Photo Squarization by Deep Multi-Operator Retargeting

Yu Song
NLPR, Institute of Automation, Chinese Academy of Sciences & University of Chinese Academy of Sciences
songyu2017@ia.ac.cn

Fan Tang*
NLPR, Institute of Automation, Chinese Academy of Sciences & University of Chinese Academy of Sciences
tangfan2013@ia.ac.cn

Weiming Dong†
NLPR, Institute of Automation, Chinese Academy of Sciences
weiming.dong@ia.ac.cn

Xiaopeng Zhang
NLPR, Institute of Automation, Chinese Academy of Sciences
xiaopeng.zhang@ia.ac.cn

Oliver Deussen
VCC SIAT Shenzhen & University of Konstanz
oliver.deussen@uni-konstanz.de

Tong-Yee Lee
National Cheng-Kung University
tonylee@mail.ncku.edu.tw

ABSTRACT

Squared forms of photos are widely used in social media as album covers or thumbnails of image streams. In this study, we realize photo squarization by modeling *Retargeting Visual Perception Issues*, which reflect human perception preference toward image ratargeting. General image retargeting techniques deal with three common issues, namely, *salient content*, *object shape*, and *scene composition*, to preserve the important information of original image. We propose a new way based on multi-operator techniques to investigate human behavior in balancing the three issues. We establish a new dataset and observe human behavior by inviting investigators to retarget images to square manually. We propose a data-driven approach composed of perception and distillation modules by using deep learning techniques to predict human perception preference. The perception part learns the relations among the three issues, and the distillation part transfers the learned relations to a simple but effective network. Our study contributes to deep learning literature by optimizing a network index and lightening its running burden. Experimental results show that photo squarization results generated by the proposed model are consistent with human visual perception results.

CCS CONCEPTS

• **Computing methodologies** → **Image processing**; *Perception*;

KEYWORDS

Image retargeting; photo squarization; retargeting visual perception issues; distillation

*Yu Song and Fan Tang contributed equally to this work.
†Corresponding author

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240623

ACM Reference Format:
Yu Song, Fan Tang, Weiming Dong, Xiaopeng Zhang, Oliver Deussen, and Tong-Yee Lee. 2018. Photo Squarization by Deep Multi-Operator Retargeting. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240623

1 INTRODUCTION

At present, people like to share their opinions, insights, and experiences on the Internet which not only can act as a record but also can send messages to the audience from all over the world. Social media, such as Instagram, Flickr, and Facebook, provide tools for users to display photos in addition to textual information. However, in many cases, the photos need to be displayed in a fixed resolution, that is, in square shape. As shown in Figure 1, a normal web page on Flickr ("Albums" page) contains dozens of photos, which are essentially square thumbnails of photos with different aspect ratios. Therefore, each square thumbnail should display the most prominent information present in the original photo. The standard operation used by most social media to perform photo squarization is cropping. Most methods use a saliency map or an object detector to identify regions in the image that can serve as effective crops in creating thumbnails [7, 27, 35]. Unfortunately, some important content may have to be discarded due to space limitation and the composition of the original photo may be destroyed.

Content-aware image retargeting (CAIR) is a possible choice to solve the photo squarization problem in social media. Various methods have been suggested for CAIR to preserve important content as much as possible [3, 18, 24]. However, most methods are developed by following one single scheme, thereby leading to adaptability problem on general displaying use for social media. Multi-operator methods [8, 25, 31] have been proposed to solve this issue, in which different mechanisms (e.g., seam carving, scaling, and cropping) are integrated to accommodate the variation in contents and compositions of images. Each operator is related to a specific visual perception-based information preserving issue (e.g., salient content, object shape, and scene composition). However, these methods allocate the utilization percentages of operators only in accordance with image similarity and fail to consider user preferences due to lack of user interaction as input; this factor is important in practical

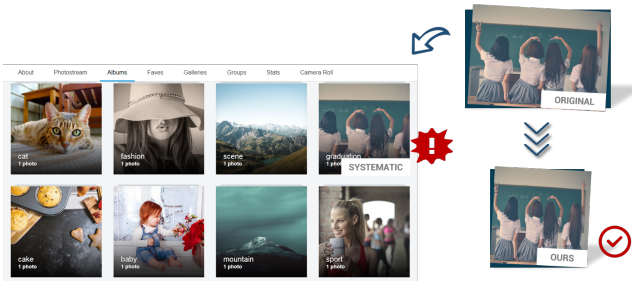


Figure 1: "Albums" page of a Flickr member. All album cover images are in square shape with a certain degree of information loss. We propose a CNN-based photo squarization method for transforming photos of arbitrary aspect ratios to square. This method can capture user's visual preference during retargeting operations.

photography [13, 23]. This drawback limits the practical application of these methods.

One the one hand, previous studies have shown that learning-based approach for image synthesis is effective in reaching the desired outcomes with low investment on time and effort during acquisition [7, 28, 37]. On the other hand, multi-operator image retargeting has been shown to capture user preference effectively [4]. Although the two techniques have shown success, methods that tightly integrate them for learning-based multi-operator image reshaping have yet to be explored.

Square shape is broadly used in social media for photo display. Hence, we focus on the photo squarization problem that has a great research value in the image retargeting field. Two major challenges are acquiring user preference information for retargeting a photo to square and predicting squarization strategy that is consistent with human visual perception. The key idea behind user preference acquisition is to understand human behavior in balancing three issues during retargeting operations: *salient content*, *object shape*, and *scene composition*. We call the three issues as *Retargeting Visual Perception Issues* (RVPIs). First, these issues are the most important information that needs to be preserved during retargeting process. Second, the positions and shapes of visual elements in different photos often vary considerably. Thus, we need to accurately transform the abstract information losses related to the three issues into concrete image retargeting operations. Moreover, the visual perception-based retargeting strategy for squarization should be predicted directly from the visual features of the original photo. However, existing CAIR methods either use iterative frameworks [8, 30] or rely on single operator and thus cannot capture different perception preferences of users [3, 7, 19].

We attempt to engage user preference into a multi-operator retargeting algorithm for overcoming the above-mentioned challenges. In this way, human behavior in balancing RVPIs for a given image and desired size can be determined. This problem is central to social media and image retargeting research. We build a synergy between visual images and human perception by transforming abstract losses into concrete image retargeting operations (Section 3). We propose a data-driven approach to predict people's perception preference by suggesting a CNN-based framework that involves two different modules: *perception* module that mines the relationship among RVPIs and *distillation* module that encodes such knowledge together with variant losses into a simple but effective network (Section 4). Convincing visual and quantitative experimental results are shown to demonstrate the effectiveness of the learned method (Section 5).

Our work makes the following technical contributions:

(1) A new way to observe human perception preferences among RVPIs during photo squarization process, which can also be attempted for general image retargeting problems.
(2) A new dataset with annotations that indicates user preferences toward photo squarization problem.
(3) A new two-module deep neural network for learning user preferences and predicting retargeting operations for photo squarization. To the best of our knowledge, we are the first to allocate the final percentages of operators directly in accordance with the features of original photo without iteration and input-output similarity comparison.

2 RELATED WORK

Image Reshaping. Transforming an image to a new shape is an interesting topic in image synthesis. Gal et al. [10] used a feature map to roughly mask the important features of an image and then performed non-homogeneous texture mapping to transform the image into arbitrary shape. Li et al. [17] proposed a geodesic-preserving method to transform panorama to rectangle. Qi et al. [22] reshaped an image into non-rectangular shape by removing a sequence of seam segments that does not considerably alter or distort the image content. In current study, we focus on reshaping a photo of arbitrary aspect ratio to square by CAIR.

Image Retargeting. Image retargeting preserves the important content of an image after resizing. Cropping [7, 27, 35] is a simple retargeting method that removes outer parts of an image while protecting the subject and edge continuity. This method will not change any area of the original image and does not result in distortion. Thus, cropping has been widely used in social media to generate square thumbnails or thumbnails of other shapes. However, cropping often destroys the completeness of the objects and causes unexpected loss of information. In recent years, CAIR technology has been extensively investigated. The methods can roughly be categorized into discrete and continuous retargeting. Discrete methods change the aspect ratio of an image by repeatedly removing or inserting pixels or patches at unimportant areas. Avidan et al. [2] introduced the concept of seam carving and solved it using dynamic programming, in which a gradient energy was used as the importance map. Rubinstein et al. [24] improved seam carving by using a forward energy. Pritch et al. [21] performed a discrete labeling over individual pixels and retargeted an image by removing segments in the net. These approaches are effective at retargeting images with rich texture content but may cause artifacts of local discontinuity. Continuous methods [14, 16, 32] focus on preserving local structure and optimize a warping from the source size to the target size in accordance with its important regions and permissible deformation. Panozzo et al. [20] minimized warping energy in the space of axis-aligned deformations to avoid unnatural distortions. Lin et al. [18] presented a patch-based scheme with

an extended significance measurement to preserve shapes of visually salient objects and structural lines. These approaches can preserve the geometric structure of image content smoothly but may also permit unwanted low important regions to appear in the retargeting result. Multi-operator methods [4–6, 31] fuse discrete and continuous methods into a unified optimization framework. Rubinstein et al. [25] defined a retargeting space as a conceptual multi-dimensional space in combination with several operators and used bi-directional warping with dynamic programming to find an optimal path in this space. Wang et al. [29] exploited complementary relationships among three condensation operators and fused them into a unified grid-based convex programming problem. Fang et al. [8] constructed the retargeting operator sequence by evaluating the similarity between the original and retargeted images at each iteration. However, these methods depend on low-level feature-based saliency maps, which can barely reflect visual semantics.

At present, deep learning- or perception-based approaches have facilitated further research on image retargeting. Esmaeil et al. [7] utilized a fully-convolutional deep neural network to develop a cropping-based thumbnail generation framework by learning specific filters for thumbnails of different sizes and aspect ratios. Liu et al. [19] developed an aggregation-based CNN to learn the deep representation for gaze shifting path and then used these features for image retargeting through a probabilistic model. Cho et al. [3] utilized a weakly- and self-supervised deep CNN to retarget source images directly to target ratio. Xia et al. [33] proposed a photo retargeting model by learning human gaze shifting process, in which a few active graphlet paths were selected on the basis of a sparsity-guided ranking algorithm. Zhou et al. [36] used photographs marked as aesthetically pleasing for training and utilized the learned priors to shrink the corresponding gaze shifting path of a retargeted photograph to maximize its similarity to those from the training photographs. Noticeably, the above-mentioned perception-guided retargeting methods still cannot capture user preferences in resizing an image in accordance with three perceptual aspects: *salient content*, *object shape*, and *scene composition*.

3 DATASET

We introduce a new scheme to observe human subjective perception toward image retargeting task by building a synergy between visual images and human perception (Section 3.1). We investigate human behavior in balancing RVPIs, which are related to information loss of images during retargeting (Section 3.2).

3.1 Perception Formulation

Perception-aware image retargeting results are generated by formulating abstract human perception into concrete representation. Defining and quantifying the distribution among RVPIs directly are difficult even for experts. However, people know good retargeting results depending on their perception toward different images. Given the lessons from *multi-operator* approaches, we observe people's behavior in balancing RVPIs by offering them three basic retargeting operators to retarget images manually. The details of the *operator selection* are given as follows:

(1) We use *seam carving* [24] to measure loss in *salient content*. This CAIR method carves one seam with the lowest energy each time in accordance with the energy functions defined beforehand. A high execution time of seam carving (R_{sc}) indicates a large proportion of salient content of the original image will be preserved and large proportions of object shape and scene composition may be damaged.
(2) We use *cropping* to measure loss in *object shape*. This simple method removes outer parts of an image to protect the subject. A high execution time of cropping (R_{cr}) indicates a large proportion of object shape of the original image will be preserved and large proportions of salient content and scene composition may be lost.
(3) We use *scaling* to measure loss in *scene composition*. This uniform method transforms the original size to the target size given a scale factor. A high execution time of scaling (R_{sl}) indicates a large proportion of scene composition of the original image will be preserved and a large proportion of object shape of the original image may be distorted.

Without loss of generality, we arrange the operator order as *seam carving*→*cropping*→*scaling*. Then, human perceptions are formulated into a fixed order that is filled with the number of times that each operator is performed to generate retargeted images, denoted as $[R_{sc}, R_{cr}, R_{sl}]$.

Notably, we do not add warping into our framework due to three reasons: first, integrating warping may introduce artifacts of boundary distortion and over-stretching of homogeneous content [29]. Second, the functionality of warping can be substituted by seam caving and scaling. Third, adding operators will increase the difficulty in data annotation.

3.2 Data Collection

We collect 5,084 images as the supporting database on the basis of the following principles: 1) *Popularity*: We collect images from Flickr, Pinterest, and Pexels under the Creative Commons license because of the aim of contributing to social media. By summarizing tags in "the most popular tag in history" block on each website, eight categories, namely, nature landscape, portrait, animal, food, art, fashion, festival, and architecture, are obtained, and used as keywords to search images. 2) *Comprehensive*: Images from "RetargetMe" [23], which is a classic benchmark for image retargeting methods, are added. 3) *Operability*: Duplicate images and images with too large or too small aspect ratio are manually picked out. Figure 2 shows some images in our dataset. The original aspect ratios of these images range from 0.52 to 3.97. The saliency map is also prepared for each image by using the method in [15].

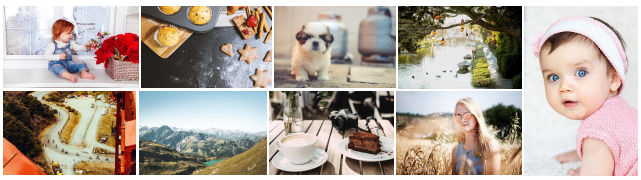


Figure 2: Images in the collected dataset.

Six expert photographers (3 males, 3 females, age range of 20-45) are invited to do the annotation. Each participant needs to square all the 5,084 images by allocating the three operators. We show the original image and a bar with two sliders (Figure 3) to participants. For each photo, the initial allocation of the three operators is calculated by a state-of-the-art multi-operator image retargeting method [4]. The squarization result will be generated and displayed in real time when the slider value is changed. Participants are asked to adjust the sliders freely with no time limitation until they see the ideal results.

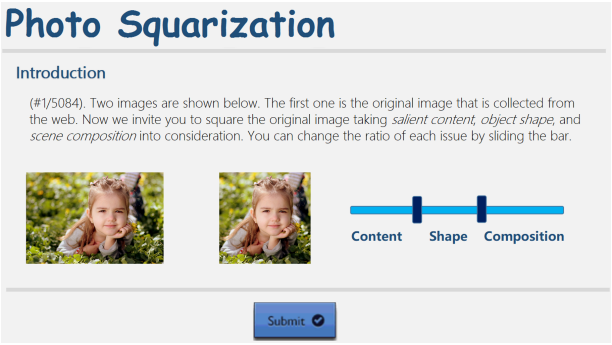


Figure 3: Annotation website for data collection.

We calculate Kendall's tau (τ) coefficient among the six participants. The results of $\tau = 0.85$ and $sig. = 3.11e^{-14}$ confirm that the participants have a general consensus with regard to the rating of three operators. We use the set of values that has the smallest difference from the average value of the six participants as the final annotation for each image. Figure 4 shows the average numbers of each operator for every 100 times of execution by manual labeling and automatic calculation following the method in [4]. From the results, we can observe that seam carving as a CAIR technique is adopted most frequently. Therefore, humans are highly sensitive to the change in salient image content. Human behaviors in balancing RVPIs are different from the automatically calculated assignments.

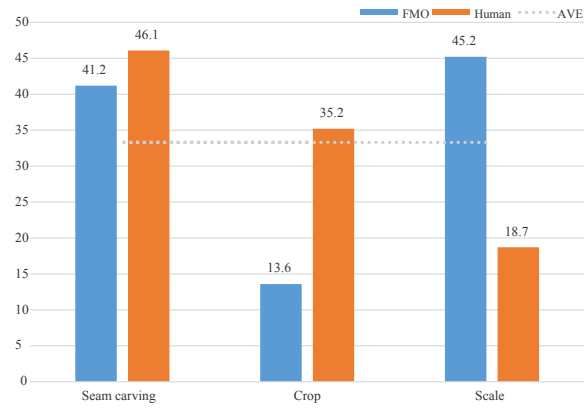


Figure 4: Allocation statistics for data collection.

4 APPROACH

On the basis of the annotated dataset, we propose a data-driven approach to learn human behavior in balancing the RVPIs when retargeting a photo to square by using multiple retargeting operators. Then, given an input photo, we use the model to predict the allocation of the percentage of each operator for squarization, which is consistent with human perception preference.

4.1 Problem Formulation

Given a photo I of size (w,h), we define $[R_{sc}, R_{cr}, R_{sl}]$ as the operation allocation performed on one dimension of I to obtain a square image T of size (t,t), where $t = min(w,h)$. Similar to all social media, we square the input photo by retargeting at the shorter dimension. This way can ensure that the maximum information of the original photo is kept. The squarization is performed using a combination of the three operators by regular sequences. Without loss of generality, we use the reduction in image width as an example. Specifically, to reduce the width of an image I by $R = max(w,h) - t$ pixels, R_{sc} seams are carved out, R_{cr} columns are cropped from the image, and the image is scaled by R_{sl} pixels, where $R_{sc} + R_{cr} + R_{sl} = R$. We let $r = [r_{sc} = \frac{R_{sc}}{R}, r_{cr} = \frac{R_{cr}}{R}, r_{sl} = \frac{R_{sl}}{R}]$ be the normalized representation of $[R_{sc}, R_{cr}, R_{sl}]$ and focus on learning the mapping function $f : f(I) \rightarrow r \in \mathbb{R}^3$. Then, our learning goal can be formulated as the following equation:

$$\begin{aligned} & avg \min ||r^* - r||_2 \\ & s.t.\ r^*_{sc}, r^*_{cr}, r^*_{sl} \geq 0, \\ & \quad r^*_{sc} + r^*_{cr} + r^*_{sl} = 1\,, \end{aligned} \tag{1}$$

where $r^* = f(I)$ is the learning target.

4.2 Learning Approach

Figure 5 shows the overall structure of our learning approach. Two modules are adopted: perception and distillation modules. We propose the perception part to learn the relationships among different operations and distill such "dark knowledge" into the second part. Different from most distillation models that focus on classification research, our learning target proposed in Section. 4.1 is an objective regression with constraints that are applied to the distillation module. Details about the two modules are given as follows.

Perception module. Instead of directly solving the optimization problem raised in Equation (1), we propose a perception module to learn the relationship among the three kinds of operations. Given the real-value vector r_i, the most frequently used operator can be treated as the superior operator OP_i for image I. The perception module focuses on the classification of superior operators. For this typical multiclass classification problem, we adopt cross entropy loss with l_2 normalization as the objective function.

The perception model is designed on the basis of VGG16 architecture [26]. The pre-trained VGG16 parameters are used as the seed, but the parameters of the last three fully connected layers are dropped. We initialize them by Xavier initialization [11]. Batch normalization is also added. Then, a mapping between the original image and the perceptive value of network will be formed. We define the output of this perception module as r_p which serves as the soft target for distillation part.

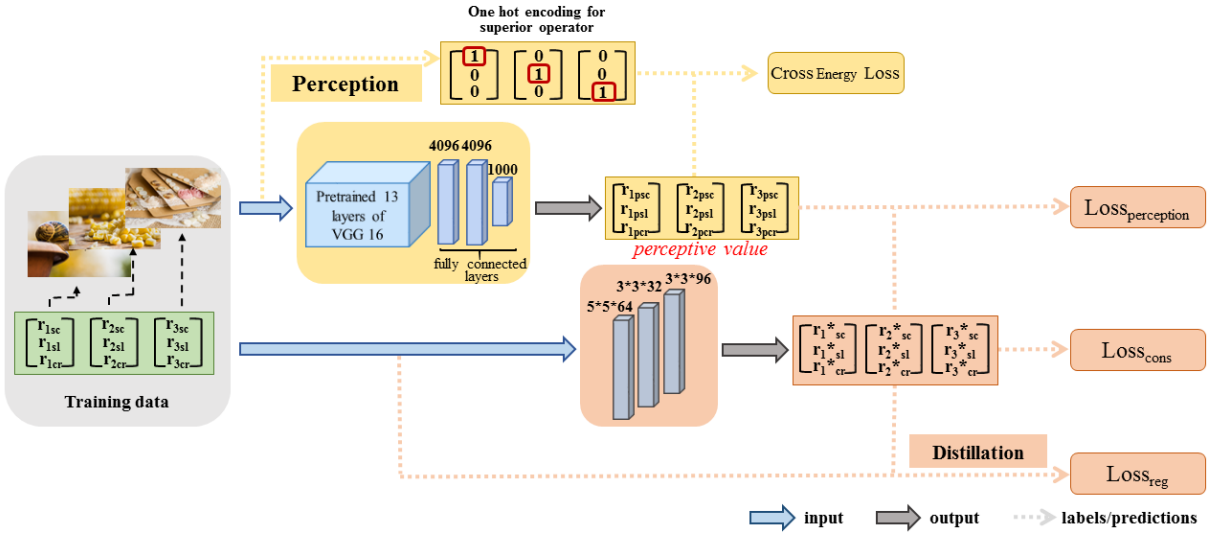


Figure 5: Overall structure of learning approach. Two modules are introduced: perception and distillation modules. The desired outputs are generated by the distillation module, which is trained using three types of losses.

Distillation module. Distillation technology proposed recently provides fresh portion to construct networks and can transform a large model to a small one [9, 34]. Distillation has difficulty reducing the network structure while keeping the performance in a small one. By adopting a reasonable distillation technology, the efficiency of network execution can be enhanced greatly and the overfitting can be relieved to some degree.

We transfer the output of the perception module to the distillation module by training it with soft targets for the non-special classes in addition to training it using Equation (1). For the distillation module, we set the optimized objective to train by considering the two aspects. One is for transferring soft targets r_p, which is also known as *dark knowledge* learned by the perception module into the distillation module, namely, $Loss_{perception}$. The other is the offset between the input label r and the output r^* together with the constraints in Equation (1), namely, $Loss_{target}$:

$$Loss(r^*, r_p, r) = Loss_{perception}(r^*, r_p) + \lambda \cdot Loss_{target}(r^*, r), \quad (2)$$

where λ balances the two kinds of losses and $Loss_{perception}(r^*, r_p)$ is the Euclidean distance between r^* and r_p. We set $\lambda = 2$ in our experiments.

$Loss_{target}$ is calculated as

$$\begin{aligned} Loss_{target}(r^*, r) &= Loss_{reg}(r^*, r) + \beta Loss_{cons}(r^*) \\ &= ||r^* - r||_2 + \beta Loss_{cons}(r^*), \end{aligned} \quad (3)$$

where $Loss_{cons}(r^*)$ encodes the constraints in Equation (1):

$$Loss_{cons}(r^*) = (\sum_{k=1}^{3} r^*(k) - 1)^2 + \sum_{k=1}^{3} max(0, -r^*(k)), \quad (4)$$

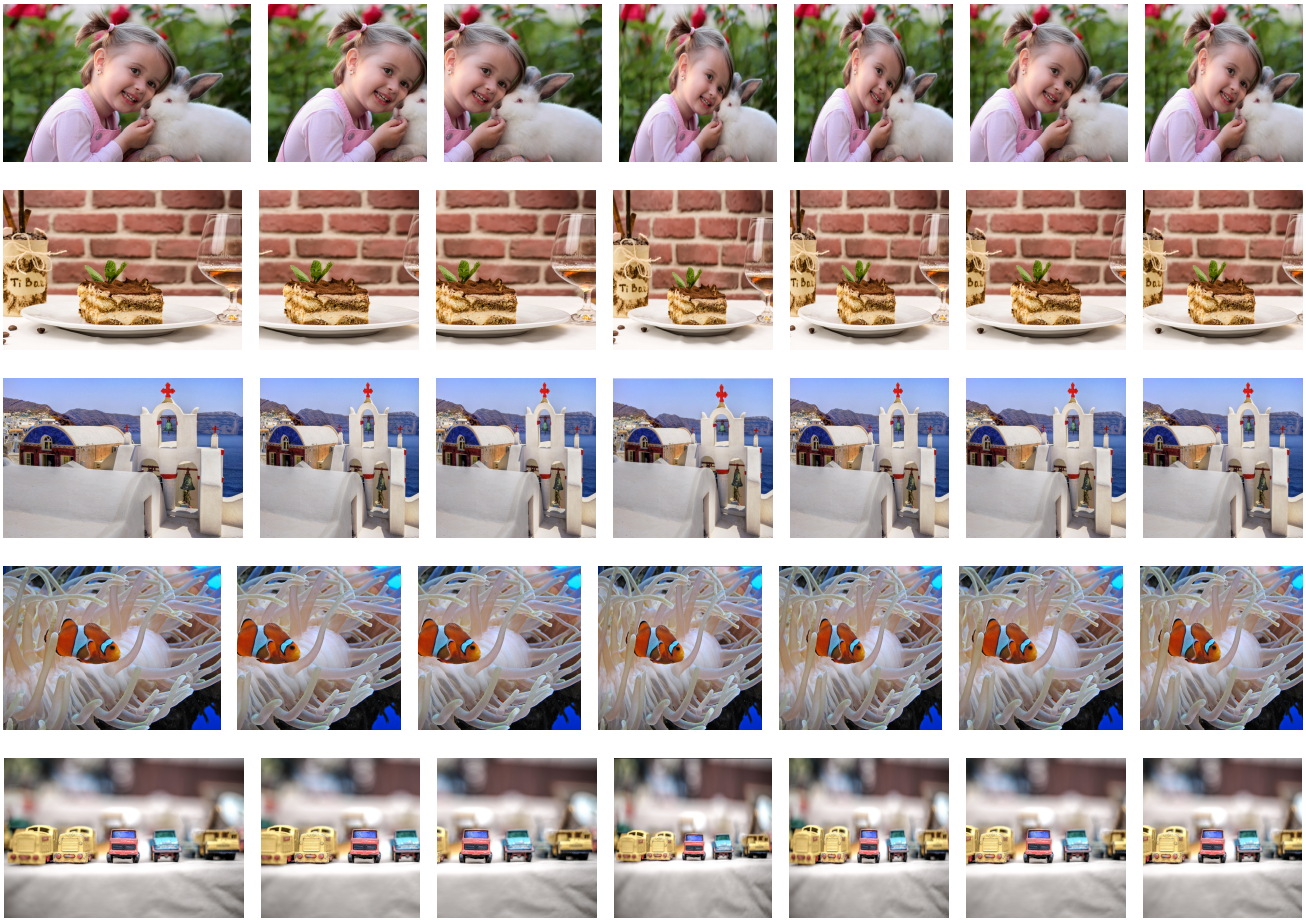
where $r^*(\cdot)$ is the prediction for one of the three operations. The first item in Equation (4) corresponds to the equality constraint. The second item equals 0 when $r^*(\cdot) \geq 0$.

Specifically, the distillation module is composed of a shallow three-layer CNN. The first convolutional layer is fed by the input image with 64 kernels of size 5×5. The second convolutional layer takes the output of the first convolutional layer as input and filters it with 32 kernels of size 3×3. The third convolutional layer is set with 96 kernels of size 3×3 with a pad of one pixel. The output layer is fully connected to the last convolutional layer with three output neurons.

5 IMPLEMENTATION AND EXPERIMENTS

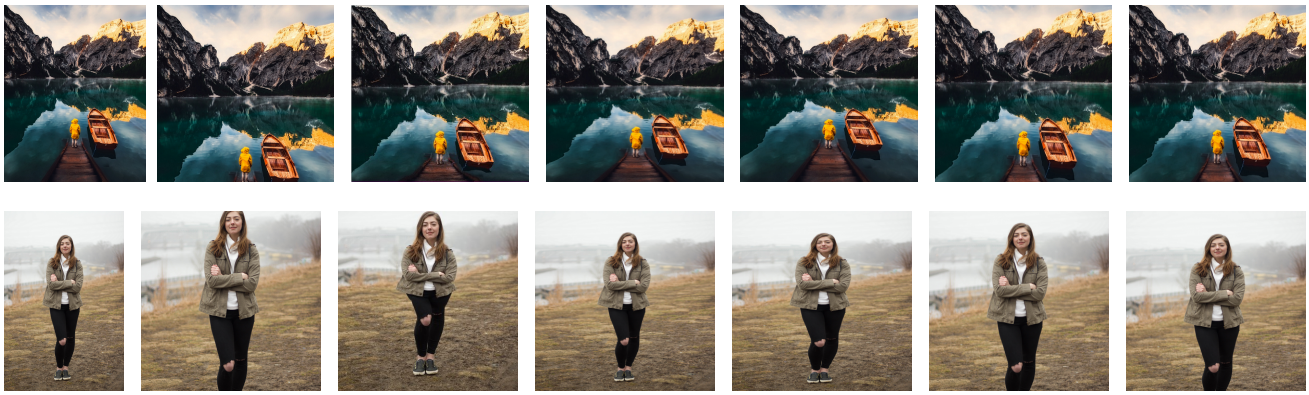
5.1 Implementation Details

We randomly choose 3,660 images from the collected dataset for training and use the rest as the evaluating set. Rotation and contrast adjustments are performed for the data argumentation. In accordance with the two-module method, our training process consists of two parts as well. The training set is sent into the perception part at the beginning to achieve the perceptive values, which will be used as the new labels to the distillation part. The configuration of the two modules is set as learning rate: 10^{-5}, batch size: 20. When evaluating, only the distillation module will be activated and the outputs of the distillation module are l_1 normalized. Then, the predicted allocation r^* is transferred to the number of times that each operator is adopted by multiplying R. Predictions less than zero are set to zero. The entire network is optimized on a PC equipped with 3.6 GHZ Intel Core i7 and Nvidia Geforce GTX 1080Ti GPU with 11172 MB memory. The implementation is based on the Tensorflow platform [1]. Figures 6 and 7 show the squarization results. Our method can generate better results than other state-of-the-art CAIR methods and these results are visually consistent with human preference results. More results are shown in the supplementary material.



(a) Original Photo (b) CR (c) SC (d) WARP (e) FMO (f) Human (g) Ours

Figure 6: Comparison of our photo squarization results with those of other state-of-the-art methods for transverse images.



(a) Original (b) CR (c) SC (d) WARP (e) FMO (f) Human (g) Ours

Figure 7: Comparison of our photo squarization results with those of other state-of-the-art methods for longitudinal images.

5.2 Experiments

To solve the regression task proposed in Equation (1), our approach adopts two modules including several losses. Comparisons on losses and net structures are provided. Furthermore, we discuss the computational time of our approach.

Evaluation metrics. We use Mean Absolute Error (MAE) and Root Means Square Error (RMSE) as evaluation metrics. MAE measures the absolute difference between r and r^*, whereas

$$RMSE = \sqrt{\frac{1}{N}\sum_{N}\sum_{k=1}^{3}(r^*(k) - r(k))^2},$$

where N is the size of testing set.

Loss variants. We test the performance of loss variants on three kinds of CNN structures: the shallow three-layer CNN used in distillation module, VGG16 [26], and ResNet [12].

- CNNs with regression loss. We fine tune CNNs on the training data to solve the regression problem directly.
- CNNs with regression loss and constraints. We fine tune CNNs on the training data to solve the regression problem with the $Loss_{cons}$ described in Equation (4).

Notably, all the outputs are l_1 normalized. We compare the proposed method with the fast multi-operator (FMO) [4] method, which uses the same three operators.

The results on testing data are reported in Table 1. Although all the predicted values are forced to meet the constraints in Equation (1) by using l_1 normalization, adopting constraint loss to train the network can yield satisfying results. We conduct paired sample t-test to evaluate whether the improvement of using variant losses is significant. the $p-values$ when using VGG16 and ResNet structures are 0.041 and 0.009, respectively, which demonstrate the significance of the improvements; by contrast, for the shallow network, the $p-value = 0.341 > 0.05$ indicates the constraint loss does not work efficiently.

Table 1: Comparisons of loss function variants.

Strategy	MAE		RMSE	
	reg.	reg.+cons.	reg.	reg.+cons.
Shallow	0.50	0.51	0.60	0.57
VGG16	0.40	0.35	0.41	0.31
ResNet	0.37	0.33	0.32	0.25
FMo	0.37		0.27	
Ours	**0.14**		**0.16**	

Net structure. Methods that try to model the RVPI directly by a one-way CNN with all the three kinds of losses have been tested. RMSE for ResNet and VGG are 0.25 and 0.31, respectively. Table 1 shows that our approach reports a better performance than the one-way structure. The reason is that the RVPIs are related to one another and difficult to be predicted directly. The solution requires relevant analysis on RVPIs. When the entire structure is separated into two parts, the outcomes of perception module can imply the inner structure of the RVPIs that are not encoded in the original regression/classification labels and are transferred into the distillation part. Thus, we convert the original regression labels into category labels and propose an additional classification task, namely, perception module.

Computational time. In terms of the computational time of the network, the training process takes about 2 h to converge. When testing, the distillation model takes about 0.5 s to process 100 images. Therefore, this model is about six times faster than a fine-tuned VGG16 style model using our dataset. This comparison of computational time exhibits the advantage of distillation structure in terms of running speed (Section 4.2). Notably, while the FMO procedure needs about 7 s per image. Our method is 1, 400 times faster than FMO and thus exhibits a better real-time character.

5.3 User Study

To quantitatively evaluate our contribution, we set up two user studies that involve 65 investigators (31 males, 34 females, age range of 20-45) with different experiences. First, we conduct a user study to compare the retargeting results with those of state-of-the-art CAIR algorithms, namely, cropping (CR hereinafter), FMO [4], AAD warping [20], and seam carving (SC hereinafter) [2]. We display 96 sets of retargeting images in turn, which contain the original image and six retargeting results generated respectively from CR, FMO, AAD, SC, humans, and our method. The retargeting results are shown in a random order except the original image that is always displayed in front. Then, we ask the investigators to select the three best results in each set by clicking the mouse. The selection has no time limitation, but the entire study needs to be completed within 1 h.

Table. 2 reports the distribution of the votes. We can observe that our method outperforms other methods and is comparable to human perception. For each method involved in the user study, the distribution of user voting obeys binomial distribution. Given that the subjects are asked to select three results out of six images each time, the expected value equals $\frac{C_5^2}{C_6^3} = 0.5$. The mean voting scores of AAD, our method and human perception are larger than 0.5. We conduct binomial test to evaluate whether the superiority is significant. The p-values are 0.031, 0.010, and 0.003. Therefore, the users think the three methods significantly perform more than the average.

Table 2: Statistics of votes in user study.

Method	CR	FMO	AAD	SC	Human	Ours
Mean Voting	40%	47%	**53**%	30%	**68**%	**62**%

Second, we perform another user study to examine the deviation between our result and human re-perception result. Re-perception indicates we run the annotation progress again with the start allocation set to our result. The subjects involved in the second user study is the same as in the first one, in which 96 images are tested. The deviation between the start point and the re-perception distribution is recorded. The entire study time is set to 2 h and has no limitation for a certain image adjustment. The average time consumption obtained by statistical analysis is 1 h 22 min 14 s,

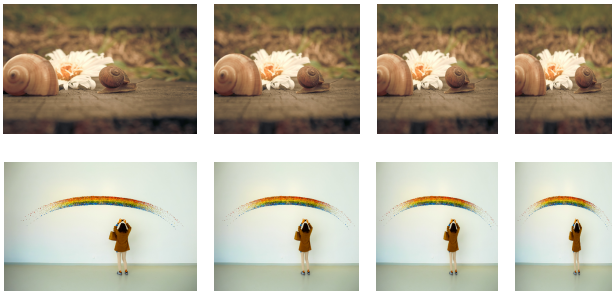
and the deviations are recorded. Table 3 shows the statistics. We can observe that people have less distribution adjustments for our results. Therefore, our results are more close to people's subjective consciousness.

Table 3: Statistics of distribution adjustments in re-perception.

Deviation of	FMO	Ours
Distribution adjustments	37%	12%

5.4 Discussion

The measure of RVPIs should be an inner property depending on the original image. Hence, we test the learned model to arbitrary target aspect ratios. Figure 8 illustrates the retargeting results by using the learned allocation of the operators on three target ratios. Although we focus on photo squarization, the proposed method can also be attempted to universal problems.



(a) Original Photo (b) 75.0% (c) 62.5% (d) 50%

Figure 8: Retargeting to arbitrary aspect ratios.

Another emphasis is that the learned model is based on the consistency of human perception. Although the overall analysis of Kendall's tau (τ) coefficient shows that people can reach agreements on most images, consistency is low in some cases. Figure 9 shows an example in which the τ coefficient is lower than 0.3. When the aspect ratio of the original image (Figure 9(a)) is abnormal, the coefficient level among users is low. For these instances, participants can not square it to good results even if they use all kinds of proportional combination. Under such circumstances, even the trained model can generate results as human. Therefore, the results of squarization are usually not ideal.

Figure 10 shows a failure case of the proposed model. The average difference of operator utilization time between the predicted operator allocation (Figure 10(b)) and the annotated allocation (Figure 10(c)) is 0.313. Additional cropping operations are used in our results to protect the object shape.

6 CONCLUSION AND FUTURE WORK

In this study, we focus on the generation of square thumbnails, that is, "photo squarization," of images on social media platform. We propose a data-driven approach that considers human perception to learn and predict human behavior in dealing with three basic issues in image retargeting, namely, *salient content*, *object shape*, and *scene composition*. We establish a new dataset with human perception information and train a two-module CNN framework that takes advantage of deep learning and multi-operator image retargeting. Experimental results show that the proposed method can create highly appealing results.

In the proposed squarization process, the length of the original short side is adopted as the target dimension. Therefore, the changing of aspect ratio acts as a down-sampling operation, thereby ignoring up-sampling operation that may have an outperforming effect. In the future, we will consider up-sampling operation. The database will also be enlarged to offer a large number of photo examples. Human behavior in dealing with different target aspect ratios will also be analyzed. Other visual media types, such as videos, will be considered as well.

(a) Original Photo (b) Human (c) Ours

Figure 9: Special cases of the dataset.

(a) Original Photo (b) r=(0.38,0.14,0.48) (c) r^*=(0.45,0.27,0.28)

Figure 10: Failure case of our method. The predicted operation allocation has low consistency with the ground truth.

ACKNOWLEDGMENTS

We thank the users on Pexels and Pixabay who share their great works under CC0 Creative Commons. This work was supported by National Natural Science Foundation of China under nos. 61832016, 61672520 and 61702488, by Beijing Natural Science Foundation under No. 4162056, by the independent research project of National Laboratory of Pattern Recognition, by the Leading Talents of Guangdong Program under No. 00201509, and by the Ministry of Science and Technology under No. MOST-104-2221- E-006-044-MY3, Taiwan.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. TensorFlow: A System for Large-Scale Machine Learning.. In *OSDI*, Vol. 16. 265–283.

[2] Shai Avidan and Ariel Shamir. 2007. Seam Carving for Content-Aware Image Resizing. *ACM Transactions on Graphics* 26, 3 (July 2007), 10:1–10:10.

[3] Donghyeon Cho, Jinsun Park, Tae-Hyun Oh, Yu-Wing Tai, and In So Kweon. 2017. Weakly- and Self-Supervised Learning for Content-Aware Deep Image

Retargeting. In *IEEE International Conference on Computer Vision (ICCV)*. 4568–4577.
[4] Weiming Dong, Guanbo Bao, Xiaopeng Zhang, and Jean-Claude Paul. 2012. Fast Multi-Operator Image Resizing and Evaluation. *Journal of Computer Science and Technology* 27, 1 (2012), 121–134.
[5] Weiming Dong, Fuzhang Wu, Yan Kong, Xing Mei, Tong-Yee Lee, and Xiaopeng Zhang. 2016. Image Retargeting by Texture-Aware Synthesis. *IEEE Transactions on Visualization and Computer Graphics* 22, 2 (2016), 1088–1101.
[6] Weiming Dong, Ning Zhou, Jean-Claude Paul, and Xiaopeng Zhang. 2009. Optimized Image Resizing Using Seam Carving and Scaling. *ACM Transactions on Graphics* 28, 5 (Dec. 2009), 125:1–125:10.
[7] Seyed A. Esmaeili, Bharat Singh, and Larry S. Davis. 2017. Fast-At: Fast Automatic Thumbnail Generation Using Deep Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4178–4186.
[8] Yuming Fang, Zhijun Fang, Feiniu Yuan, Yong Yang, Shouyuan Yang, and Neal N. Xiong. 2017. Optimized Multioperator Image Retargeting Based on Perceptual Similarity Measure. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 47, 11 (Nov 2017), 2956–2966.
[9] Nicholas Frosst and Geoffrey E. Hinton. 2017. Distilling a Neural Network Into a Soft Decision Tree. *CoRR* abs/1711.09784 (2017). arXiv:1711.09784 http://arxiv.org/abs/1711.09784
[10] Ran Gal, Olga Sorkine, and Daniel Cohen-Or. 2006. Feature-Aware Texturing. In *Proceedings of the 17th Eurographics Conference on Rendering Techniques (EGSR '06)*. Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 297–303.
[11] Xavier Glorot and Yoshua Bengio. 2010. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Vol. 9. PMLR, Chia Laguna Resort, Sardinia, Italy, 249–256.
[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.
[13] Bert P. Krages. 2005. *Photography: The Art of Composition*. Allworth Press.
[14] Philipp Krähenbühl, Manuel Lang, Alexander Hornung, and Markus Gross. 2009. A System for Retargeting of Streaming Video. *ACM Transactions on Graphics* 28, 5 (2009), 126:1–126:10.
[15] Gayoung Lee, Yu-Wing Tai, and Junmo Kim. 2016. Deep Saliency with Encoded Low Level Distance Map and High Level Features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 660–668.
[16] Bing Li, Ling-Yu Duan, Jinqiao Wang, Rongrong Ji, Chia-Wen Lin, and Wen Gao. 2014. Spatiotemporal Grid Flow for Video Retargeting. *IEEE Transactions on Image Processing* 23, 4 (April 2014), 1615–1628.
[17] Dongping Li, Kaiming He, Jian Sun, and Kun Zhou. 2015. A Geodesic-Preserving Method for Image Warping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 213–221.
[18] Shih Syun Lin, I Cheng Yeh, Chao Hung Lin, and Tong Yee Lee. 2013. Patch-Based Image Warping for Content-Aware Retargeting. *IEEE Transactions on Multimedia* 15, 2 (2013), 359–368.
[19] Zhenguang Liu, Zepeng Wang, Luming Zhang, Rajiv Ratn Shah, Yingjie Xia, Yi Yang, and Xuelong Li. 2017. FastShrinkage: Perceptually-aware Retargeting Toward Mobile Platforms. In *Proceedings of the 2017 ACM on Multimedia Conference (MM '17)*. ACM, New York, NY, USA, 501–509.
[20] Daniele Panozzo, Ofir Weber, and Olga Sorkine. 2012. Robust Image Retargeting via Axis-Aligned Deformation. *Computer Graphics Forum* 31, 2 (2012), 229–236.
[21] Yael Pritch, Eitam Kav-Venaki, and Shmuel Peleg. 2009. Shift-Map Image Editing. In *IEEE International Conference on Computer Vision (ICCV)*. 151–158.
[22] Shaoyu Qi, Yu-Tseh Jason Chi, Adrian M. Peter, and Jeffrey Ho. 2016. CASAIR: Content and Shape-Aware Image Retargeting and Its Applications. *IEEE Transactions on Image Processing* 25, 5 (May 2016), 2222–2232.
[23] Michael Rubinstein, Diego Gutierrez, Olga Sorkine, and Ariel Shamir. 2010. A Comparative Study of Image Retargeting. *ACM Transactions on Graphics* 29, 6 (Dec. 2010), 160:1–160:10.
[24] Michael Rubinstein, Ariel Shamir, and Shai Avidan. 2008. Improved Seam Carving for Video Retargeting. *ACM Transactions on Graphics* 27, 3 (2008), 16:1–16:10.
[25] Michael Rubinstein, Ariel Shamir, and Shai Avidan. 2009. Multi-operator Media Retargeting. *ACM Transactions on Graphics* 28, 3 (July 2009), 23:1–23:11.
[26] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations (ICLR)*. Vancouver, BC, Canada.
[27] Jin Sun and Haibin Ling. 2013. Scale and Object Aware Image Thumbnailing. *International Journal of Computer Vision* 104, 2 (2013), 135–153.
[28] Wei-Tse Sun, Ting-Hsuan Chao, Yin-Hsi Kuo, and Winston H. Hsu. 2017. Photo Filter Recommendation by Category-Aware Aesthetic Learning. *IEEE Transactions on Multimedia* 19, 8 (Aug 2017), 1870–1880.
[29] Jinqiao Wang, Zhan Qu, Yingying Chen, Tao Mei, Min Xu, La Zhang, and Hanqing Lu. 2016. Adaptive Content Condensation Based on Grid Optimization for Thumbnail Image Generation. *IEEE Transactions on Circuits and Systems for Video Technology* 26, 11 (Nov 2016), 2079–2092.
[30] Wenguan Wang and Jianbing Shen. 2017. Deep Cropping via Attention Box Prediction and Aesthetics Assessment. In *IEEE International Conference on Computer Vision (ICCV)*. 2205–2213.
[31] Yu-Shuen Wang, Hui-Chih Lin, Olga Sorkine, and Tong-Yee Lee. 2010. Motion-based Video Retargeting with Optimized Crop-and-Warp. *ACM Transactions on Graphics* 29, 4 (July 2010), 90:1–90:9.
[32] Yu-Shuen Wang, Chiew-Lan Tai, Olga Sorkine, and Tong-Yee Lee. 2008. Optimized Scale-and-Stretch for Image Resizing. *ACM Transactions on Graphics* 27, 5 (2008), 118:1–118:8.
[33] Yingjie Xia, Luming Zhang, Richang Hong, Liqiang Nie, Yan Yan, and Ling Shao. 2017. Perceptually Guided Photo Retargeting. *IEEE Transactions on Cybernetics* 47, 3 (March 2017), 566–578.
[34] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. 2017. A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7130–7138.
[35] Luming Zhang, Mingli Song, Yi Yang, Qi Zhao, Chen Zhao, and Nicu Sebe. 2014. Weakly Supervised Photo Cropping. *IEEE Transactions on Multimedia* 16, 1 (2014), 94–107.
[36] Yinzuo Zhou, Luming Zhang, Chao Zhang, Ping Li, and Xuelong Li. 2018. Perceptually Aware Image Retargeting for Mobile Devices. *IEEE Transactions on Image Processing* 27, 5 (May 2018), 2301–2313.
[37] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *IEEE International Conference on Computer Vision (ICCV)*. 2242–2251.