Image Retargetability

Fan Tang, Weiming Dong, *Member, IEEE*, Yiping Meng, Chongyang Ma, Fuzhang Wu, Xinrui Li, and Tong-Yee Lee, *Senior Member, IEEE*

***Abstract*—Real-world applications could benefit from the ability to automatically retarget an image to different aspect ratios and resolutions while preserving its visually and semantically important content. However, not all images can be equally processed. This study introduces the notion of image retargetability to describe how well a particular image can be handled by content-aware image retargeting. We propose to learn a deep convolutional neural network to rank photo retargetability, in which the relative ranking of photo retargetability is directly modeled in the loss function. Our model incorporates the joint learning of meaningful photographic attributes and image content information, which can facilitate the regularization of the complicated retargetability rating problem. To train and analyze this model, we collect a dataset that contains retargetability scores and meaningful image attributes assigned by six expert raters. The experiments demonstrate that our unified model can generate retargetability rankings that are highly consistent with human labels. To further validate our model, we show the applications of image retargetability in retargeting method selection, retargeting method assessment and generating a photo collage.**

***Index Terms*—Image retargetability, visual attributes, multi-task learning, deep convolutional neural network.**

Manuscript received August 23, 2018; revised January 4, 2019 and May 5, 2019; accepted July 22, 2019. Date of publication August 1, 2019; date of current version February 21, 2020. This work was supported in part by National Key R&D Program of China under 2018YFC0807500, and in part by National Natural Science Foundation of China under Grants 61832016, 61672520, and 61702488, and in part by Ministry of Science and Technology under 108-2221-E-006-038-MY3, Taiwan, and in part by CASIA-Tencent Youtu joint research project. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xavier Giro-i-Nieto. (*Corresponding author: Weiming Dong.*)

F. Tang is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100864, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: tangfan2013@ia.ac.cn).

W. Dong is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100864, China (e-mail: weiming.dong@ia.ac.cn).

Y. Meng is with the Didi Chuxing, Beijing 100094, China (e-mail: mengyipingkitty@didichuxing.com).

C. Ma is with the Kuaishou Technology, Beijing 100085, China (e-mail: chongyangm@gmail.com).

F. Wu is with the Institute of Software, Chinese Academy of Sciences, Beijing 100864, China (e-mail: fuzhang@iscas.ac.cn).

X. Li is with the Department of Mathematics and Physics, North China Electric Power University, Beijing 102206, China (e-mail: szyclxr@163.com).

T.-Y. Lee is with the National Cheng Kung University, Tainan 701, Taiwan (e-mail: tonylee@mail.ncku.edu.tw).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2019.2932620

I. Introduction

CONTENT-aware image retargeting (CAIR) addresses the increasing demand of display images on devices with varying resolutions and aspect ratios while preserving its visually important content and avoiding observable artifacts [1]–[7]. Although state-of-the-art image retargeting techniques can successfully handle numerous images, whether a specific image can be successfully retargeted beforehand remains unclear. CAIR techniques typically expect that the input image contains a mid-sized salient object and a relatively simple background, for which the majority of information can be presented in a small space. The retargeting results may present severe artifacts if the input images contain rich contents or geometric structures that may be damaged. Furthermore, not all CAIR methods work equally well for the same input. The optimal approach which considers quality and robustness depends on the input image and target resolution. For example, warping-based retargeting methods [4], [5], [8] are effective and popular, but tends to overstretch or oversqueeze some contents when salient shapes should be preserved.

To address the problems of the CAIR method selection and result evaluation, we introduce the notion of "image retargetability" to quantitatively compute how well the image can be retargeted on the basis of its visual content. Fig. 1 shows the predicted retargetability scores of several input images and the corresponding results of the "best" retargeting method selected by our system.

We are inspired by some recent studies on quantifying qualitative image properties, such as interestingness [9], memorability [10], synthesizability [11], and mirrorability [12]. To compute image retargetability, we adopt a data-driven methodology and collected a dataset of 13,584 sample images from Internet photos (Section III). For each image in the dataset, we apply multiple retargeting methods and request six expert raters to label the quality of each retargeting result in one of the following three levels: good, acceptable, and bad (see Fig. 5 for examples). We also ask the raters to annotate a set of high-level visual attributes for each sample image in the dataset, including repeating patterns, specific geometric structures, perspective, fuzzy, text, and shading contrast.

We propose to quantitatively measure and analyze image retargetability on the basis of the collected dataset with manual annotations. We demonstrate that there is a strong correlation between image retargetability and other visual attributes. We use this observation as basis to leverage a deep convolutional neural network (NN) and propose a multi-task learning approach by

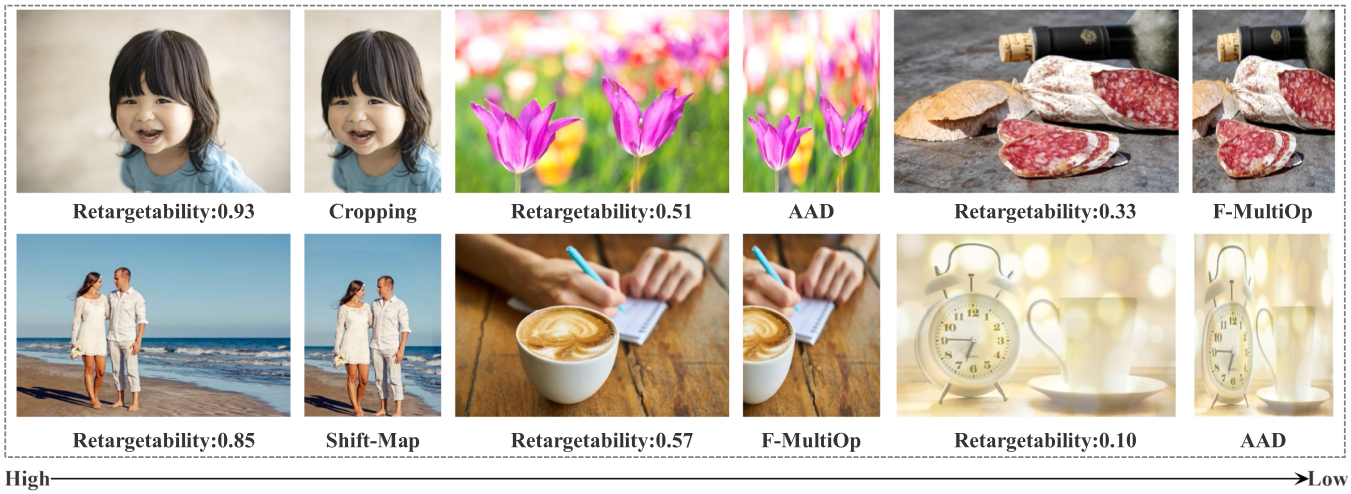


Fig. 1. Retargetability of images predicted by our system. The values are in $[0, 1]$, in which a high value indicates that the image is easy to retarget. For each group, left is the original image and right is the retargeted result generated by CAIR method suggested by our system.

jointly learning visual attributes from deep features and feature sharing for retargetability prediction (Section IV).

We evaluate the effectiveness of our framework for image retargetability prediction by comparing against a baseline approach in Section V. Given that each CAIR method exhibits its own advantages and limitations, no single CAIR algorithm that works better than other algorithms in all the cases has been produced. We demonstrate how to select the "best" CAIR method using our system. We also show that image retargetability is useful for retargeting method assessment and photo collage generation (Section VI).

In summary, our main contributions are as follows.

- We introduce image retargetability as a new quantitative property for image analysis.
- We collect a large image dataset to learn deep features for image retargetability prediction. The dataset and source code will be released upon final publication.
- We adopt a deep NN and propose a novel multi-task learning architecture to compute the retargetability of a given image.
- We demonstrate that image retargetability can facilitate several applications for image analysis/processing, such as retargeting method assessment/selection and generating a photo collage.

II. Related Work

Image retargeting algorithms: The concept of CAIR aims to preserve the important content of an image after resizing. Cropping has been widely used to eliminate unimportant information from the image periphery or improve the overall composition of an image [6], [13], [14]. However, cropping often destroys object completeness and causes unexpected information losses. Discrete methods remove or insert pixels or patches judiciously to preserve content. Seam carving methods iteratively remove a seam in the input image to preserve visually salient content [1], [15]. Shift-map method [3] performs a discrete labeling over individual pixels and retargets an image by removing segments in the net. These approaches are good at retargeting images with rich texture content but may occasionally cause local discontinuity artifacts. Continuous methods focus on preserving local structure and often optimize a warping from the source size to the target size, constrained on its important regions and permissible deformation [8], [16]–[18]. Panozzo *et al.* [4] minimized warping energy in the space of axis-aligned deformations (AAD) to avoid unnatural distortions. Kaufmann *et al.* [19] adopt a finite element method to formulate image warping. Lin *et al.* [5] presented a patch-based scheme with an extended significance measurement to preserve the shapes of visually salient objects and structural lines. Tan *et al.* [20] generated feature-preserving constraints in the space of AAD by calculating a feature salience map to guide the warping process. These approaches can smoothly preserve the image content geometric structure but may also permit minimally important and unwanted regions to appear in the retargeting result. Multi-operator methods [2], [21], [22] fuse three condensation operators (i.e., seam caring, cropping, and scaling) into a unified optimization framework. Different operators influence one another and are simultaneously optimized to retarget images. Summarization-based methods measure patch similarity and select patch arrangements that fit well together to change image size [23]–[25].

Researchers recently adopted deep learning techniques to solve CAIR and related tasks [26]–[29]. Guo *et al.* [30] cropped aesthetically pleasing regions on the basis of a novel cascaded cropping regression method. Song *et al.* [31] proposed a two-module deep architecture to encode the human perception for image retargeting task and perform a multi-operator-based photo squarization solution. These deep learning-based approaches are the extension of traditional cropping, warping, or multi-operator-based methods. Section III-B provides the details of the advantages and disadvantages of the different types of CAIR methods.

Image retargeting evaluations: Rubinstein *et al.* [32] present the first comprehensive perceptual study and analysis of image retargeting, created the RetargetMe benchmark, and conducted

a user study to compare the retargeted images generated by numerous state-of-the-art methods. An overall ranking of the retargeting methods has been provided on the basis of user study. Liu *et al.* [33] proposed an objective quality assessment metric that simulates the human vision system to compare image quality with different retargeting methods. Their experiments also suggest that no single method is absolutely superior to others in all the cases. Ma *et al.* [34] built an image retargeting quality data set to analyze different retargeting factors, including scales, methods, and image content. Zhang *et al.* [35] analyzed three determining factors for the human visual quality of experience, namely, global structural distortion, local region distortion, and loss of salient information. Fang *et al.* [36] generated a structural similarity map to evaluate if the structural information is well preserved in the retargeted image. Hsu *et al.* [37] proposed a novel full-reference objective metric for assessing the visual quality of a retargeted image on the basis of perceptual geometric distortion and information loss. Bare *et al.* [38] proposed a new feature and predicted the retargeted image quality by training an RBF NN. Wang *et al.* [39] analyzed human-scenery position relationship, which can be used to evaluate content composition, in retargeted images. Zhang *et al.* [40] adopted a novel aspect ratio similarity metric to measure the geometric change of the images as proven by how the original image is retargeted. Liang *et al.* [41] evaluated image retargeting quality through multiple factors, including preservation of salient regions, symmetry, and global structure, influence of artifacts, and aesthetics. Eye tracking data are also used to improve the performance of the objective quality metrics for retargeted image [42]. Rawat *et al.* [43] focused on the visual balance of social media images and provide real-time feedback on the relative size of image frame. Several studies [37], [38] have also learned to predict a score to discover how well the retargeted images are to indicate whether the quality of a specific retargeting result is good. However, these studies evaluated the image retargeting quality by comparing the original and retargeted images. By contrast, the current study focuses on predicting the quality of retargeting result from the input image itself, thereby possibly indicating if an image can be well retargeted.

Image property analysis: Various semantic properties of images have been widely analyzed. Rosenholtz *et al.* [44] measured the visual clutter of an image, which is useful for the retrieval of visual content. Recently, unusual photographs are found to be interesting [9], and images of indoor scenes with people are found to be memorable, whereas scenic and outdoor scenes are not [10]. Other qualitative image properties such as popularity [45], colorfulness [46], and aesthetics [47] have been also studied. Dai *et al.* [11] used the techniques of example-based texture synthesis as bases to quantify texture synthesizability as an image property, which can be learned and predicted. In text-based image retrieval, image specificity [48] based on image content and properties has been used to discriminate easily describable images.

The current research defines image retargetability as a semantic property to quantify the probability that an image can be well retargeted. We show that this notion is closely related to deep relative attributes [49].

III. Dataset Preparation

This section introduces our data set preparation for image retargetability investigation. First, we collect a large set of images and manually label each image with a few attributes on the basis of visual content (Section III-A). Second, we apply four typical CAIR methods to all the images in the dataset and manually annotate the quality of each retargeting result (Section III-B).

A. Images and Attributes

Our framework is designed to measure image retargetability on a wild spectrum of natural images. Accordingly, the dataset should be considerable variability in terms of contents and compositions. Although the "RetargetMe" benchmark [32] has been widely used in image retargeting works for quality assessment, this dataset only contains 80 images, which are inadequate for the reliable learning of image attributes. To learn retargetability prediction, we prepare an image dataset and manually annotate the sample inputs in terms of retargetability. We collect 14,000 images from Flickr, Pinterest, 500px, and Pexels under Creative Commons license by providing 26 keywords acquired from 500px photo categories (see https://500px.com). The keywords cover the most common categories, such as animals, food, nature, sport, travel, still-life, fashion, and urban exploration. All images are homogeneously scaled by truncating their long sides to 500 pixels. Images smaller than this size are not used. We remove some images that are of low quality or heavily watermarked. Lastly, we add the "RetargetMe" images and ended up with a data set of 13,584 images.

The CAIR methods work best on images with disposable content. These images typically include either smooth or some regularly textured areas, such as sky, water, or grass. Challenges are present when the input image contained either rich semantic contents, salient texts, or geometric structures that may be damaged during retargeting. We use this observation and photography theories [50] as bases to choose a set of attributes that can be mapped to the several major retargeting objectives (preserving content, structure, and aesthetics and preventing artifacts) and manually annotate collected images with these attributes. The selected attributes are *people and faces*, *lines and/or clear boundaries*, salient *single object*, salient *multiple objects*, *diagonal composition*, *texture*, *repeating patterns*, specific *geometric structures*, *perspective*, *fuzzy*, *text*, *shading contrast*, *content rich*, and *symmetry*. Fig. 2 shows some examples in our data set with the attributes assigned to each image. Fig. 3 shows the correlation between image attributes. Apart from some attributes with opposite meaning (e.g., a single object versus multiple objects), the majority of the numbers in Fig. 3 contain relatively low absolute values, thereby demonstrating that most attributes are uncorrelated to one another.

B. Retargeting Methods and Annotations

To evaluate the retargetability of the collected images in the dataset, we select and implement the four most typical and commonly used CAIR methods, namely, multi-operators, inhomogeneous warping, shift-map, and cropping. We apply these four methods to all the collected images in the dataset.

Fig. 2. Example images in our dataset with manually annotated attributes.

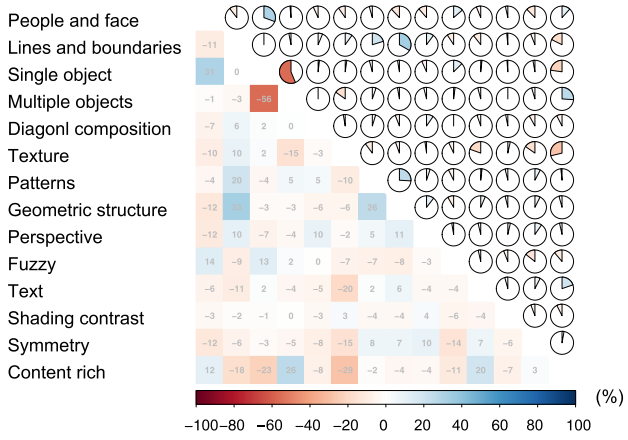


Fig. 3. Correlation among the visual attributes.

- *Multi-operator* method outperforms the majority of the other approaches according to the comparative study [32]. A typical multi-operator method integrates seam carving, homogeneous scaling, and cropping to resize an image and can be considered a generalized version of seam carving. Our study adopts the fast multi-operator method [22], which is sufficiently rapid for practical applications.
- *Inhomogeneous warping*-based method is known for its real-time performance and local continuity preservation. We use the AAD method [4], which has been recently verified to be one of the most effective warping methods. Other state-of-the-art warping-based methods [5], [18], [19] can also be used as the representative method, which does not affect the effectiveness of our retargetability learning and prediction framework.
- *Shift-map*-based method can selectively stitch some contents together and often works well for input with salient contents distributed in the different parts of the image. Our study applies the original shift-map method [3] to the images in the data set.
- *Cropping*-based CAIR algorithm is preferred in many cases because this type of method does not introduce any distortion in the retargeting results [32]. In particular, we use the SOAT_{cr} method [6].

Discussions: We did not add summarization-based retargeting methods, such as BDS [23] and PM [24], during data collection for two reasons. This method typically requires several minutes to generate a good result, while the results of BDS/PM often present a structural mismatch of spatial content [25]. These two artifacts limit their practical use in many applications, particularly for some systems that require real-time performance. We did not integrate some other CAIR methods that focus on images containing specific contents, such as symmetry structure [51], semantically-rich information [7], and textures [25], because our goal was to evaluate image retargetability in a generic manner.

Fig. 4. Example of the importance map used in our framework.

Although the selected methods are not recently proposed, they represent the mainstream CAIR framework. The recently proposed deep learning-based methods are an extension of these methods. We choose these classic approaches because they have been widely tested and proven to be stable and effective.

Given that the majority of the CAIR methods are carefully designed for one-dimensional retargeting, we restrict the change to either the width or height of an image. For each image in the data set, we resized the long dimension to 50% using the four CAIR methods described in Section III-B, all with fixed parameters. In particular, we choose to retarget the images to half their size, which is similar to the methods performed in previous research on CAIR, because the majority of the images can be handled well for small-sized changes, while causing poor results for large changes. We further guided the CAIR methods by computing an importance map for each image. We adopt state-of-the-art saliency detection approaches [52], face segmentation [53] and body detector [54] to generate the importance map. Note that the output of a body detector is the bounding box of the body region. We use GrabCut [55] to generate the importance map when a body is detected. We use the average of these maps as our importance map (see Fig. 4).

We ask six expert raters to independently evaluate the quality of all the retargeted images and annotate the result as one of the following three levels (see Fig. 5): *good*, *acceptable*, and *poor*, which correspond to scores of 1, 0.5 and, 0, respectively. Thereafter, we compute the average score from the six raters as the evaluation of each retargeting result in the dataset.

Consistency Analysis of Annotations: We measure the inter-rater consistency to verify the objectivity of the annotation data. For each image in our data set, we adopt Kendall's coefficient of concordance (Kendall's W) [56] to study the rating consistency among different subjects. Kendall's W is a non-parametric statistical measure that ranges from 0 (no consistency) to 1 (completely consistent). The overall average Kendall's W is 0.562 with a standard deviation of 0.0192. Moreover, raters can obtain significant concordance on 87.69% images at the 0.05 significance level.

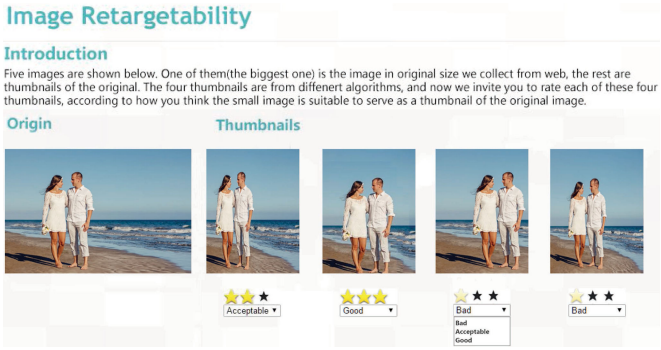


Fig. 5. Our web interface for data annotation. Expects were invited to rate each of the retargeted images on the basis of their own opinion. No further information on the definition of retargetability is released.

IV. Modeling Retargetability

First, we use our dataset as basis to analyze the correlation between image attributes and retargetability (Section IV-A). Second, we introduce our framework to employ deep learning and multi-task learning to learn and predict image retargetability (Section IV-B).

A. Measuring Retargetability

For each sample image in the dataset, we define its retargetability as the max score of the four average user-rated scores (each image possesses four retargeted outputs, while each output image contains six user-rated scores; see Section III).

Given this quantitative measurement of retargetability, we can analyze the relationship between different visual attributes and the image retargetability using Ridit analysis [57], which is commonly used in the study of ordered categorical data. In Fig. 7, the dashed horizontal line is the reference unit 0.5 and the deviation from reference unit represents the influence of the attribute. Evidently, some visual attributes are closely related to image retargetability. For example, the groups of lines, text, symmetry, geometry, and patterns are under the dashed line, thereby indicating that images with these attributes are likely to exhibit low retargetability scores or equivalently, worse retargeting results annotated by our raters. Images with content rich, diagonal structure, and texture often correspond to high scores. We use this key observation as basis to propose to learn and predict retargetability on the basis of visual attributes.

B. Learning and Predicting Retargetability

Although the retargetability of an image is calculated by the ratings of the retargeted images, retargetability itself is a high-level property of such an image. Thus, we aim to learn retargetability directly from the source image rather than the retargeted images by utilizing the pre-selected attributes to regularize a pair-wise retargetability ranking training. Fig. 6 shows the overall structure of our model, including a three-level feature representation mechanism and two types of loss functions that correspond to binary visual features or relative retargetability. First, we use the output from deep convolutional network as the low-level representation of the image. Second, we learn attribute-specific features for each attribute and eventually use this information to learn retargetability. To boost training phase, we simultaneously learn the visual attribute features with retargetability.

In the following section, we demonstrate the multi-task learning approach by jointly learning visual attributes from deep features and feature sharing for retargetability.

1) Deep Features: We use a VGG-19 [58] style model pre-trained on ImageNet [59] for image classification to extract deep representations for input images. The network consists of a stack of convolution layers with pooling and ReLU, followed by three fully-connected layers and softmax with loss. After isotropically re-scaling the input image's short side to 224, we densely crop the image to obtain 224^2 sub-images and feed the square sub-images to the convolutional network as [58]. The need to re-scale or crop the image to the same size for the learning appears to defeat the purpose of studying the effect of changing the aspect ratio of an image. However, compared with image retargetability, other visual attributes are more robust to the size change of the input image. That is, if an image contains a face, then the face will continue to exist even if the aspect ratio of the image has been changed. We denote Fm_i as the last convolutional layer's output of i_{th} sub-image. The low-level deep feature of an input is as follows:

$$Fm = \frac{\sum_{i=1}^{K} Fm_i}{K},$$

where K is the number of sub-images and is set to 10 in our implementation. We use the output of the convolution layers instead of the fully connected layers to acquire the low-level image representation. Hence, the output is not considerably related to the pre-trained classification task.

2) Learning Retargetability: We attempt to learn the middle-level features for visual attributes and share these features for retargetability to boost learning performance. In our task, all visual attributes are labeled as 1 or -1, in which binary attributes are often learned using the classification method. We formulate retargetability as a type of relative attribute that is powerful in uniquely identifying an image and offer a semantically meaningful method to describe and compare images in the wild [49]. We design different losses for different types of attributes. Given a low-level deep feature space together with annotated attributes and retargetability-labeled image data, we learn a shared attribute-level feature representation by optimizing a joint loss function that favors a pair-wise relative loss and squared hinge loss function with sparsity patterns across binary attributes.

Binary attribute features learning: Given M semantic attributes, the goal is to learn M binary classifiers jointly. Each binary classifier is a four-layer NN with one input layer and two hidden layers of 4096 and 1000 nodes, respectively, and a one-node output layer, followed by squared hinge loss function. Inspired by [60], [61], we utilize $l_{2,1}$-norm minimization to boost feature sharing among the different attributes. Multi-task feature learning via $l_{2,1}$-norm regularization has been studied in

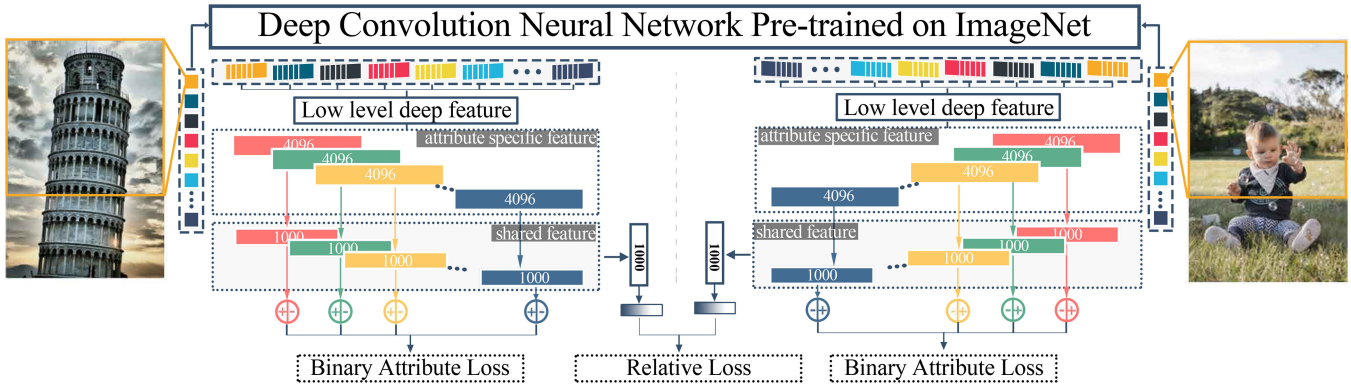


Fig. 6. Overall structure of our method. A siamese network with a three-level feature representation mechanism and two types of loss function corresponding to binary visual features or relative retargetability was adopted for retargetability learning.

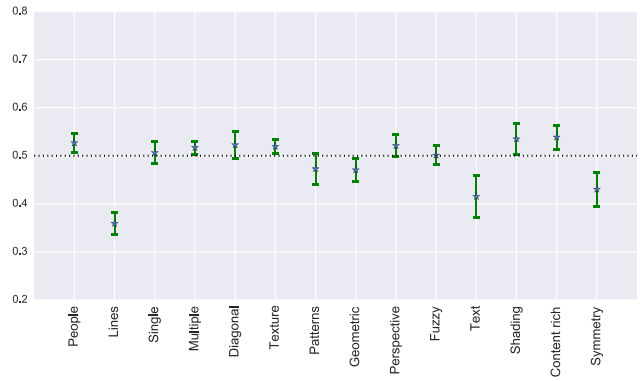


Fig. 7. Ridit scores with 95% confidence interval for visual attributes.

many approaches and encourages multiple predictors from different tasks to share similar parameter sparsity patterns. Given an image i, with a M-dimension label vector L_i element that is only 1 or -1, we proposed the following equation by supposing that the parameters between the two hidden layers for the k_{th} attribute learning MLP server as w_k:

$$loss_{binary}(i) = \sum_{k=1}^{M} \frac{1}{2}[\max(0, 1 - L_{ik} \cdot L^*_{ik})]^2 + \frac{1}{2}\alpha\|W\|_{2,1},$$

where $W = [w_k]$ and L^*_{ik} is the output of the k_{th} MLP for image i and $\|W\|_{2,1} = \sum \|w_k\|$ is the $l_{2,1}$-norm of the matrix W. We apply $l_{2,1}$-norm to W, thereby indicating that the outputs of the first hidden layer in different MLPs are relatively independent (see "attribute specific feature" in Fig. 6). By contrast, the outputs of the last hidden layer are boosted by multi-task feature learning technique (see "shared feature" in Fig. 6).

Relative retargetability learning: In general, the goal of relative attribute learning is to learn ranking functions for labeled image pairs. The existing relative attribute learning approaches learn linear functions to map hand-crafted features to relative scores. Inspired by [62], we collect features learned for each visual attribute as mid-level visual features and use these attribute-related features to train retargetability by a three-layer NN with 1000 hidden nodes. All the shared features are concatenated as the input of the three-layer NN. We define the relative loss as the sum of the contrastive and similar constraints. Given a pair of images i and j ($i \neq j$), with retargetability y_i and y_j predicted as y^*_i and y^*_j, the loss for the image pair (i, j) is as follows:

$$loss_{relative}(i, j) = I(i, j) \cdot l_p(i, j) + (1 - I(i, j)) \cdot l_q(i, j),$$

where

$$I(i, j) = \begin{cases} 1, & y_i > y_j \\ 0, & y_i \sim y_j \end{cases}$$

$$l_p(i, j) = \max(0, \tau - (y^*_i - y^*_j)),$$

$$l_q(i, j) = \frac{1}{2}(y^*_i - y^*_j)^2,$$

where $I(i, j)$ is a binary function that indicates whether images i and j exhibit similar retargetability, $l_p(i, j)$ denotes the contrastive constraint for ordered image pair (i, j), and $l_q(i, j)$ donates similar constraint for unordered pairs. The parameter τ controls the relative margin among the attribute values when $I(i, j) = 1$.

Formulations and implementations: Given the pair-wise relative loss, we use a two-channel Siamese network as the overall structure [63]. Each channel of the network predicts 14 visual attributes and retargetability together with the attribute-specific features. The hinge-based binary loss is calculated among each group of the M attributes while relative loss is computed by two predicted retargetability. The goal of the entire two-channel network is as follows:

$$\min_{\Theta} J_{\Theta} = \sum_{i,j}^{i \neq j} loss_{binary}(i) + loss_{binary}(j) + loss_{relative}(i, j) + \beta\|\Theta\|_F,$$

where Θ stands for all the parameters to be optimized and $\|\Theta\|_F$ is a regression term to penalize overfitting. Table I summarizes the configuration of the proposed method. Note that the majority of the parameters come from the MLPs for binary attribute learning. Due to these MLPs are trained in a multi-task learning

TABLE I
NET CONFIGURATION FOR PROPOSED ARCHITECTURE

Net Configuration (One-way)				
CNN part	Structure			Params
	224*224 Input RGB Image			[n, k, h, w]
	conv3-64*2			[64, 3, 3, 3], [64, 64, 3, 3]
	pooling			-
	conv3-128*2			[128, 64, 3, 3], [128, 128, 3, 3]
	pooling			-
	conv3-256*4			[256, 128, 3, 3], [256, 256, 3, 3]*3
	pooling			-
	conv3-512*4			[512, 256, 3, 3], [512, 512, 3, 3]*3
	pooling			-
	conv3-512*4			[512, 512, 3, 3]*4
	pooling			-
	4096	...	4096	≈1028M
	1000	...	1000	≈41M
MLP part	Binary Arrtibute * 10		1000 Retarget-ability	Total≈1180M

manner, the parameters are learned "separately". Given N images for training, we could obtain $14 \times N$ independent labels for training these MLPs separately. Each attribute is trained on one single small model. Different attributes are further boosted by feature sharing technique. For calculating the relative loss, we randomly draw a pair of samples from these N images. Totally we could draw $\mathcal{C}_N^2$ sample pairs for training. These advantages can possibly train the big network on the proposed data set. We adopt mini-batch stochastic gradient descent with a batch-size of 64 and an initial learning rate of 0.01. For the full connected layers, we adopt ReLU as the activation function. In the training stage, we randomly drop out 30% parameters to push the network to learn additional general features. In the test stage, we use one way of the siamese network to generate the outputs and truncate the predicted value into $[0, 1]$.

V. EVALUATIONS

This section presents the evaluation and analytical results of the image retargetability prediction. All experiments are performed on a PC equipped with 3.6 GHZ Intel Core i7 and Nvidia Geforce GTX 1080Ti. The implementation is based on Caffe2. The proposed network is trained for approximately 6 hours. Given the densely cropping operation, the testing speed is approximately 0.3 fps, which is slower than the current state-of-the-art CAIR method. Song *et al.* [31] report that their model takes about 0.5 s to process 100 images. However, the current study did not focus on the testing speed. Accordingly, the speed can be highly improved by taking the model quantization or distillation techniques as [31]. We randomly select half of the annotated images to train the retargetability predictor and use the remainder for testing. This process is performed five times and the average results are as follows.

A. Experimental Settings

Given that our research is the first study of image retargetability, finding direct comparisons with any previous study is difficult. To demonstrate the effectiveness of our framework, we compare our framework with the following CNN structures:

TABLE II
CONFIGURATIONS FOR CONTRAST METHODS

Method	Dense crop	Binary attribute	$l_{2,1}$ loss	Relative loss
Net_{-}	-	-	-	-
Net_{+}	✓	-	-	✓
Net_{*}	✓	✓	-	✓
$Net_{\&}$	-	✓	✓	✓
$Net_{@}$	✓	✓	✓	-
$Ours$	✓	✓	✓	✓

TABLE III
RMSE FOR DIFFERENT NET STRUCTURES

Method	Net_{-}	Net_{+}	Net_{*}	$Net_{\&}$	$Net_{@}$	Ours
RMSE	0.334	0.296	0.248	0.246	0.228	0.209

- Net_{-}. A straightforward end-to-end VGG19, which is fine-tuned on the training data to directly solve the regression problem.
- Net_{+}. A siamese network without binary attribute loss. All the other configurations are the same as the proposed approach, including the relative loss and low level feature extraction.
- Net_{*}. A siamese network without $l_{2,1}$ normalization in binary attribute loss. In the proposed method, we utilize $l_{2,1}$ to boost the feature sharing among different binary attributes. Net_{+} is tested to evaluate the performance of $l_{2,1}$ normalization.
- $Net_{\&}$. A siamese network without dense cropping in low level feature extraction. We re-scale the input image's short side to 224 and cropped K sub-images to calculate low-level deep representation. By contrast, $Net_{\&}$ directly re-scales the input image to 224^2.
- $Net_{@}$. A one-way network without relative loss. All the other configurations are the same as the proposed approach, including the binary attribute loss and low-level feature extraction.

Table II shows the details of these configurations.

B. Qualitative Analysis

Figs. 8 and 9 show the predicted retargetability and corresponding retargeted output, respectively, for several input images. The quality of the retargeted images is consistent with the predicted retargetability score. The results indicate that images with large homogeneous regions, blurry background, and single object lead to increased scores. By contrast, low scores are caused by several factors, including salient lines, clear boundaries, geometric structures, and symmetry.

C. Quantitative Analysis

We use root-mean-square error (RMSE) as measurement to evaluate the accuracy of our retargetability prediction approach. Assuming N images in the testing set are present, the overall $\text{RMSE} = \sqrt{\frac{1}{N}\sum_N (y - y^*)^2}$.

Table III shows the results. Accordingly, provide the following observation.

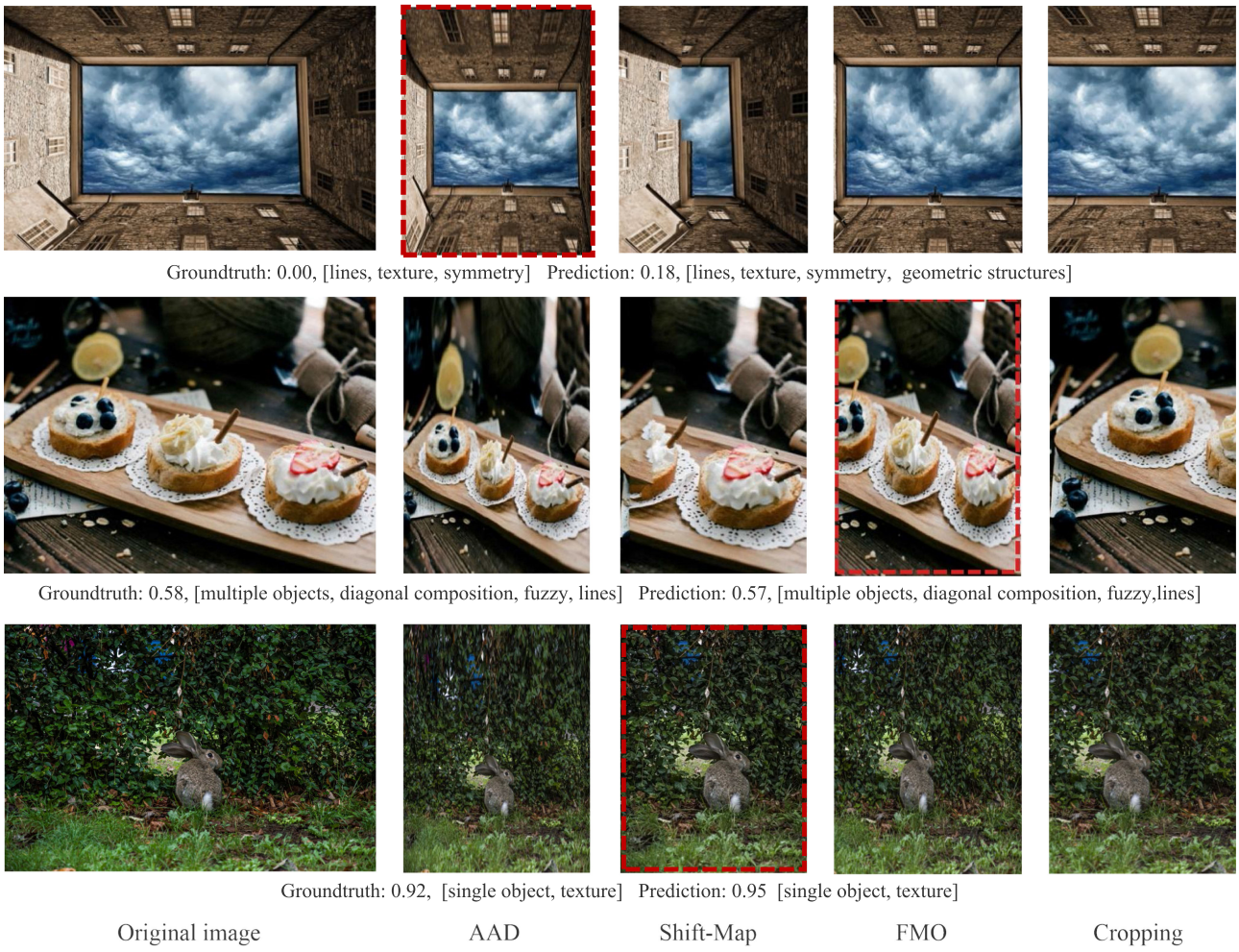


Fig. 8. Examples for retargetability prediction, visual attributes prediction, and method selection. In each example, the input image and the results of the four CAIR methods are shown on the left and right respectively. The visual attribute is predicted by the output of MLP for binary attribute learning: if $L_k^* > 0.5$, then the image is labeled as the k_{th} attribute. The results of the predicted best method are highlighted with red dash lines.

- As a baseline approach, Net_- reported the largest RMSE, thereby indicating that the proposed image property and retargetability can not be well-learned using traditional deep convolutional network.
- The RMSE improvement between Net_+ and Net_* demonstrates the model benefits from the representative ability of the extracted features by joint learning with binary attributes.
- Dense cropping in low level feature extraction promotes the performance of the feature learning process, which can be proven by the comparisons between Net_--Net_+, $Net_{\&}$-Ours. Such observations confirm that retargetability is a property dealing with the ability to be resized, retargetable operations to the original images may cause uncertain results. In our pipeline, retargetability is learned on features related to visual attributes which are considerably insensitive to the input size changing.
- The proposed model reported the lowest RMSE by embedding all the losses thereby confirming that sharing visual knowledge with high-level image attributes in the predictive model is a compelling method for boosting the learning process. Compared with dense cropping or binary attribute, the improvement using relative loss is not definite because the proposed relative loss tends to rank the images according to their retargetability. In training phase, our model is more likely to "compare" images with a pair-wise loss rather than learning absolute scores.

D. Discussion of Definition

In building the data set, we use the max rating (MAX-De) of the four methods from the six raters as the measurement of image retargetability. One alternative method is to use the mean value, MEAN-De. We trained the proposed siamese network together with Net_- under the definition of MEAN-De. RMSE of MEAN-De is as follows: ours $= 0.27$ and $Net_- = 0.42$. We label the testing samples as 1 or -1 according to its ground truth retargetability: 1 if the score is above than a threshold σ, and -1 otherwise. Thereafter, the regression task can be evaluated as a binary classification task. In Fig. 10, we plot the receiver operating characteristic curve and report the area under curve (AUC) value by setting σ to 0.95 and 0.7, respectively. Due

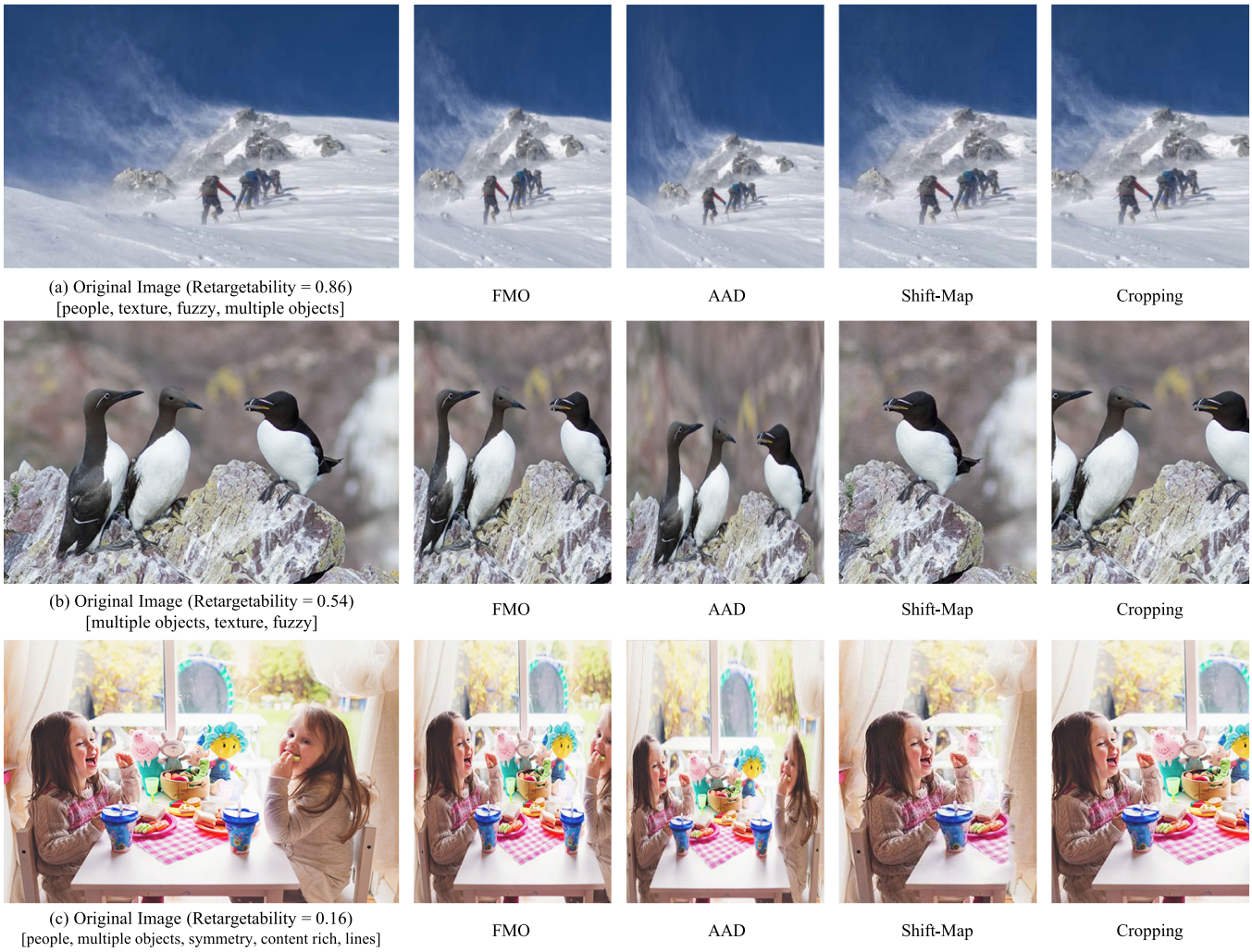
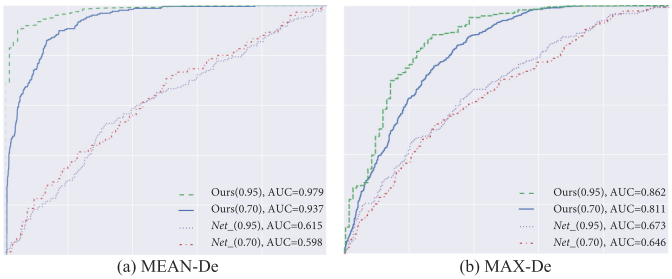
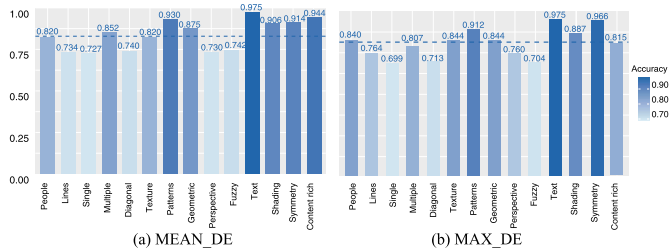


Fig. 9. Images with different retargetability scores and the corresponding results from four selected CAIR methods. (b) and (c) are more reliable for assessing new retargeting methods for these images are difficult for existing CAIR methods.



(a) MEAN-De (b) MAX-De

Fig. 10. Comparison of the retargetability prediction accuracy between our method and the baseline approach.



(a) MEAN_DE (b) MAX_DE

Fig. 11. Accuracy rate for attribute prediction.

to the plot of ROC curve is based on the prediction of each test samples, we could not repeat this result by randomly splitting the data set five times. For definition discussion, we randomly select 2000 testing samples from one of the five testing processes. Although the AUC value of MEAN-De is higher than that of MAX-De, RMSE of MAX-De is lower than that of MEAN-De.

Fig. 11 reports the results of the binary visual attribute learning. Additional discussions about the definition are provided in the supplemental materials.

VI. Applications

This section shows several applications of the proposed method and the dataset, including retargeting method selection, retargeting method assessment, and generating photo collage.

0.26 0.16 0.57 0.13 0.69

(a) Images with low or moderate retargetability scores, which means that most or all existing retargeting method cannot generate good results for them.

0.88 0.96 0.85 0.98 0.83

(b) Images with high retargetability scores, which means that most or all existing retargeting method can generate good results for them.

Fig. 12. Images with different retargetability scores. The images in the first row are reliable for assessing new retargeting method.

A. Joint Representations for Retargeting Tasks

This study proposes a unified framework for joint learning visual attributes and image retargetability. Traditional CAIR (assessment) works highly depend on hand-crafted features. However, image retargetability is based on the features learned by a proposed end-to-end siamese network. The outputs of the last hidden full connected layer in the retargetability branch embed the image visual attributes and features related to the CAIR tasks. These joint representations offer insight into the possible connections between CAIR research and deep learning approaches.

We adopt the learned image representations for another CAIR task: *retargeting method selection*, thereby suggesting the "best" retargeting method for a given image. First, we collect images that have either "good" or "acceptable" retargeting results in the training set and record the "best" method(s) of each image based on manually annotations (Section III-B). Second, we train the SVM classifier [64] for each CAIR method to learn whether an image can be well retargeted by the method. The inputs of these classifiers are the learned representations by our approach. During testing, the method with the highest predicted value are suggested as the "best" retargeting method for a given image. The average precision of the "best" retargeting method classification task is 82.12% on the testing set. Fig. 8 shows three such examples, in which the results of the suggested methods and results of other methods are compared. We conduct a user study to evaluate the perceived quality from an observer's perspective. In accordance with the experimental setting by [32], we adopt the paired comparisons technique, in which the participants are shown the original image and two retargeted images side by side. One of retargeted images is the "best" result predicted by our classifiers and the other is randomly chosen from the four retargeted results. The subjects are asked to compare the two results and choose the one they like better. A third option called "comparable" is offered when participants find no marked difference between the two results. A total of 300 images from the testing set are selected for the user study.

TABLE IV
STATISTICS FOR USER STUDY

Option	Adaptive Selection	Random Selection	Comparable	Vigilance
Counts	13, 164	3, 110	4, 786	2, 340
%	56.26%	13.29%	20.45%	10.00%

During the survey, we set up a vigilance comparison every 10 tests, in which the two retargeted images are the same. The results are discarded if one subject fails 50% of the vigilance comparisons. The vigilance comparisons ensure that workers are focusing, thereby leading to high-quality results. A total of 82 participants (age range of 20–45) from different backgrounds are involved. Among which, 95.12% results are valid and we obtain 23,400 votes. Table IV shows the statistics. The ratio of the participants' selection between "adaptive selection" and "random selection" was $56.26\% : 13.29\% \approx 4.23$, which is consistent with that of the quantitative analysis ($82.12\% : 17.88\% \approx 4.59$). The quantitative analysis and user study show that "adaptive selection" is superior to random guessing. This finding is due to the fact that all the CAIR methods exhibit their own philosophies and each one worked better than the others for some images. This result necessitates an adaptive selection for the "best" retargeting method for a given image example.

B. Retargeting Method Assessment

Although the CAIR methods have recently drawn considerable attention, the most popular assessment benchmark, namely, "RetargetMe" [32], was introduced approximately 10 years ago. The annotated dataset is relatively small and the current state-of-the-art CAIR methods report near-perfect results on this dataset. The current study offers a relatively large image dataset together with retargeting annotations, which can be used to augment other datasets, such as "RetargetMe" [32]. With the help of retargetability, people can easily collect a suitable testing set,

(a) Input photos with various retargetability from low to high

(b) Our photo collage result by considering retargetability

(c) Photo collage result without considering retargetability

Fig. 13. Example of photo collage generation. We generated photo collage by using retargetability to guide the placement of photos. Our result preserved more salient content and presentd less retargeting artifact compared with the result without considering retargetability.

which contains a wide range of images with different retargeting difficulties, to help the assessment of retargeting methods. This scheme has also been used in image retrieval method evaluation by organizing the evaluation set to three different levels of difficulties [65]. To evaluate if a new proposed CAIR method is effective, the new method must be tested on images which are difficult to existing methods. Therefore, during assessment, we can just use images with low or moderate retargetability scores, such as the images in Figs. 9(b), 9(c) and 12(a). People can use retargetability to filter some examples, which can be well retargeted by exiting methods such as the image in Figs. 9(a) and 12(b). Through our experiments, we find that the images with retargetability arranging between $(0.0, 0.75]$ are reliable for the new CAIR method assessment.

C. Photo Collage Generation

Photo collage is often created by placing multiple photo images on a canvas of limited size. The input images can be fitted on the canvas by retargeting them at the risk of losing important visual information and making the collage dull. Therefore, optimally selecting image examples for different sizes of canvas regions is important because any input image cannot merely be well retargeted in a given scale. Given the notion of image retargetability, automatic photo collage can be reliable We present an example of using image retargetability to guide the generation of photo collage (see Fig. 13). The origin photos with varying retargetability are shown in Fig. 13(a). With the help of retargetability, the collage can be created in a simple but effective manner. We first sort all the images based on their retargetability and place the images thereafter into the canvas in increasing retargetability order. Images with relatively low retargetability are preferentially placed into regions where the aspect ratio can be retained to the maximum extent. Fig. 13(b) shows the result of our collage generation using this strategy. Without considering retargetability, the collage may result in Fig. 13(c), which causes severe content loss or boundary discontinuity artifacts

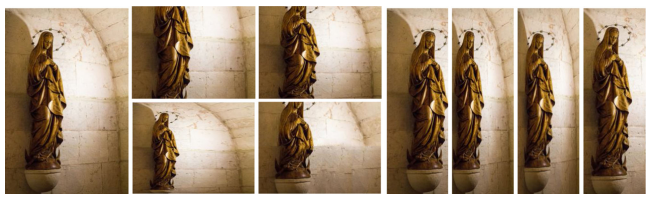


Fig. 14. Retargeting along different dimensions. The input image is shown on the left, the retargeting results along the long and short dimension using the four selected CAIR methods are shown in the middle and right respectively.

(a) Original Image (b) 75% (c) 50% (d) 25%

Fig. 15. Retarget one image to different scales. Targeting scale affects quality.

to the photos (see the images in the left-top and right-bottom corners of Fig. 13(c)). We observe that considering retargetability can preserve salient content and present less unnatural retargeting artifact.

VII. Conclusion and Future Work

This research presents retargetability as a novel image property and develops a computational predictor on the basis of multi-task learning. We construct a large image data set and annotate the retargetability of each image according to the quality of its retargeted results. We propose a siamese network structure that jointly learns attribute features and the relative retargetability. Our experiments show that image retargetability can be learned and predicted computationally and can be used to adaptively select a retargeting method for an image, find feasible image samples for retargeting method evaluation, and optimize collage layout for graphic design.

Our experiments only consider the retargetability of an image in one dimension (i.e., long side), thereby indicating the we restrict the change to either the width or height of the image. However, the retargetability of an image on these two dimensions may not consistently be similar. Fig. 14 shows that when we retarget the long side, the resulting images may not be as satisfactory as retargeting the short side. Therefore, we eventually plan to investigate the computation of image retargetability in both dimensions. Another limitation of our method is that we only retargeted source images to a fixed scale (50%), but the retargetability of an image may vary with the changing of target scale. Fig. 15 shows that we retarget one image to 75%, 50% and 25% and we could see that the quality of the retargeting results is related to the targeting scale.

We can augment the resulting images and annotate them to analyze the relationship between retargetability and target size in the future. Furthermore, we will attempt to generalize retargetability for analysis and processing of video data.

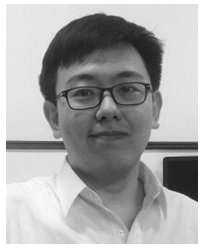
Acknowledgment

The authors thank the anonymous reviewers for valuable comments.

References

[1] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Trans. Graph.*, vol. 26, no. 3, pp. 10:1–10:10, 2007.
[2] M. Rubinstein, A. Shamir, and S. Avidan, "Multi-operator media retargeting," *ACM Trans. Graph.*, vol. 28, no. 3, pp. 23:1–23:12, 2009.
[3] Y. Pritch, E. Kav-Venaki, and S. Peleg, "Shift-map image editing," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 151–158.
[4] D. Panozzo, O. Weber, and O. Sorkine, "Robust image retargeting via axis-aligned deformation," *Comput. Graph. Forum*, vol. 31, no. 2, pp. 229–236, 2012.
[5] S.-S. Lin, I.-C. Yeh, C.-H. Lin, and T.-Y. Lee, "Patch-based image warping for content-aware retargeting," *IEEE Trans. Multimedia*, vol. 15, no. 2, pp. 359–368, Feb. 2013.
[6] J. Sun and H. Ling, "Scale and object aware image thumbnailing," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 135–153, 2013.
[7] L. Zhang *et al.*,"Retargeting semantically-rich photos," *IEEE Trans. Multimedia*, vol. 17, no. 9, pp. 1538–1549, Sep. 2015.
[8] Y.-S. Wang, C.-L. Tai, O. Sorkine, and T.-Y. Lee, "Optimized scale-and-stretch for image resizing," *ACM Trans. Graph.*, vol. 27, no. 5, pp. 118:1–118:8, 2008.
[9] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. V. Gool, "The interestingness of images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1633–1640.
[10] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva, "What makes a photograph memorable?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1469–1482, Jul. 2014.
[11] D. Dai, H. Riemenschneider, and L. V. Gool, "The synthesizability of texture examples," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3027–3034.
[12] H. Yang and I. Patras, "Mirror, mirror on the wall, tell me, is the error small?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4685–4693.
[13] J. Yan, S. Lin, S. B. Kang, and X. Tang, "Learning the change for automatic image cropping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 971–978.
[14] L. Zhang *et al.*, "Weakly supervised photo cropping," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 94–107, Jan. 2014.
[15] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 16:1–16:10, 2008.
[16] L. Wolf, M. Guttmann, and D. Cohen-Or, "Non-homogeneous content-driven video-retargeting," in *Proc. 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–6.
[17] Y. F. Zhang, S. M. Hu, and R. R. Martin, "Shrinkability maps for content-aware video resizing," *Comput. Graph. Forum*, vol. 27, no. 7, pp. 1797–1804, 2008.
[18] P. Krähenbühl, M. Lang, A. Hornung, and M. Gross, "A system for retargeting of streaming video," *ACM Trans. Graph.*, vol. 28, no. 5, pp. 126:1–126:10, 2009.
[19] P. Kaufmann *et al.*, "Finite element image warping," *Comput. Graph. Forum*, vol. 32, no. 2pt1, pp. 31–39, 2013.
[20] W. Tan, B. Yan, K. Li, and Q. Tian, "Image retargeting for preserving robust local feature: Application to mobile visual search," *IEEE Trans. Multimedia*, vol. 18, no. 1, pp. 128–137, Jan. 2016.
[21] W. Dong, N. Zhou, J.-C. Paul, and X. Zhang, "Optimized image resizing using seam carving and scaling," *ACM Trans. Graph.*, vol. 28, no. 5, 2009, Art. no. 125.
[22] W. Dong, G. Bao, X. Zhang, and J.-C. Paul, "Fast multi-operator image resizing and evaluation," *J. Comput. Sci. Technol.*, vol. 27, no. 1, pp. 121–134, 2012.
[23] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani, "Summarizing visual data using bidirectional similarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[24] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: a randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, 2009, Art. no. 24.
[25] W. Dong *et al.*, "Image retargeting by texture-aware synthesis," *IEEE Trans. Visualization Comput. Graph.*, vol. 22, no. 2, pp. 1088–1101, Feb. 2016.
[26] Y.-L. Chen, J. Klopp, M. Sun, S.-Y. Chien, and K.-L. Ma, "Learning to compose with professional photographs on the web," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 37–45.
[27] W. Wang and J. Shen, "Deep cropping via attention box prediction and aesthetics assessment," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2205–2213.
[28] D. Li, H. Wu, J. Zhang, and K. Huang, "A2-RL: Aesthetics aware reinforcement learning for automatic image cropping," in *Proc. IEEE Comput. Vis. Pattern Recognit. Conf.*, 2018, pp. 8193–8201.
[29] H. Chen, B. Wang, T. Pan, L. Zhou, and H. Zeng, "Cropnet: Real-time thumbnailing," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 81–89.

[30] G. Guo, H. Wang, C. Shen, Y. Yan, and H. Y. M. Liao, "Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2073–2085, Aug. 2018.
[31] Y. Song *et al.*, "Photo squarization by deep multi-operator retargeting," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 1047–1055.
[32] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir, "A comparative study of image retargeting," *ACM Trans. Graph.*, vol. 29, no. 6, 2010, Art. no. 160.
[33] Y.-J. Liu, X. Luo, Y.-M. Xuan, W.-F. Chen, and X.-L. Fu, "Image retargeting quality assessment," in *Comput. Graph. Forum*, vol. 30, no. 2, pp. 583–592, 2011.
[34] L. Ma, W. Lin, C. Deng, and K. N. Ngan, "Image retargeting quality assessment: A study of subjective scores and objective metrics," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 626–639, Oct. 2012.
[35] J. Zhang and C.-C. J. Kuo, "An objective quality of experience (QoE) assessment index for retargeted images," in *Proc. 22nd ACM Int. Conf. Multimedia.*, 2014, pp. 257–266.
[36] Y. Fang *et al.*, "Objective quality assessment for image retargeting based on structural similarity," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 4, no. 1, pp. 95–105, Mar. 2014.
[37] C.-C. Hsu, C.-W. Lin, Y. Fang, and W. Lin, "Objective quality assessment for image retargeting based on perceptual geometric distortion and information loss," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 3, pp. 377–389, Jun. 2014.
[38] B. Bare, K. Li, W. Wang, and B. Yan, "Learning to assess image retargeting," in *Proc. 22nd ACM Int. Conf. Multimedia.*, 2014, pp. 925–928.
[39] Y. Wang *et al.*, "Where2Stand: A human position recommendation system for souvenir photography," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 1, Oct. 2015, Art. no. 9.
[40] Y. Zhang, Y. Fang, W. Lin, X. Zhang, and L. Li, "Backward registration based aspect ratio similarity (ARS) for image retargeting quality assessment." *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4286–4297, Sep. 2016.
[41] Y. Liang, Y.-J. Liu, and D. Gutierrez, "Objective quality prediction of image retargeting algorithms," *IEEE Trans. Visualization Comput. Graph.*, vol. 23, no. 2, pp. 1099–1110, Feb. 2017.
[42] S. Castillo, T. Judd, and D. Gutierrez, "Using eye-tracking to assess different image retargeting methods," in *Proc. ACM SIGGRAPH Symp. Appl. Perception Graph. Visualization*. 2011, pp. 7–14.
[43] Y. S. Rawat, M. Song, and M. S. Kankanhalli, "A spring-electric graph model for socialized group photography," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 754–766, Mar. 2018.
[44] R. Rosenholtz, Y. Li, and L. Nakano, "Measuring visual clutter," *J. Vis.*, vol. 7, no. 2, pp. 17:1–17:22, 2007.
[45] A. Khosla, A. Das Sarma, and R. Hamid, "What makes an image popular?" in *Proc. 23rd Int. Conf. World Wide Web.*, 2014, pp. 867–876.
[46] C. Amati, N. J. Mitra, and T. Weyrich, "A study of image colourfulness," in *Proc. Workshop Comput. Aesthetics.*, 2014, pp. 23–31.
[47] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rating image aesthetics using deep learning," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2021–2034, Nov. 2015.
[48] M. Jas and D. Parikh, "Image specificity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2727–2736.
[49] X. Yang *et al.*, "Deep relative attributes," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1832–1842, Sep. 2016.
[50] L. Yao, P. Suryanarayan, M. Qiao, J. Z. Wang, and J. Li, "OSCAR: On-site composition and aesthetics feedback through exemplars for photographers," *Int. J. Comput. Vis.*, vol. 96, no. 3, pp. 353–383, 2012.
[51] H. Wu *et al.*, "Resizing by symmetry-summarization," *ACM Trans. Graph.*, vol. 29, no. 6, 2010, Art. no. 159.
[52] M. M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S. M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.
[53] S. Saito, T. Li, and H. Li, "Real-time facial segmentation and performance capture from rgb input," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 244–261.
[54] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2015.
[55] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive foreground extraction using iterated graph cuts," in *Proc. ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.
[56] M. G. Kendall and B. B. Smith, "The problem of m rankings," *Ann. Math. Stat.*, vol. 10, no. 3, pp. 275–287, 1939.
[57] J. L. Devore, *Probability and Statistics for Engineering and the Sciences*. Boston, MA, USA: Cengage Learning, 2015.
[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Comput. Sci.*, 2014.
[59] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[60] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient $l_{2,1}$-norm minimization," in *Proc. 25th Conf. Uncertainty Artif. Intell.*, 2009, pp. 339–348.
[61] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-task CNN model for attribute prediction," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1949–1959, Nov. 2015.
[62] S. J. Hwang, F. Sha, and K. Grauman, "Sharing features between objects and their attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1761–1768.
[63] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4353–4361.
[64] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, 2011, Art. no. 27, [Online]. Available: http://www.csie.ntu.edu.tw/∼cjlin/libsvm
[65] F. Radenovic, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Revisiting oxford and paris: Large-scale image retrieval benchmarking," in *Proc. IEEE Comput. Vis. Pattern Recognit. Conf.*, 2018, pp. 5706–5715.



Fan Tang received the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2019. He received the B.Sc. degree in computer science from North China Electric Power University, Beijing, China, in 2013. He is currently a Postdoctoral Scholar with Institute of Software, Chinese Academy of Sciences. His research interests include image synthesis and image recognition.

Weiming Dong received the the B.Eng. and M.Eng. degrees in computer science in 2001 and 2004, respectively, both from Tsinghua University, Beijing, China. He received his Ph.D. in computer science from the University of Lorraine, France, in 2007. He is a Professor in the Sino-European Lab in Computer Science, Automation and Applied Mathematics (LIAMA) and National Laboratory of Pattern Recognition (NLPR) at the Institute of Automation, Chinese Academy of Sciences. His research interests include image synthesis and image recognition. Dr. Dong is a member of the ACM.

Yiping Meng received the B.Sc. degree in software engineering from the University of Electronic Science and Technology of China, Chengdu, India, in 2013. She received the M.Eng. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2017. She is currently a Research Outreach Manager with Didi Chuxing. Her research interests include image synthesis and image recognition.

Chongyang Ma received the B.S. degree from the Fundamental Science Class (Mathematics and Physics) of Tsinghua University, Beijing, China, in 2007 and the Ph.D. degree in computer science from the Institute for Advanced Study of Tsinghua University, in 2012. He is currently a Research Lead at Kuaishou Technology. His research interests include computer graphics and computer vision.

Fuzhang Wu received the B.Sc. in computer science from Northeast Normal University, China, in 2010. He received the Ph.D. candidate from the Sino-European Lab in computer science, automation and applied mathematics and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include image synthesis and image analysis. He is currently a Postdoctoral Scholar with the Institute of Software, Chinese Academy of Sciences.

Xinrui Li received the M.Sc. degree from North China Electric Power University, in 2019. She received the B.Ec. degree in statistics from Shandong University of Finance and Economics, China, in 2016. She is currently a Researcher at the State Power Investment Corp. Research Institute. Her research interests include applied statistics and mathematical statistics.

Tong-Yee Lee received the Ph.D. degree in computer engineering from Washington State University, Pullman, USA, in May 1995. He is currently a Chair Professor in the Department of Computer Science and Information Engineering, National Cheng-Kung University, Tainan, Taiwan. He leads the Computer Graphics Group, Visual System Laboratory, National Cheng-Kung University, Tainan City, Taiwan. His current research interests include computer graphics, non-photorealistic rendering, medical visualization, virtual reality, and media resizing. Prof. Lee is a member of the ACM.