are then utilized to guide the photo/video retargeting process. Theoretically, we can enforce that the GSP of the test photo/video is maximally similar to those from the well-aesthetic Flickr photos. Computational time analysis have demonstrated that our proposed retargeting system can run in real-time on the state-of-the-art iOS/Android devices. Moreover, comprehensive user studies have shown that photos/videos retargeted by our method are more visually attractive and better preserve semantically important objects than its competitors.

The main contributions of this work can be summarized as follows. First, we propose a geometry-preserving graph ranking algorithm which efficiently and effectively select visually/semantic patches for building a GSP. Second, an aggregation-based deep model is developed for learning the deep feature of each GSP, which is more descriptive than the shallow features. Third, a unified probabilistic model is proposed for photo/video retargeting, wherein experiences of multiple Flickr users and auxiliary visual clues can be flexibly encoded.

## 2 RELATED WORK

Many content-aware retargeting algorithms have been proposed in the literature. They can roughly be categorized into the discrete and continuous retargeting[1]. For the former, a seam (8-connected path of pixels from top to bottom or from left to right) is iteratively removed to preserve the important pixels within a photo. Further, Avidan *et al.* [1] formulated seam detection as dynamic programming, where a gradient energy is employed as the importance map. Bubinstein *et al.* [32] introduced a forward energy criterion to improve Avidan *et al.*'s work. As a variant of seaming, Pritch *et al.* [30] proposed to discretely remove repeated patterns in homogenous image regions. For continuous retargeting, Wolf *et al.* [43] proposed to merge less important pixels in order to reduce distortion. Wang *et al.* [42] proposed an optimized scale-and-stretch approach, which iteratively wraps local regions to match the optimal scaling factors as close as possible. In [36], Sun *et al.* proposed an algorithm to create thumbnails from input images. Two thumbnailing algorithms, termed SOATtp and SOATcr, have been designed to combine the scale and object aware saliency with image retargeting and thumbnail cropping respectively. In [10], Guo *et al.* presented an effective image retargeting method using saliency-based mesh parametrization, which optimally preserves image structures. Since many approaches cannot effectively preserve structural lines, Lin *et al.* [20] presented a patch-based photo retargeting model which preserves the shapes of both visually salient objects and structural lines. It is worth noticing that, the above content-aware retargeting methods depend merely on low-level feature-based saliency maps, which can hardly reflect visual semantics. Rubinstein *et al.* [33] presented a retargeting algorithm focusing on searching the optimal path in the resizing space. Wang *et al.* [42] introduced a scale-and-stretch warping algorithm that allows resizing images into different aspect ratios while preserving visually prominent features. In [51], Zhang *et al.* proposed a content-aware dynamic video retargeting algorithm. A pixel-level shrinkable map is constructed that indicates both the importance of each pixel and its continuity, based on which a scaling function calculates the new pixel location of the retargeted video. In [14], Krähenbühl *et al.* developed a content-aware interactive video retargeting system. It combines key frame-based constraint editing with numerous automatic algorithms for video analysis.

In recent years, Castillo *et al.* [4] evaluated the impact of photo retargeting on human fixations, by experimenting on the RetargetMe data set [31]. Their work revealed that: 1) even strong artifacts in the retargeted photo cannot influence human gaze shifting if they are distributed outside the regions of interest; 2) removing contents in photo retargeting might change its semantics, which influences human perception of photo aesthetics accordingly; and 3) employing eye-tracking data can more accurately capture the

[1]There are a large body of retargeting-related methods and discussing them enumeratively would be too lengthy (e.g., [10, 11, 17, 21, 28, 32, 34, 37, 45, 48, 49]). Readers can refer to [2, 34, 37] for a more comprehensive survey.

regions of interest, which might be informative for photo retargeting. In [45], Zhang *et al.* proposed a photo retargeting model by learning human gaze allocation, wherein a few salient graphlets are selected based on a sparsity-guided ranking algorithm. Noticeably, the above perception-guided retargeting models may not be applied onto mobile platforms. The reason lies in that there is no exact solution to the sparse ranking algorithm, and the approximate solution might be intolerably time-consuming.

# 3 OUR PROPOSED METHOD

## 3.1 Fast Human Gaze Behavior Modeling

Biological and psychological studies [3, 44] have shown that, in human visual system, only a small fraction of distinctive sensory information is selected for further processing. More specifically, before understanding each real-world scenery, humans will first perceive objects, *i.e.*, selecting possible object locations. Subsequently, human vision system will process only part of an image/video in detail, while leaving the others nearly unprocessed. Apparently, it is important to incorporate such human perception into the retargeting process. Toward a mobile retargeting system, a fast object proposals generation associated with an efficient geometry-preserved graph ranking algorithm is developed for simulating how humans selectively allocating their gazes.

**BING-based fast object patches [6]:** Humans typically attend to those foreground semantic objects, *e.g.*, human/animal faces. Optimally preserving these semantic objects are essential during retargeting, since heavily shrinking them may cause visual distortion. To effectively recognize these semantic objects which may draw human attention, we employ an objectness measure to produce a succinct set of object proposals. During the system design, we believe that an optimal objectness measure should have the following advantages: 1) achieving a high object detection accuracy and ultra-low computational cost; 2) generating a succinct set of object proposals which will facilitate the subsequent salient object patches detection; and 3) exhibiting a good generalization ability to unknown object categories, thereby the model can be flexibly applied onto different data sets.

Taking the above criteria into consideration, we adopt the BING feature proposed by Cheng *et al.* [6] as the objectness measure. The BING feature resizes each image window to $8 \times 8$ and subsequently uses the binarized norm of gradient as its descriptor. It can achieve a high object detection accuracy and maintain an extraordinarily fast speed at the same time.

**Geometry-preserved graph ranking:** We observe that there are still a number of object patches output from [6]. To mimic the actively viewing mechanism of human visual system, an efficient geometry-preserved graph ranking algorithm is proposed for selecting object patches based on their representativeness to a photo/video. These highly representative object patches are more likely to draw human attention, which are sequentially connected to form the gaze shifting path (GSP).

We denote a set of object patches as $\{x_1, \cdots, x_N\} \in \mathbb{R}^{137}$, where each $x_i$ is the 137-D appearance feature (128-D HOG [7] plus 9-D color moment [35]) of the $i$-th object patch. To preserve the geometrical characteristics of a photo/video, we construct a kNN graph $\mathcal{G}$, wherein each vertex represents an object patch and each edge


**Figure 2: Left: preserving all the relative distances between object patches and implicitly maintaining the image/video geometrical characteristics; Right: GSP constructed using the geometry-preserved graph ranking, wherein $M = 5$ top-ranked object patches are selected.**

links pairwise spatially adjacent object patches as shown on the left of Fig. 2. Specifically, the edge weight of graph $\mathcal{G}$ is:

$$\mathbf{W}_{ij} = \exp(-\frac{||x_i - x_j||^2}{\sigma^2}), \tag{1}$$

In our implementation, each object patch is linked with its three nearest neighbors. If pairwise object patches are not connected, we simply set the edge weight to zero. As shown on the left of Fig. 2, preserving all the pairwise distances between object patches during our proposed graph ranking can implicitly maintain the image/video geometrical feature.

Let $\phi : x \rightarrow \mathbb{R}$ be a ranking function which assigns to each object patch $x_i$ a ranking score, we define an initial vector $y = [y_1, \cdots, y_N]^T$, wherein $y_i = 1$ if the $i$-th object patch is salient and $y_i = 0$ otherwise. Based on this, the cost function associated with $\phi$ can be formulated as:

$$f(\phi) = \frac{1}{2}(\sum_{i,j=1}^{N} \mathbf{W}_{ij}||\frac{1}{\mathbf{D}_{ii}}\phi_i - \frac{1}{\mathbf{D}_{jj}}\phi_j||^2 + \mu \sum_{i=1}^{N} ||r_i - y_i||^2), \tag{2}$$

where $\mu > 0$ is the regularization parameter; matrix $\mathbf{D}$ is a diagonal matrix whose $i$-th diagonal element is $\mathbf{D}_{ii} = \sum_{j=1}^{N} \mathbf{W}_{ij}$.

The first term in (2) is a smoothness constraint that enforces the adjacent object patches have similar ranking scores. The second term is a fitting constraint which means that the ranking result should maximally fit the initial label assignment. Notably, the initial labels are assigned according to a well-known fast visual saliency model proposed by Hou *et al.* [12].

By minimizing object function (1), we obtain the optimal $\phi$ using the following closed-form solution:

$$\phi^* = (\mathbf{I}_N - \mathbf{S}/(\mu + 1))y, \tag{3}$$

where $\mathbf{S}$ is the symmetrical normalization of matrix $\mathbf{W}$, *i.e.*, $\mathbf{S} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$; $\mathbf{I}_N$ is an $N \times N$-sized identity matrix.

## 3.2 Deep Network for GSP Representation

By constructing the GSP from each image/video, a deep architecture is formulated to efficiently learn its representation, which is more descriptive than that produced by shallow models. As shown in Fig. 3, the deep architecture contains two key components: 1) deep CNN for representing each object patch, and 2) statistical-aggregation-based GSP representation.

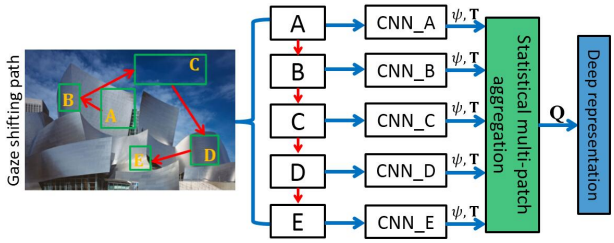


**Figure 3: Structure of our designed deep model, wherein an ordered set of object patches are sequentially aggregated to form the final deep representation.**

First, thorough experimental validations [25] have shown that maintaining the original image resolution and aspect ratio is essential to visual quality modeling. Moreover, arbitrarily-sized objects are more descriptive to aesthetic quality [5]. To this end, we upgrade the conventional five-layer CNN [16] to support arbitrarily-sized inputs. The key technique is an adaptive spatial pooling (ASP) layer whose pooling size can be dynamically adjusted in order to support input patches with various sizes.

Each of the $M$ deep CNNs is detailed as follows. Starting from a large quantity of top-ranked object patches selected by our graph ranking algorithm, we randomly jetter each object patch and flip it horizontally/vertically with probability 0.5 to improve its generality. The network contains four stages of convolution, ASP, and local response normalization, followed by a fully-connected layer with 1024 hidden units. Afterward, the network branches out one fully connected layer containing $H$ 128-D units to describe the corresponding $H$ latent aesthetics-related topics, *e.g.*, "colorful" and "harmony". It is worth emphasizing that, the bottom CNN layers are shared to: 1) decrease the number of parameters, and 2) take advantage of the common low-layer CNN structure.

Second, as shown in Fig. 3, given a GSP which involves multiple sequentially-connected object patches, we extract the $L$-D deep feature for each object patch using the above patch-level deep CNN. Then, these patch-level deep features are statistically aggregated into the deep representation for each GSP.

We denote $\Theta = \{\theta_i\}_{i\in[1,M]}$, where $\theta_i \in \mathbb{R}^L$ is the deep feature corresponding to each of the $M$ object patches from a GSP. Then, we represent $T_k$ as the set of values of the $k$-th component of all $\theta_i \in \Theta$, *i.e.*, $T_l = \{\theta_{lj}\}_{j\in[1,M]}$. The statistical aggregation involves a set of statistical functions: $\Psi = \{\psi_u\}_{u\in[1,U]}$. Each $\psi$ specifies a particular statistical function toward the set of patch-level deep feature output from the $M$ CNNs. Herein, we set $\Psi = \{\min, \max, \text{mean}, \text{median}\}$. The outputs of the functions in $\Psi$ are concatenated and aggregated using a fully-connected layer to generate a $R$-D vector to deeply describe a GSP. The entire flowchart of the above process can be formulated as:

$$f(\Psi) = \mathbf{Q} \times (\oplus_{u=1}^{U} \oplus_{l=1}^{L} \psi_u(T_l)), \tag{4}$$

where $\mathbf{Q} \in \mathbb{R}^{R\times UL}$ represents the parameters of the fully-connected aggregation layer, and $U = 4$ is the number of statistical functions.

**Deep model training:** During the forward propagation, the output $o_i$ of each the $i$-th neuron at the statistical layer can be formulated as $o_i = \sum_{m=1}^{M}\sum_{l=1}^{L} p_{ml\to i} o'_{ml}$, where $p_{ml\to i}$ can be considered as the "contribution" of the neuron $p_{ml}$ to the $i$-th neuron at the statistical layer. Denoting $\eta_i$ as the error propagated to the $i$-th neuron at the statistical layer, the error $\eta'_{ml}$ back-propagated to the neuron $p_{ml}$ is calculated by $\eta'_{ml} = \sum_i p_{ml\to i}\eta_{ml}$.

The overall architecture of our deep model is trained based on the standard back-propagation of the error, associated with a stochastic gradient decent as the loss function, *i.e.*, the sum of the log-loss of each object patch from the training stage.

Time cost of the deep model is briefed as follows. The training stage takes about 17 hours on a desktop PC, wherein object patches from 20,000 well-aesthetic photos are manually selected as the training data. The training is conducted off-line. Comparatively, the test stage is carried out rapidly. It takes nearly 11.435ms and 8.767ms to calculate the deep feature for each GSP, on iPhone 6S and Samsung Galaxy S6 respectively.

## 3.3 Probabilistic Model for Retargeting

Due to the subjectivity of visual aesthetics perception, people with different backgrounds, experiences, and eduction might bias for retargeted photo/video with certain styles. To reduce such bias, it is necessary to exploit the aesthetic experiences of multiple users. Specifically, to make the retargeted photo/video unbiased, we use a probabilistic model to describe the aesthetic experience of professional photographers. As a widely used statistic tool, Gaussian mixture models (GMMs) have been shown to be effective for learning the distribution of a set of data. In our work, GMMs are used to uncover the distribution of GSPs from all training aesthetically pleasing photos. The training photos are collected by googling images using the keywords such as "iPhone wallpaper". For each GSP, we use a 5-component GMM to learn its distribution:

$$p(f|\theta) = \sum_{i=1}^{5} \alpha_i \mathcal{N}(f|\pi_i, \Sigma_i), \tag{5}$$

where $f$ denotes the $R$-D deep feature for each GSP, and $\theta = \{\alpha_i, \pi_i, \Sigma_i\}$ represents the GMM parameters.

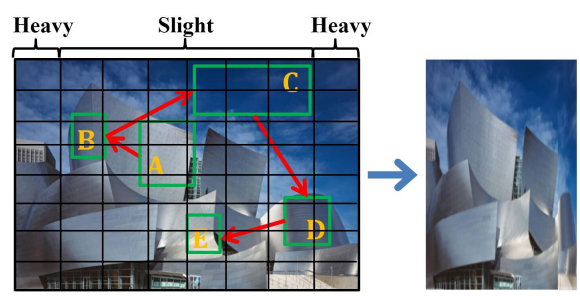


**Figure 4: An example of the grid-based retargeting. The left is the original photo and the right is the retargeted one.**

After learning the GMM priors, we shrink a test photo (or video frame) to make its GSP most similar to those from the training photos. That is, given the GSP of a test photo/video, we calculate the probability of its GSP. To avoid the triangle mesh as the control mesh in shrinking which may result in distortions in triangle orientations, we use grid-based shrinking. Particularly, we decompose a

photo into equal-sized grids (Grid size is a user-tuned parameter and we set it to 20×20 based on cross validation), and the horizontal weight of grid $g$ is calculated as:

$$w_h(g) = p(f|\theta), \text{ if } f \cap g^h \neq \emptyset, \tag{6}$$

where $f \cap g^h \neq \emptyset$ denotes that GSP $f$ is horizontally overlapping with grid $g$.

Similarly, the vertical weight of grid $g$ is calculated as:

$$w_v(g) = p(f|\theta), \text{ if } f \cap g^v \neq \emptyset, \tag{7}$$

For grids not overlapping with a GSP, we set the grid weights a sufficiently low one (0.05 in our work) because these regions will not be attended by human eye. After obtaining the horizontal (resp. vertical) weight of each grid, a normalization operation is carried out to make them sum to one, *i.e.*, $\bar{w}_h(\phi_i) = w_h(\phi_i)/\sum_i w_h(\phi_i)$. Thereafter, given the size of the retargeted photo/video whose size is $W \times H$, the horizontal dimension of the $i$-th grid is shrunk to $[W \cdot \bar{w}_h(\phi_i)]$, and the vertical one of the $i$-th grid is shrunk to $[H \cdot \bar{w}_v(\phi_i)]$, where $[\cdot]$ rounds a real number to the nearest integer. The above retargeting process can be elaborated in Fig. 4. Grids covered by the central architecture are semantically significant, and are thus preserved in the retargeted photo with slight scaling. In contrast, grids covered by the surrounding architecture are less semantically important, thereby they are heavily shrunk in both horizontal and vertical directions. Notably, the above steps are for photo retargeting. For video retargeting, we follow the operations in [46], where the shrinking weight of the current frame is utilized to guide the shrinkage of the next frame.

The time consumption of the above probabilistic retargeting model is as follows. The GMM training is moderately time-consuming due to the iterative EM algorithm (*i.e.*, about 130s on iPhone 6S and Galaxy S6 respectively). Comparatively, the grid-based shrinking is conducted very fast (about 2.321ms and 2.431 per image on iPhone 6S and Galaxy S6 respectively). Fortunately, the GMM training is usually conducted off-line, thereby the probabilistic retargeting is real time on mobile platforms.

Based on our discussions from Sec 3.1 to Sec 3.3, the proposed photo/video retargeting on mobile platforms can be summarized in Algorithm 1.

**Algorithm 1** Perceptual Retargeting on Mobile Platforms

**input**: $N$ well-aesthetic photos from multiple professional photographers, parameters: $\mu$, $M$, and the test photo/video;
**output**: Retargeted photo/video;

1) Extract a set of object patches using the BING feature [6], then utilize the geometry-preserved ranking to construct GSP;
2) Calculate the deep representation of each GSP based on our aggregation-based deep model;
3) Retarget each photo/video using the grid-based probabilistic model as shown in (5).

# 4 EXPERIMENTS AND ANALYSIS

All the baseline retargeting models were implemented based on the C++. Except for our method, all the baseline retargeting algorithms are experimented on the workstation HP Z840, which is equipped with a dual Intel E5-2600 CPU, 32GB RAM, 256GB SSD, and HP Z24X LED monitor. For our mobile retargeting algorithm, we implement two versions on both iOS 10.1 and Android 6.0.1 platforms respectively. Two popular mobile devices, iPhone 6S and Samsung Galaxy S6, are employed for experiments.

## 4.1 Comparative Study

**Photo retargeting evaluation:** We compare our retargeting method against several representative approaches in the state-of-the-art, including three cropping methods: omni-range context-based cropping (OCBC) [5], probabilistic graphlet-based cropping (PGC) [50], describable attribute for photo cropping (DAPC) [8], as well as four content-aware retargeting methods: seam carving (SC) [1] and its improved version (ISC) [32], optimized scale-and-sketch (OSS) [42], and saliency-based mesh parametrization (SMP) [10]. We experiment on the standard retargeting image set, RetargetMe [31]. The resolution of the resulting photos is fixed to: 640 × 960.

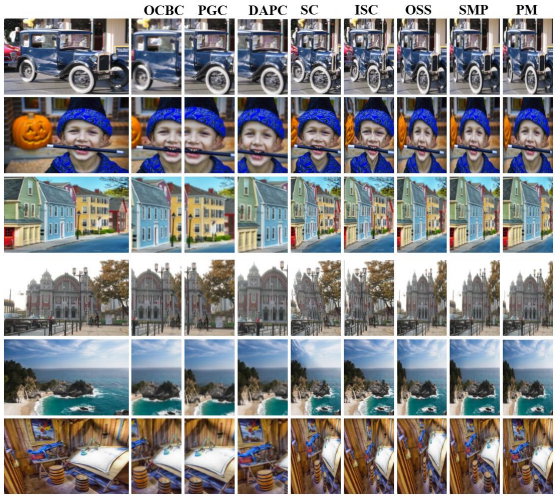


**Figure 5: Comparison of our approach with well-known photo retargeting methods (PM: our proposed method)**

In order to make the evaluation comprehensive, we adopt a paired-comparison-based user study to evaluate the effectiveness of the proposed retargeting algorithm. This strategy was also used in [50] to evaluate the quality of a cropped photo. In the paired comparison, each subject is presented with a pair of retargeted photos from two different approaches, and is required to indicate a preference as of which one they would choose for a phone wallpaper. The participants are 35 ∼ 45 amateur/professional photographers.

As the comparative results shown in Fig. 5, we made the following observations. First, compared with the three content-aware retargeting methods, our approach preserves the semantically important objects in the original photo well, such as the barrels from the first photo, the wheels from the second vehicle wheels. In contrast, the compared retargeting methods may shrink the semantically important objects, such as the vehicles wheels and human face. Even worse, SC and its variant ISC, as well as OSS may result in visual distortions, *i.e.*, the human faces and drawing papers. Moreover, only our retargeting method well preserves the spatial composition of the orginal photo. For example, in the last photo, the left barrel is larger than the right one. For photos retargeted by different methods, only our method accurately captures this

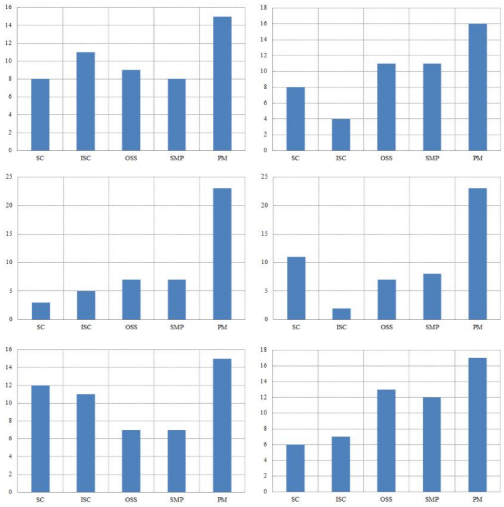


**Figure 6: Statistics of user study from the six sets of retargeted photos in Fig. 5 (the vertical axes denote the user votes for each retargeting method)**

clue. Second, although cropping methods preserve important regions without visual distortions, they abandon regions that are less visually salient but still capture the global spatial layout. Specifically, the vehicle from the first photo, the entire human face and church from the second and fourth photo, the left door from the last photo. Third, as the statistical results displayed in Fig. 6, user study demonstrates that our method outperforms its competitors on the resulting photos. It is noticeable that, when the resulting photos appear without distortion, the content-aware retargeting outperforms the cropping technique, and vice versa. On all the six photos, our approach produces non-distorted photos and the semantically significant objects are nicely preserved. Therefore, the best resulting photos are consistently achieved by our approach.

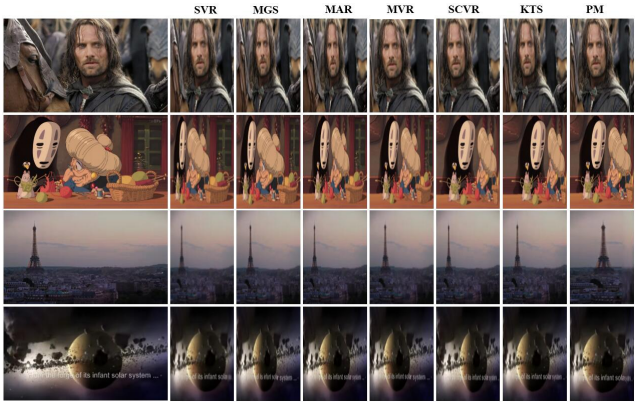


**Figure 7: Comparative video retargeting results**

**Video retargeting evaluation:** We select six representative algorithms as the baseline for testifying the video retargeting performance. They are streaming video retargeting (SVR) [15], mosaic-guided scaling (MGS) [48], motion-aware video retargeting (MAR) [39], motion-based video retargeting (MVR) [41], scalable and coherent video resizing (SCVR) [40], and key-frames using grid flows (KTS) [17]. We crawl nearly 500 video clips from Youtube, wherein the resolution is fixed to $1280 \times 720$. These videos contain semantic contents from eight categories (*i.e.*, "human face", "architecture", "landscape", "vehicle", "boat", "park", "pedestrian", and "river") and each lasts from 46s to 98s. As the qualitative results shown in Fig. 7, our method can best preserve the foreground salient objects. And no obvious visual distortions are observed in retargeted videos produced by us. To quantitatively compare the retargeting results in Fig. 7, we follow the paired-comparison-based user study above. As shown in Fig. 8, users consistently consider that videos retargeted by our method is the most aesthetically-pleasing.

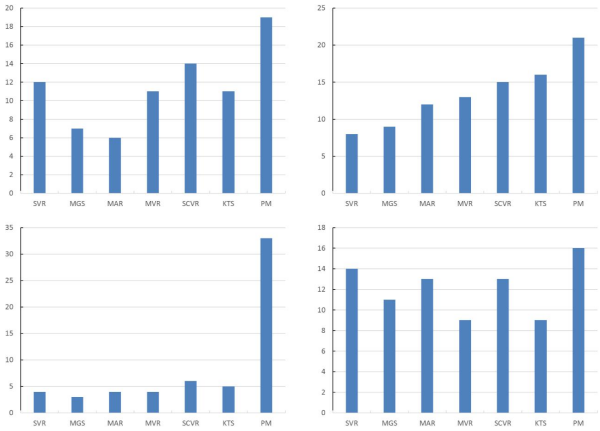


**Figure 8: Statistics of user study from the four sets of retargeted videos as displayed in Fig. 7**

**Time consumption analysis:** In retrospect, our retargeting framework contains three key components, fast GSP construction, deep network for GSP representation, and probabilistic model for retargeting. For photo retargeting, time consumptions of the three steps are 13.212ms, 11.435ms, and 12.114ms respectively on the iOS platform. On the Android platform, it takes 11.212ms, 8.767ms, and 11.231ms to conduct the three steps respectively. Totally, it consumes about 30ms to retarget each photo, which is sufficiently fast. Comparatively, even on the desktop platform, time consumptions of the baseline photo retargeting algorithms are: 2.432s (SC), 3.332s (ISC), 32.321s (OSS), and 13.211s (SMP) respectively.

For video retargeting, on the iOS platform, time consumptions of the three steps are 0.231s, 0.321s, and 0.123s respectively when retargeting each 1s video clip. On the Android platform, time costs of the three operations are 0.254s, 0.221s, and 0.165s when retargeting each 1s video clip. That is to say, our retargeting method is real-time on mobile platforms. Contrastively, it takes nearly ten seconds to retarget each 1s video for the other methods.

## 4.2 Parameter Analysis

This experiment reports the influences of important parameters on retargeting a specific photo. Totally, there are four key parameters in our approach: 1) $\mu$, the regularization parameter, 2) $M$, the number of object patches within each GSP, 3) $L$ and $R$, the dimensions

of patch-level and image-level deep features respectively, and 4) the grid size for probabilistic retargeting. The default values of the above parameters are: $\mu = 0.2$, $K = 5$, $L = 256$, $R = 256$, and $GridSize = 20$.

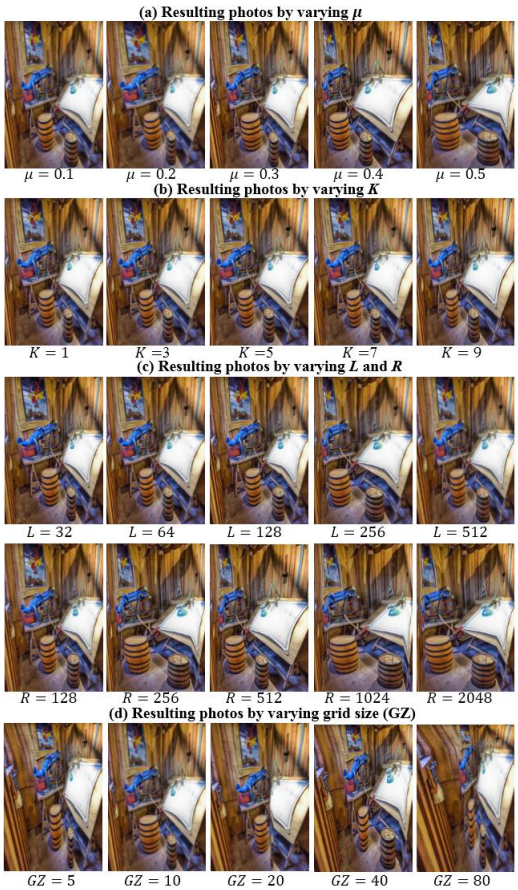


**Figure 9: Retargeted photos produced under different parameter settings**

Retargeting results under different parameter settings are shown in Fig. 9. First, we tune the value of $\mu$ and observe that the most aesthetically-pleasing retargeted photo is achieved when $\mu = 0.3$. This might because a larger $\mu$ will enforce too much on the locality-preserving attribute, which will make the foreground object too large. Even worse, slight visual distortion is observed when $\mu = 0.5$. Second, we present the retargeted photos when different numbers of object patches are selected for GSP construction. As seen, by increasing the number of selected object patches $M$ from one to five, the semantically significant objects, such as the barrels and the drawing board, are better preserved in the retargeted photo. When $M$ is larger than five, however, the resulting photo remains almost unchanged. Thereby, we set $M = 5$ for this photo. Third, we retarget a photo using different dimensional patch-level and image-level deep features, *i.e.*, $L$ and $R$. We observe that a larger $L$ will make more semantically important regions retained in the retargeted photo. But emphasizing the foreground objects too much might not be a good choice and may decrease the global composition, *e.g.*, $L = 256$ or 512. Similarly, a too large $R$ will also inappropriately emphasize the foreground objects. In this way, we set $R = 512$. Finally, we change the grid size and display the corresponding retargeted photo. As can be seen, when the gird size is set to $5 \times 5$ and $10 \times 10$ respectively, the resulting photos are both distorted. When the grid size is larger than $20 \times 20$, the distortion disappears but the left barrel becomes disharmonically large. Therefore, we set the gird size to $20 \times 20$.

## 4.3 GSP Evaluation using Eye Tracker

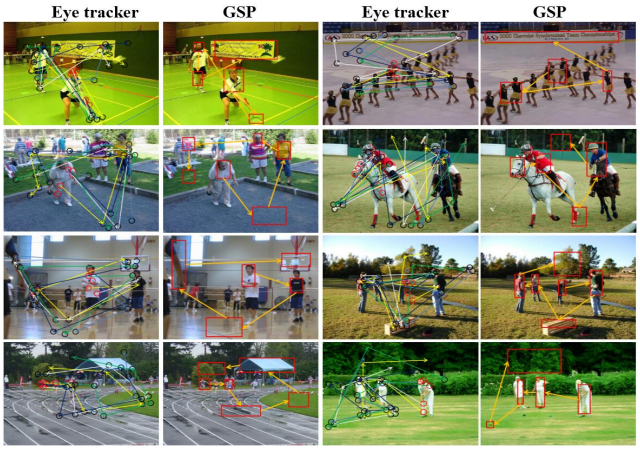


**Figure 10: Comparison of gaze shifting paths from five observers (differently colored) and our calculated GSPs**

In this subsection, we quantitatively and qualitatively compare the calculated GSPs with real human gaze shifting paths. More specifically, we record the eye fixations from five observers by leveraging the eye-tracker EyeLink II[2], and then link the fixations into a path in a sequential manner. As shown in Fig. 10, for most scene images, our calculated GSPs are consistent with the real human gaze shifting paths. Moreover, we calculate the percentage of the human gaze shifting path which overlaps with our calculated GSPs. In detail, given each of the five real human gaze shifting paths, we connect all the segmented regions along the path and then obtain the human gaze shifting path with the segmented regions. Thereafter, the similarity between a GSP and a real human gaze shifting path is measured as follows:

$$s(P_1, P_2) = \frac{\mathcal{N}(P_1 \cap P_2)}{\mathcal{N}(P_1) + \mathcal{N}(P_2)}, \tag{8}$$

where $P_1$ and $P_2$ denote a calculated GSP and the real human gaze shifting path with segmented regions respectively, $\mathcal{N}$ counts the pixels inside each image region, and $P_1 \cap P_2$ denotes the shared regions between $P_1$ and $P_2$. According to (8), we observe that the overlapping percentage between our calculated GSPs and real human gaze shifting paths is 89.321% on average. This result shows that our predicted paths can effectively capture the real human gaze shifting process.

[2] www.sr-research.com/

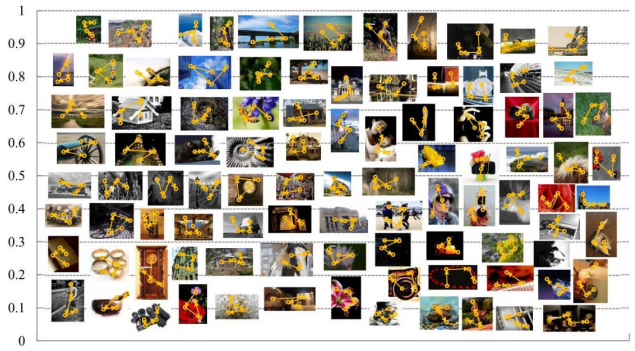


**Figure 11: Visualized GSPs from a set of AVA images [26]. The yellow paths denote the GSPs predicted by our method, where each circle indicates the location of a region.**

Additionally, we visualize GSPs calculated from AVA scene images [26]. AVA contains a large number of images with their quality scores. As shown in Fig. 11, the following observations can be made. First, as shown in the photos whose quality levels ranged from 0.8 and 1, the high quality scene pictures with multiple interacting objects are assigned with very high scores, which shows that our calcualted GSPs can well predict how humans perceive local/global composition in these beautiful pictures. Second, as shown in images whose quality levels are between 0.5 and 0.8, the high quality pictures with a single object are also appreciated by the proposed methods. This is because our graph ranking algorithm can naturally reveal local scene composition. Third, the objects from photos whose quality levels are ranked between 0 and 0.5 are either spatially disharmoniously distributed or blurred. Therefore, they are considered as low quality by our model.

Last but not least, we analyze the GSPs extracted from both low quality and high quality scene images. As can be seen from Fig. 11, neither low nor high quality scene images have a particular path geometry, *e.g.*, the angle between pairwise shifting vectors (yellow arrows). It is worth emphasizing that, for high quality scene pictures, the fixation points (yellow circles) are aesthetically pleasing and the objects along the path are harmoniously distributed.

## 4.4 Quality Prediction Evaluation

The key of our probabilistic retargeting model is a quality measure which discovers the most beautiful candidate retargeted photo. The first experiment compares our approach with a series of shallow/deep media quality methods. The shallow models include three global feature-based approaches proposed by Dhar *et al.* [8], Luo *et al.* [24], and Marchesotti *et al.* [27], respectively; as well as two local patch integration-based methods proposed by Cheng *et al.* [5] and Nishiyama *et al.* [29], respectively. At the same time, three deep quality models proposed by Lu *et al.* [22, 23] and Mai *et al.* [25] are also testified. In the comparative study, we notice that the source codes of the five shallow quality models are not provided and some experimental details are not mentioned, therefore it is difficult to strictly implement them. We thus adopt the following implementation settings. For Dhar's approach, we use the public codes from Li *et al.* [18] to extract the attributes from each photo.

**Table 1: Comparison of quality prediction performance**

| | Models | CUHK | PNE | AVA | LIVE-IQ |
|---|---|---|---|---|---|
| Shallow | Dhar *et al.* | 0.7386 | 0.6754 | 0.6435 | 0.8943 |
| | Luo*et al.* | 0.8004 | 0.7213 | 0.6879 | 0.8854 |
| | Marchesotti*et al.* | 0.8767 | 0.8114 | 0.7891 | 0.8784 |
| | Cheng *et al.* | 0.8432 | 0.7754 | 0.8121 | 0.9021 |
| | Nishiyama *et al.* | 0.7745 | 0.7341 | 0.7659 | 0.8657 |
| Deep | Lu *et al.* [22] | 0.9154 | 0.8034 | 0.7446 | 0.8832 |
| | Lu *et al.* [23] | 0.9237 | 0.8034 | 0.7446 | 0.9023 |
| | Mai *et al.* | 0.9276 | 0.8432 | 0.7710 | 0.8943 |
| | Ours | **0.9321** | **0.8676** | **0.8256** | **0.9312** |

These attributes are combined with the low-level features proposed by Yeh *et al.* [47] to train the classifier. For Luo *et al.*'s approach, not only the low-level and high-level features in their publication are implemented, but also the six global features from Getlter *et al.* [9]'s work are used to strengthen the aesthetic prediction ability. For Marchesotti *et al.*'s approach, similar to the implementation of Luo *et al.*'s method, the six additional features are also adopted. For Cheng *et al.*'s approach, we implement it as a simplified version of our approach, *i.e.*, only 2-sized graphlets are employed for aesthetics measure. Notably, for the three probabilistic model-based quality (*i.e.*, Cheng *et al.*'s, Nishiyama *et al.*'s, and our method), if the quality score is larger than 0.5, then this image/video is deemed as high quality, and vice versa. For the three deep quality models, we notice that source codes of Lu *et al.* [22, 23]'s approaches are unavailable. Thereby, we implemented them by ourselves. According to their publications, some detailed experimental configurations, such as the CDUA-Convnet, are missing. Therefore, we carefully tune the parameters until the performance on the AVA [26] is close to that reported publicly.

We report the quality prediction accuracies on the CUHK, PNE, AVA, and the LIVE-IQ in Table 1. On the four data sets, our approach outperforms its competitors remarkably, which demonstrates the advantages of our quality prediction. First, accurately modeling human gaze shifting is informative in predicting media quality, since human visual perception can be well encoded. Second, deep models have remarkable advantage over shallow models in image/video quality modeling. Noticeably, the previous deep quality models based on the entire images or randomly-cropped image patches might be less effective. Discovering visually/semantically salient object patches for deep quality model training can receive a significant performance gain.

## 5 CONCLUSIONS

In this work, a mobile platform is designed which effectively retargets photos/videos by deeply encoding human gaze behavior. More specifically, given a set of photos or video clips, we first construct their GSPs based on the fast graph ranking. Afterward, a deep architecture is proposed which converts each GSP into its deep representation using an aggregation scheme. Finally, these deep GSP features are integrated through a probabilistic model for photo/video retargeting. Comprehensive experimental results on both the iOS and Android devices have demonstrated the efficiency and effectiveness of our method.

## 6 ACKNOWLEDGMENT

Prof. Yingjie Xia (the 5th author) is the correspondence author.

## REFERENCES


[1] Shai Avidan and Ariel Shamir. 2007. Seam carving for content-aware image resizing. *ACM Trans. Graph.* 26, 3 (2007), 10.

[2] Francesco Banterle, Alessandro Artusi, Tunc O. Aydin, Piotr Didyk, Elmar Eisemann, Diego Gutierrez, Rafal Mantiuk, and Karol Myszkowski. 2011. Spatial Image Retargeting. In Multidimensional Image Retargeting. In *SIGGRAPH Asia Courses*.

[3] Neil D. B. Bruce and John K. Tsotsos. 2009. Saliency, Attention, and Visual Search: An Information Theoretic Approach. *Journal of Vision* 9, 3 (2009), 5.1–24.

[4] Susana Castillo, Tilke Judd, and Diego Gutierrez. 2011. Using eye-tracking to assess different image retargeting methods. In *Proc. of APGV*. 7–14.

[5] Bin Cheng, Bingbing Ni, Shuicheng Yan, and Qi Tian. 2010. Learning to photograph. In *ACM Multimedia*. 291–300.

[6] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip H. S. Torr. 2014. BING: Binarized Normed Gradients for Objectness Estimation at 300fps. In *Proc. of CVPR*. 3286–3293.

[7] Navneet Dalal and Bill Triggs. 2005. Histograms of Oriented Gradients for Human Detection. In *Proc. of CVPR*. 886–893.

[8] Sagnik Dhar, Vicente Ordonez, and Tamara L. Berg. 2011. High level describable attributes for predicting aesthetics and interestingness. In *Proc. of CVPR*. 1657–1664.

[9] Peter V. Gehler and Sebastian Nowozin. 2009. On feature combination for multiclass object classification. In *Proceedings of ICCV*. 221–228.

[10] Yanwen Guo, Feng Liu, Jian Shi, ZhiHua Zhou, and Michael Gleicher. 2009. Image Retargeting Using Mesh Parametrization. *IEEE Trans. Multimedia* 11, 5 (2009), 856–867.

[11] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of WWW*. 173–182.

[12] Xiaodi Hou, Jonathan Harel, and Christof Koch. 2012. Image Signature: Highlighting Sparse Salient Regions. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 1 (2012), 194–201.

[13] Johannes Kiess, Daniel Gritzner, Benjamin Guthier, Stephan Kopf, and Wolfgang Effelsberg. 2014. GPU video retargeting with parallelized SeamCrop. In *ACM MMSYS*. 139–147.

[14] Philipp Krähenbühl, Manuel Lang, Alexander Hornung, and Markus H. Gross. 2009. A system for retargeting of streaming video. *ACM Trans. Graph.* 28, 5 (2009), 126:1–126:10.

[15] Philipp Krähenbühl, Manuel Lang, Alexander Hornung, and Markus H. Gross. 2009. A system for retargeting of streaming video. *ACM Trans. Graph.* 28, 5 (2009), 126:1–126:10.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proc. of NIPS*. 1106–1114.

[17] Bing Li, Ling-Yu Duan, Jinqiao Wang, Rongrong Ji, Chia-Wen Lin, and Wen Gao. 2014. Spatiotemporal Grid Flow for Video Retargeting. *IEEE Trans. Image Processing* 23, 4 (2014), 1615–1628.

[18] Fei-Fei Li and Pietro Perona. 2005. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In *Proceedings of CVPR*. 524–531.

[19] Li-Jia Li, Hao Su, Eric P. Xing, and Fei-Fei Li. 2010. Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification. In *Proc. of NIPS*. 1378–1386.

[20] Shih-Syun Lin, I-Cheng Yeh, Chao-Hung Lin, and Tong-Yee Lee. 2013. Patch-Based Image Warping for Content-Aware Retargeting. *IEEE Trans. Multimedia* 15, 2 (2013), 359–368.

[21] Anan Liu, Yuting Su, Weizhi Nie, and Mohan S. Kankanhalli. 2017. Hierarchical Clustering Multi-Task Learning for Joint Human Action Grouping and Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1 (2017), 102–114.

[22] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Zijun Wang. 2014. RAPID: Rating Pictorial Aesthetics using Deep Learning. In *ACM MM*. 457–466.

[23] Xin Lu, Zhe Lin, Xiaohui Shen, Radomír Mech, and James Zijun Wang. 2015. Deep Multi-patch Aggregation Network for Image Style, Aesthetics, and Quality Estimation. In *Proc. of ICCV*. 990–998.

[24] Wei Luo, Xiaogang Wang, and Xiaoou Tang. 2011. Content-based photo quality assessment. In *Proceedings of ICCV*. 2206–2213.

[25] Long Mai, Hailin Jin, and Feng Liu. 2016. Composition-Preserving Deep Photo Aesthetics Assessment. In *Proc. of CVPR*. 497–506.

[26] Luca Marchesotti, Naila Murray, and Florent Perronnin. 2014. Discovering beautiful attributes for aesthetic image analysis. *IJCV* (2014). DOI:https://doi.org/10.1007/s11263-014-0789-2

[27] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka. 2011. Assessing the aesthetic quality of photographs using generic image descriptors. In *Proceedings of ICCV*. 1784–1791.

[28] Liqiang Nie, Shuicheng Yan, Meng Wang, Richang Hong, and Tat-Seng Chua. 2012. Harvesting visual concepts for image search with complex queries. In *ACM Multimedia*. 59–68.

[29] Masashi Nishiyama, Takahiro Okabe, Imari Sato, and Yoichi Sato. 2011. Aesthetic quality classification of photographs based on color harmony. In *Proceedings of CVPR*. 33–40.

[30] Yael Pritch, Eitam Kav-Venaki, and Shmuel Peleg. 2009. Shift-map image editing. In *Proc. of ICCV*. 151–158.

[31] Michael Rubinstein, Diego Gutierrez, Olga Sorkine, and Ariel Shamir. 2010. A comparative study of image retargeting. *ACM Trans. Graph.* 29, 6 (2010), 160:1–160:10.

[32] Michael Rubinstein, Ariel Shamir, and Shai Avidan. 2008. Improved seam carving for video retargeting. *ACM Trans. Graph.* 27, 3 (2008), 16:1–16:9.

[33] Michael Rubinstein, Ariel Shamir, and Shai Avidan. 2009. Multi-operator media retargeting. *ACM Trans. Graph.* 28, 3 (2009), 23:1–23:11.

[34] Ariel Shamir, Alexander Sorkine-Hornung, and Olga Sorkine-Hornung. 2012. Modern Approaches to Media Retargeting. In *SIGGRAPH Asia Courses*.

[35] Markus A. Stricker and Markus Orengo. 1995. Similarity of Color Images. In *Storage and Retrieval for Image and Video Databases*. 381–392.

[36] Jin Sun and Haibin Ling. 2013. Scale and Object Aware Image Thumbnailing. *International Journal of Computer Vision* 104, 2 (2013), 135–153.

[37] Daniel Vaquero, Matthew Turk, Kari Pulli, Marius Tico, and Natasha Gelfand. 2010. A Survey of Image Retargeting Techniques. In *Proc. of SPIE*.

[38] Manuela Vasconcelos, Nuno Vasconcelos, and Gustavo Carneiro. 2006. Weakly Supervised Top-down Image Segmentation. In *Proc. of CVPR*. 1001–1006.

[39] Yu-Shuen Wang, Hongbo Fu, Olga Sorkine, Tong-Yee Lee, and Hans-Peter Seidel. 2009. Motion-aware temporal coherence for video resizing. *ACM Trans. Graph.* 28, 5 (2009), 127:1–127:10.

[40] Yu-Shuen Wang, Jen-Hung Hsiao, Olga Sorkine, and Tong-Yee Lee. 2011. Scalable and coherent video resizing with per-frame optimization. *ACM Trans. Graph.* 30, 4 (2011), 88:1–88:8.

[41] Yu-Shuen Wang, Hui-Chih Lin, Olga Sorkine, and Tong-Yee Lee. 2010. Motion-based video retargeting with optimized crop-and-warp. *ACM Trans. Graph.* 29, 4 (2010), 90:1–90:9.

[42] Yu-Shuen Wang, Chiew-Lan Tai, Olga Sorkine, and Tong-Yee Lee. 2008. Optimized scale-and-stretch for image resizing. *ACM Trans. Graph.* 27, 5 (2008), 118:1–118:8.

[43] Lior Wolf, Moshe Guttmann, and Daniel Cohen-Or. 2007. Non-homogeneous Content-driven Video-retargeting. In *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*. 1–6.

[44] Jeremy M. Wolfe and Todd S. Horowitz. 2004. What Attributes Guide the Deployment of Visual Attention and How Do They Do It? *Nature Reviews Neuroscience* 5, 6 (2004), 495–501.

[45] Yingjie Xia, Luming Zhang, Richang Hong, Liqiang Nie, Yan Yan, and Ling Shao. 2017. Perceptually Guided Photo Retargeting. *IEEE Trans. Cybernetics* 47, 3 (2017), 566–578.

[46] Bo Yan, Kairan Sun, and Liu Liu. 2013. Matching-Area-Based Seam Carving for Video Retargeting. *IEEE Trans. Circuits Syst. Video Techn.* 23, 2 (2013), 302–310.

[47] Che-Hua Yeh, Yuan-Chen Ho, Brian A. Barsky, and Ming Ouhyoung. 2010. Personalized photograph ranking and selection system. In *Proceedings of ACM Multimedia*. 211–220.

[48] Tzu-Chieh Yen, Chia-Ming Tsai, and Chia-Wen Lin. 2011. Maintaining Temporal Coherence in Video Retargeting Using Mosaic-Guided Scaling. *IEEE Trans. Image Processing* 20, 8 (2011), 2339–2351.

[49] Hanwang Zhang, Zheng-Jun Zha, Yang Yang, Shuicheng Yan, Yue Gao, and Tat-Seng Chua. 2013. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In *ACM Multimedia*. 33–42.

[50] Luming Zhang, Mingli Song, Qi Zhao, Xiao Liu, Jiajun Bu, and Chun Chen. 2013. Probabilistic Graphlet Transfer for Photo Cropping. *IEEE Trans. Image Processing* 22, 2 (2013), 802–815.

[51] Yi-Fei Zhang, Shi-Min Hu, and Ralph R. Martin. 2008. Shrinkability Maps for Content-Aware Video Resizing. *Comput. Graph. Forum* 27, 7 (2008), 1797–1804.