Composing Semantic Collage for Image Retargeting

Si Liu, Zhen Wei, Yao Sun, Xinyu Ou, Junyu Lin, Bin Liu, and Ming-Hsuan Yang, *Senior Member, IEEE*

***Abstract*— Image retargeting has been applied to display images of any size via devices with various resolutions (e.g., cell phone and TV monitors). To fit an image with the target resolution, certain unimportant regions need to be deleted or distorted, and the key problem is to determine the importance of each pixel. Existing methods predict pixel-wise importance in a bottom-up manner via eye fixation estimation or saliency detection. In contrast, the proposed algorithm estimates the pixel-wise importance based on a top-down criterion where the target image maintains the semantic meaning of the original image. To this end, several semantic components corresponding to foreground objects, action contexts, and background regions are extracted. The semantic component maps are integrated by a classification guided fusion network. Specifically, the deep network classifies the original image as object or scene oriented, and fuses the semantic component maps according to classification results. The network output, referred to as the semantic collage with the same size as the original image, is then fed into any existing optimization method to generate the target image. Extensive experiments are carried out on the *RetargetMe* data set and *S-Retarget* database developed in this paper. Experimental results demonstrate the merits of the proposed algorithm over the state-of-the-art image retargeting methods.**

***Index Terms*— Image retargeting, semantic component, semantic collage, classification guided fusion network.**

I. Introduction

IMAGE retargeting is a widely studied problem that aims to display an original image of arbitrary size on a target device with different resolution by cropping and resizing. Considering a source image is essentially a carrier of visual information, we define the image retargeting problem as a task to generate the target image that preserves the semantic information of the original image. For example, the image in Figure 1 shows a boy kicks a ball on a pitch (sports field), which contains four semantic components including boy, kicking, ball and pitch. Based on the source image, four target images can be generated as shown in Figure 1. The first three target images are less informative as certain semantic components are missing. The last target image is the only one that preserves all four semantic components. Existing retargeting methods [1]–[4] operate based on an importance map which indicates pixel-wise importance. To generate a target image in Figure 1 that preserves semantics well, the pixels corresponding to semantic components, e.g., boy and ball, should have higher weights in the importance map such that these are preserved in the target image. In other words, an importance map needs to preserve semantics of the original image well.

Fig. 1. Motivation of the proposed algorithm. The original image shows a boy kicks a ball on the pitch. The first three target images are less informative as important semantic components are missing, e.g., ball is missing, main foreground object is distorted, and background region is not well retained. The semantic meaning of the original image is well preserved in the fourth target image generated by the proposed algorithm.

The proposed semantics preserving deep image retargeting (SP-DIR) algorithm consists of two main steps (see Figure 2).

A. Extracting Semantic Components

Three semantic components including foreground, action context and background are extracted from an image. For example, in the image of Figure 2, the boy and ball are foreground components, kick and pitch belong to the action context, and the rest is background. Semantic components are extracted by using the stage-of-the-art modules based on deep learning. Foreground objects can be localized via image parsing [5] and classification [6]. The action context can be localized by the classification method [7], and background regions are identified by the scene classification [8] method. Each semantic component is represented by a map.

Manuscript received July 17, 2017; revised January 3, 2018 and April 6, 2018; accepted April 27, 2018. Date of publication May 15, 2018; date of current version July 12, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant U1536203 and Grant 61572493, in part by the IIE Project under Grant Y6Z0021102 and Grant Y7Z0241102, in part by the Strategy Cooperation Project under Grant AQ-1701. The work of M.-H. Yang was supported in part by the NSF CAREER under Grant 1149783, in part by Adobe, and in part by NEC. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Dacheng Tao. *(Corresponding author: Yao Sun.)*

S. Liu is with the Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, Beijing 100191, China, and also with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China.

Z. Wei, Y. Sun, and J. Lin are with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China (e-mail: sunyao@iie.ac.cn).

X. Ou is with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China, and also with the School of Media and Information Engineering, Yunnan Open University, Kunming 650223, China.

B. Liu is with the Yi+ AI Laboratory, Moshanghua Tech Co., Ltd., Beijing 100080, China.

M.-H. Yang is with the School of Engineering, University of California at Merced, Merced, CA 95343 USA.

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/TIP.2018.2836313

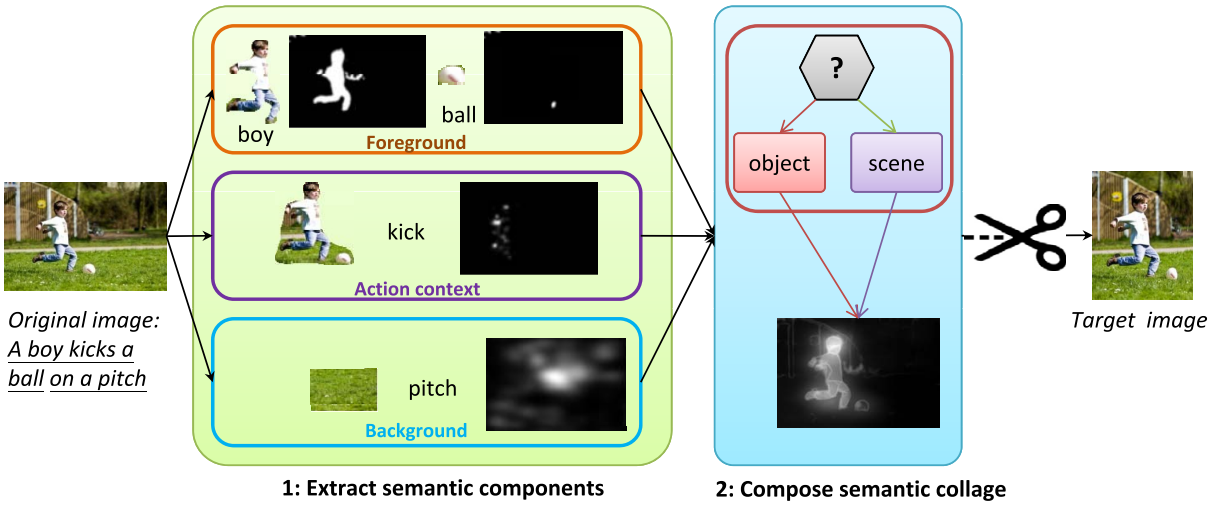


Fig. 2. Main steps of the SP-DIR algorithm. The semantic meaning of the original image is: a boy kicks a ball on a pitch. Three semantic components including boy, ball, kick and pitch are extracted first. These are fused via a classification guided fusion network to generate a semantic collage, which is fed into the carrier to render the target image.

B. Composing Semantic Collage

Although the state-of-the-art modules are used, semantic components may not be extracted well in an image. Thus, we combine all the semantic component maps via a classification guided fusion network to generate the semantic collage. As object and scene images have different properties [8], [9], the fusion network first classifies an image into two types. The semantic component maps are then fused by the corresponding sub-network based on the specified category. In contrast to existing methods, we exploit the semantic collages based on three defined components for image retargeting. The generated semantic collage is fed into a carrier method, e.g., [3], [4], [10], to generate the target image.

In this work, we make the following contributions:

- Different from existing retargeting methods, we propose to explicitly preserve semantics of the source image by first extracting multiple kinds of semantic components and then combining them automatically.
- We propose a classification guided fusion network to fuse the semantic component maps into a semantic collage with pixel-wise importance measures. In addition, object and scene images are considered differently.
- We develop a large *S-Retarget* dataset containing 1,527 images with pixel-wise labels. The dataset is one order larger than the existing dataset and available at http://www.spretarget.com.

II. Related Work

Numerous image retargeting methods have been developed including the scale and object aware thumbnailing (SOAT) [11], seam carving (ISC) [12], multi-operator [4], warp [1], optimized scale-and-streth (OSS) [2], shape-preserving [13] and any existing carrier (AAD) [3] schemes.

A. Conventional Image Retargeting

Early image retargeting methods are developed based on saliency detection that models the human eye fixation process [10], [14]–[17]. As these bottom-up methods are driven by low-level visual cues, edges and corners in images are detected rather than semantic regions. Although the thumbnailing method uses similar images in an annotated dataset to construct a saliency map for cropping [15], this task-driven approach does not exploit or preserve high-level visual semantics. In contrast, the proposed SP-DIR algorithm can better preserve semantic meanings for image retargeting. Other retargeting methods [18]–[20] crop images to improve visual quality of photographs [21], [22]. However, these schemes do not explicitly preserve visual semantics, which may discard important contents for the sake of visual quality and aesthetics.

B. Semantic-Based Image Retargeting

In recent years, more efforts have been made to analyze image contents for retargeting. Luo [14] detects a number of classes, e.g., skin, face, sky and grass, to crop photos. In [19] Yan *et al.* extend the foreground detection method of [23] with a human face detector to crop images. Similarly, Ding *et al.* [10] combine face detection results to compute importance maps for image retargeting. Goferman et al. [24] propose context-aware saliency which detects the important parts of the scene. On the other hand, Huang *et al.* [25] present a thumbnail generation scheme based on model expressiveness and recognizability of foreground objects after cropping. In contrast, the proposed SP-DIR algorithm considers semantics from objects to scenes at multiple scales as well as action contexts.

In [26], Jain *et al.* propose an end-to-end deep fully convolutional network for foreground object segmentation and apply it to image retargeting. Hou et al. [27] propose a saliency method by introducing short connections to the skip-layer structures within the HED architecture. The Fast-AT scheme [28] is recently developed for generating thumbnail images based on deep neural networks. Fan et al. [29] show that high-level semantic understanding is essential for saliency evaluation and propose the structure-measure to evaluate the

Fig. 3. Semantic foreground component maps constructed from image parsing. The input images and corresponding semantic component maps from parsing are shown in the first two columns. In the third column, the input images are overlaid with foreground component maps for visualization.

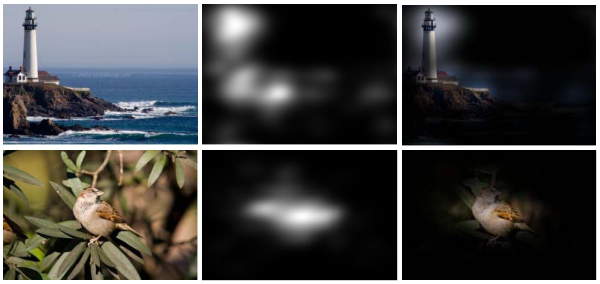


Fig. 4. Semantic foreground component maps constructed from image classification. The images are classified to contain the beacon and bird object classes, respectively. On each row, the input image, semantic component maps from classification and overlaid image are shown from left to right.

foreground maps. In [30] Cho et al. propose a weakly- and self-supervised deep CNN for content-aware image retargeting. This network generates a retargeted image directly from an input image and a target aspect ratio by learning a semantic map (attention map) implicitly. We note that these deep learning based methods predict binary maps whereas the SP-DIR algorithm predicts soft probabilities. Furthermore, these methods operates on the implicit assumption that each image contains one salient object. In contrast, we demonstrate that the proposed SP-DIR algorithm is able to generate target images containing multiple small objects in diverse scenes.

III. Composing Semantic Collage

In this section, we present the SP-DIR algorithm which extracts semantic components and composes a semantic collage for image retargeting. Each collage is fed into a carrier to generate the target image by removing or distorting less important pixels. The semantic collage can be combined with any carrier, i.e., AAD [3], multi-operator [4] and importance filtering (IF) [10].

A. Semantic Component

The semantic components including foreground, action context and background are extracted to describe an image for retargeting.

1) Semantic Foreground Components: The salient objects in an image are considered as the semantic foreground components. For example, the image in Figure 2 contains two main foreground components, i.e., boy and ball. We use the state-of-the-art image parsing and classification modules to locate foreground components.

a) Image parsing: We apply the pre-trained fully convolutional network [5] to parse each input image into 59 common categories defined in the Pascal-Context [31] dataset. The 59 categories, though still limited, include common objects that frequently occur in general images. We use all 59 parsing confidence maps where each semantic component map is denoted by M_p. As shown in Figure 3, the semantic component maps highlight the objects, i.e., person and building, well.

First, for concreteness we use 59 categories defined in the Pascal-Context dataset [31] to demonstrate the effectiveness of the proposed algorithm. While limited, they include common objects that frequently occur in general images. Second, several larger semantic segmentation datasets are released recently. For example, the ADE20K dataset contains 150 object and stuff classes with diverse annotations of scenes, objects, parts of objects, and in some cases even parts of parts. Third, it requires extensive manual labeling work to extend to a large number of categories, i.e., 3000 categories. One feasible approach is to resort to the weakly supervised semantic segmentation methods where bounding box [32] or image level annotations [33] are available.

b) Image classification: We use the VGG-16 network [34] pre-trained on the ILSVRC-2012 dataset to predict a label distribution over $1,000$ object categories in an image. As each classification is carried out on the image level, an importance map is obtained via a back propagation pass from the VGG network output [35]. The semantic component map induced by the classification output using 1-channel image is denoted by M_c. Figure 4 shows the support of the main objects, e.g., beacon and birds, can be visualized although they occupy small areas in the original image. The importance maps derived from classification are complementary to those induced from image parsing since more categories ($1,000$ vs. 59) are considered.

2) Action Context: We consider the action context surrounding the foreground objects for image retargeting. If there is no action in the scene, all pixels of the corresponding semantic component map are close to zero.

Action Recognition: Figure 2 shows an image where a boy *kicks* a ball. The action context in this scene is the kicking action between two objects (i.e., boy and ball). We train a deep model to classify 10 fine-grained actions in a way similar to the method by Oquab *et al.* [7]. The action recognition process is carried out on the detected the bounding box surrounding a human by the Faster R-CNN method [36], and the error back propagation is restricted inside the bounding box. Two representative examples of playing instruments are shown in Figure 5. In both examples, the action contexts with all involved objects are highlighted in the second column. The input images overlaid with the action contexts are shown in the third column. Given an image, the semantic component map derived from the action context is denoted as M_a.

3) Semantic Background Component: We consider the background component of the image for retargeting.

Fig. 5. Semantic component maps constructed from action contexts. On each row, the input image, semantic component map from the action context and input image overlaid with the map are shown.

Fig. 6. Semantic component maps constructed from scene classification. On each row, the input image, semantic component map from scene classification and overlaid image are shown.

Scene Classification: Figure 2 shows a scene containing a pitch. Scene classification provides holistic understanding of an image. We use the deep model [8], which is trained on the Places dataset with 2.5 million images to classify 205 categories. The semantic component map constructed from scene classification M_s is obtained similarly as M_c. As shown in Figure 6, two images are predicted as kitchen and island respectively. The obtained M_s highlights the most representative subjects that can explain the scene labels. In the kitchen scene, the hearth is highlighted. In the island scene, the rock and surrounding water are discovered. These results agree with the work by Zhou *et al.* [37] which shows that object detectors learned from the training process are responsible for scene classification. Thus, the semantic component map constructed from scene classification highlight regions of detected objects. In contrast, unimportant objects, such as lights on the ceiling, are ignored.

B. Semantic Collage

The semantic components introduced in Section III-A have several limitations. First, although the state-of-the-art deep modules are used, the semantic component maps may not be accurate. For example, the detection module are likely to generate false positives or negatives. Second, the context information between different semantic components is missing. For example, in Figure 2, the spatial relationship between boy and ball is missing in the individual semantic component maps. To address these issues, we propose a classification guided fusion network to integrate all component maps. While the importance maps have been used in the existing image retargeting methods, we emphasize the semantic collage in this work effectively preserves semantics and integrates multiple semantic component maps based on different cues.

1) Classification-Guided Fusion Network: It has been recently shown that object as well as scene images have drastically different properties and should be independently treated for the classification [8] or aesthetic evaluation [9] tasks. Motivated by these observations, our network explicitly classifies an image as either object-oriented or scene-oriented, and fused by separate weights.

Figure 7 shows the network architecture. The inputs of the network are 62-channel semantic component maps. The action, scene, classification maps all have 1 channel, while the segmentation map has 59 channels. In addition, the original images are also used in the CRF-RNN [38] modules. The concatenated semantic component maps are fed into a $128\times1\times1$ conv layer, followed by two sub-networks. First, the classification sub-network predicts the image as either scene-oriented or object-oriented. It contains a global average pooling layer and a fully connected layer for prediction. Second, the regression sub-network carries out convolutions in a way similar to the inception network [39]. The CRF-RNN modules are then added to smooth predictions. We fuse both convolutional maps and CRF-RNN output by pixel-wise average pooling to regress the target within the range of [0, 1]. The scene and object oriented maps from the regression sub-network are weighted by the classification results to generate the semantic collage. Note that in the regression sub-network, object-oriented and scene-oriented image have separate conv layers and CRF-RNN layers.

The semantic collage M_g is obtained by

$$M_g = c(o|M)\cdot r_o(M) + c(s|M)\cdot r_s(M) \tag{1}$$

where $M = \{M_p, M_c, M_s, M_a\}$ is the concatenation of all semantic component maps to be fused and contains 62 channels. In the above equation, $r_o(\cdot)$ and $r_s(\cdot)$ are regression functions for object-oriented and scene-oriented, respectively. In addition, $c(o)$ and $c(s)$ are the confidences that the image belongs to object or scene-oriented one. The semantic collage can be generated by a soft or hard fusion based on whether $c(\dot{})$ is the classification confidence or binary output.

2) Network Training: The training process involves 3 stages by increasingly optimizing more components of the network.

Stage 1: The classification sub-network is trained first as its results guide the regression sub-network. Here only the parameters related to the classification sub-network are updated. The loss function L_1 at this stage is a weighted multinomial logistic loss:

$$L_1 = \frac{1}{N}\sum_{i=1}^{N}\omega_i \log(\hat{\omega}_i) \tag{2}$$

where $\omega_i \in \{0, 1\}$ is ground truth classification label, $\hat{\omega}_i$ is the probability predicted by the classification sub-network, and N is the training set size.

Stage 2: We train both classification and regression sub-networks without CRF-RNN layers in this stage. The loss

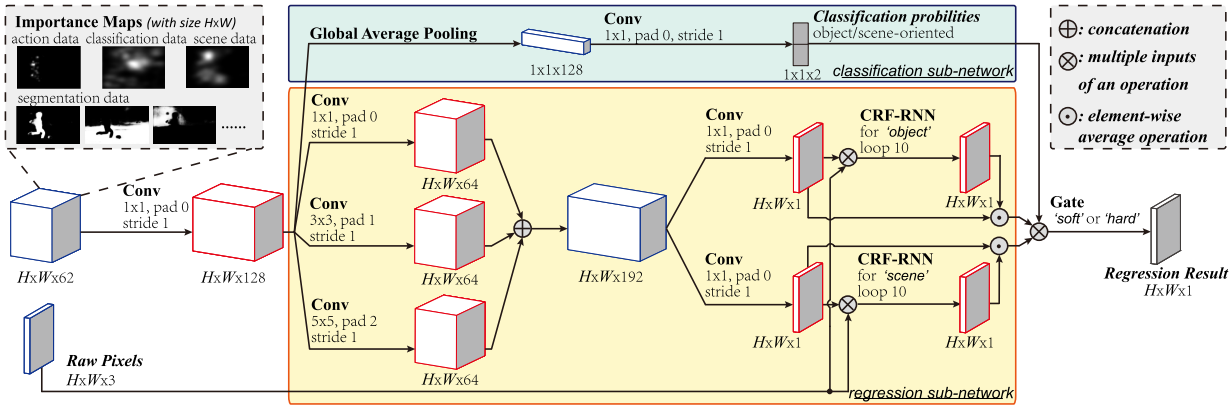


Fig. 7. Classification guided fusion network. The inputs are the multiple semantic component maps and the output is the semantic collage. The classification sub-network predicts the image as either scene or object oriented. Then the regression sub-network fuses the semantic component maps according to the classification results.

function L_2 is:

$$L_2 = L_1 + \frac{1}{N}\sum_{i=1}^{N}\sum_{x=1}^{W}\sum_{y=1}^{H}\|I_{i,x,y} - \hat{I}_{i,x,y}\|^2 \tag{3}$$

where I and $\hat{I}$ are the ground truth and estimated semantic collages. In addition, W and H are width and height of input image, respectively.

Stage 3: The CRF-RNN layers are activated. The loss function of this stage is the same as L_2.

IV. S-Retarget Dataset

Several retargeting benchmark datasets have been developed including the *RetargetMe* [40] dataset and the one collected by Mansfield *et al.* [41]. However, these datasets contains only 80 or 100 images. For comprehensive evaluation of image retargeting methods, we construct the Semantic-Retarget (*S-Retarget*) dataset which contains 1, 527 images.

A. Image Collection

We select images from the Pascal VOC [42], SUN [43], and BSR [44] datasets. In addition, we collect images from Google and Bing search engines. Based on the contents, all images are divided into 6 categories including single person, multiple people, single as well as multiple objects, and indoor as well as outdoor scenes. The images in single person, multiple people, single object and multiple objects classes are object-oriented while other images are scene-oriented. Table I shows the properties of the *S-Retarget* dataset. Some representative images are shown in Figure 8(a). The dataset is split into train/val/test subsets, containing 1, 237, 145 and 145 images respectively. The distribution of the 6 categories are almost the same in the three sets.

B. Semantic Collage

We ask 5 subjects to annotate the pixel relevance based on the semantics of an image. The labeling process consists of two stages. In the first stage, each subject annotates the caption

TABLE I

Properties of the *S-Retarget* Dataset

Classes	single person	multiple people	single object
No. images	252	265	247
Classes	multiple objects	indoor scene	outdoor scene
No. images	259	252	247

of an image. Several image captions are shown in Figure 8(b). In the second stage, the annotators rate all pixels by referring to the image caption provided in the first stage. To facilitate labeling, each image is over-segmented 5 times using multiple over-segmentations methods including SLIC [45] 3 times and Quick Shift [46] twice with different segmentation parameters, e.g., number of superpixels and compactness factors. Each annotator then assigns a value to each image segment where a higher score corresponds to high relevance. For ease of labeling, the relevance score can only take three discrete values, i.e., 1 is closed to foreground, 0 is closed to background and 0.5 is in the middle. The relevance score of one pixel is obtained by averaging all relevance scores of the segments covering the pixel. Figure 8(b) shows the semantic collage marked by two annotators. The semantic collages of the 5 annotators are averaged to generate the ground truth maps, which are shown in Figure 8(c).

The *S-Retarget* dataset can also be used as a semantic saliency dataset. Different with the saliency datasets, e.g., MSRA-5000 [47] or ECSSD [48], which mainly contain dominant objects, the images in *S-Retarget* are quite diverse. Furthermore, as shown in Figure 8(c), the ground truth are labeled with soft rather than binary annotations. The annotated dataset will be made available to the public.

V. Experimental Results

We conduct five series of experiments. First, we evaluate the effectiveness of key modules of the deep fusion network. Second, we compare our semantic collage with the state-of-the-art methods for generating importance maps. Third, we compare the target images generated by the SP-DIR algorithm

Fig. 8. Some semantic collages in the *S-Retarget* dataset. (a) original images. (b) annotations (including image captions and semantic collages) from two annotators (c) ground truth annotations obtained by averaging the semantic collages from 5 annotators.

with the ones generated by the state-of-the-art retargeting methods. We then report the results of applying the model trained on the *S-Retarget* images to the *RetargetMe* dataset. At last, we train the SP-DIR and MC method [16] on two different datasets: S-Retarget and ECSSD [48], and compare the retargeted results generated by these importance maps. This comparison aims to find out whether the new proposed S-Retarget dataset brings improvements for the retargeted results. For all experiments, the width of the target image is half of the original ones while the height remains unchanged. More results are available at www.spretarget.com.

A. Experimental Settings

1) Implementation Details: In the training process, we use 3×10^{-5} as learning rate in the first two stages and 3×10^{-6} in the last stage (see Section III-B.2).

2) Datasets and Baseline Methods: We carry out experiments on the RetaretMe [40] and *S-Retarget* datasets (see Section IV).

3) Evaluation Metric: We use the metrics of the MIT saliency benchmark dataset [49] for evaluation including the Earth Mover's Distance (EMD), Pearson linear coefficient (CC), Kullback-Leibler divergence (KL), histogram intersection (SIM), and mean absolute error (MAE). For EMD, KL, MAE, the lower the better while for CC and SIM, the higher the better. The other three metrics in the MIT saliency benchmark are not adopted as they require eye fixation as ground truth.

We carry out user study to evaluate the retargeted results from different methods using the Amazon mechanical turk (AMT). Each AMT worker is shown with two images, i.e., the target image generated by our method as well as the one generated by a randomly selected baseline method, and asked to select the preferred image. Each pair is shown to 3 different AMT workers, and we record the numbers of votes that prefer our results. The evaluation results are shown in the form of A(B) which means that in the total (A+B) comparisons, our method is preferred A times.

TABLE II
EFFECTIVENESS OF THE CLASSIFICATION GUIDANCE

Metrics / Models	Classification accuracy	EMD	CC	KL	SIM	MAE
SP-DIR w classification task	0.9110	**1.0115**	**0.6960**	**0.3004**	**0.7498**	**0.1890**
SP-DIR w/o classification task	–	1.0601	0.6613	0.3206	0.7300	0.2010

B. Ablation Study

We conduct ablation studies to evaluate different modules of the deep fusion network.

1) Classification Guidance Fusion Network: We remove classification subnetwork in the fusion network to evaluate its effectiveness. Specifically, the global pooling layer, classification output layer and switch are removed while two parallel regression branches are merged into a single one.

Table II shows classification guidance fusion network significantly helps generate better semantic collages in terms of all

TABLE III
FUSION NETWORK WITH A SOFT OR HARD SWITCH

Models \ Metrics	EMD	CC	KL	SIM	MAE
SP-DIR w soft switch	**1.0115**	**0.6960**	**0.3004**	**0.7498**	**0.1890**
SP-DIR w hard switch	1.0150	0.6933	0.3067	0.7484	0.1899

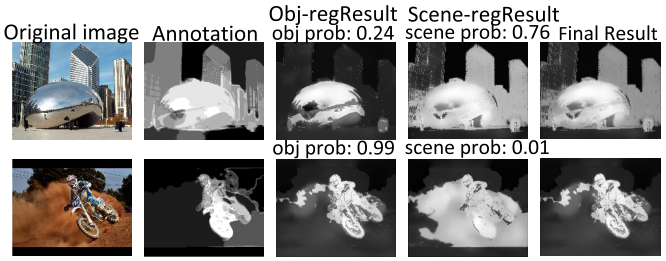


Fig. 9. Effectiveness of the classification-guided fusion network.

TABLE IV
FUSION NETWORK WITH ONE MODULE LEFT OUT

Models \ Metrics	Classification accuracy	EMD	CC	KL	SIM	MAE
SP-DIR	**0.9110**	**1.0115**	**0.6960**	**0.3004**	**0.7498**	**0.1890**
SP-DIR w/o segmentation data	0.6580	1.4610	0.3699	0.4638	0.6398	0.2891
SP-DIR w/o action data	0.9110	1.0354	0.6823	0.3098	0.7437	0.1925
SP-DIR w/o scene data	0.8970	1.0214	0.6773	0.3076	0.7453	0.1971
SP-DIR w/o classification data	0.9040	1.0669	0.6653	0.3136	0.7392	0.1979

evaluation metrics. This can be attributed to that scene-oriented and object-oriented images have different properties and thus require separate parameters to fuse the semantic components into collages. By binarizing the fusion weights, i.e., the classification confidence scores, the soft fusion becomes the hard decision. Table III shows that the soft combination, namely the linear combination of scene-oriented and object-oriented maps weighted by classification probabilities, performs better that the hard one where only one map is selected. Figure 9 shows the effectiveness of soft fusion. From the figure, we find that the final result (the last column) is better than the object-regression result (third column) and the scene-regression result (fourth column).

2) Semantic Components: The inputs of the fusion network are the semantic components generated by four deep modules including image parsing, classification, action recognition and scene classification. We analyze the role of each modules in a leave-one-out fashion. Table IV shows that the method by leaving out segmentation module performs worst as that it generates most details with clear boundary. The other three modules are also indispensable as the results are worse than the proposed SP-DIR algorithm. Figure 10 shows qualitative comparisons by discarding one module. For example, in the second row, without the segmentation module the semantic collage for the stalls are not well delineated. Similarly, a part of the balloon is missing in the third row when the classification module is not used.

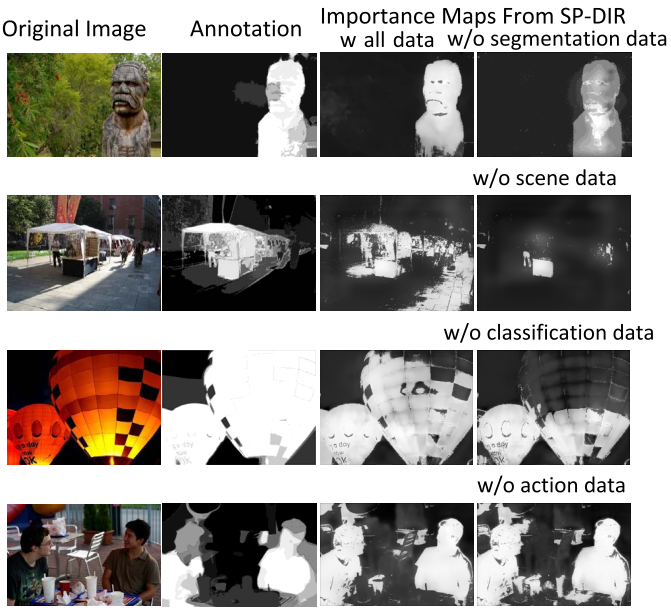


Fig. 10. Results from the whole SP-DIR network and the one with one module left out.

TABLE V
COMPARISONS BETWEEN OUR SEMANTIC COLLAGE AND 6 BASELINE MAPS WHEN COMBINED WITH 3 DIFFERENT CARRIERS

Baselines \ Classes		Single person	Multiple people	Single object	Multiple objects	Indoor scene	Outdoor scene	All
AAD	eDN	155(25)	140(40)	148(32)	151(29)	143(37)	141(39)	878(202)
	GC	143(37)	146(34)	146(34)	152(28)	143(37)	146(34)	876(204)
	oriIF	151(29)	134(46)	152(28)	154(26)	146(34)	138(42)	875(205)
	DNEF	154(26)	160(20)	163(17)	163(17)	166(14)	160(20)	966(114)
	RCC	152(28)	163(17)	149(31)	150(30)	147(33)	156(24)	917(163)
	MC	153(27)	160(20)	136(44)	152(28)	160(20)	156(24)	917(163)
Multi Op	eDN	163(17)	157(23)	150(30)	158(22)	152(28)	164(16)	944(136)
	GC	154(26)	162(18)	162(18)	163(17)	157(23)	161(19)	959(121)
	oriIF	166(14)	159(21)	155(25)	156(24)	145(35)	158(22)	939(141)
	DNEF	174(6)	173(7)	166(14)	161(19)	161(19)	169(11)	1004(76)
	RCC	134(46)	147(33)	144(36)	138(42)	146(34)	143(37)	852(228)
	MC	155(25)	163(17)	154(26)	149(31)	155(25)	163(17)	939(141)
IF	eDN	143(37)	145(35)	150(30)	147(33)	153(27)	148(32)	886(194)
	GC	161(19)	152(28)	154(26)	146(34)	147(33)	154(26)	914(166)
	oriIF	167(13)	150(30)	168(12)	164(16)	157(23)	156(24)	962(118)
	DNEF	172(8)	159(21)	170(10)	168(12)	161(19)	158(22)	988(92)
	RCC	142(38)	146(34)	149(31)	142(38)	149(31)	143(37)	871(209)
	MC	146(34)	142(38)	137(43)	153(27)	154(26)	162(18)	894(186)

3) Sensitivity Analysis: Each generated semantic collage is fed into a carrier to generate the target image by removing or distorting less important pixels. In this experiment, we randomly select 60 images from each subsets in the *S-Retarget* to evaluate the proposed semantic collage with 6 baseline importance map generation methods using 3 carriers, i.e., AAD [3], multi-operator [4] and importance filtering (IF) [10]. The baseline map generation methods and carriers are the same as discussed in Section V-A. The results of all 6 subsets are presented in Table V where we use AMT scores for evaluation. For the Single person subset, the semantic collage + AAD method is preferred by 155 persons while the eDN + AAD scheme is favored for 50 times. Overall, the proposed semantic collage performs favorably against all the baselines in all subsets.

TABLE VI

EVALUATION OF IMPORTANCE MAPS ON THE VALIDATION SET IN THE *S-RETARGET* DATASET

Models \ Metrics	EMD	CC	KL	SIM	MAE
SP-DIR	**1.0115**	**0.6960**	**0.3004**	**0.7498**	**0.1890**
fine-tuned MC-sc	1.0393	0.6202	0.5979	0.7374	0.2346
fine-tuned MC-mc	1.0412	0.6249	0.5819	0.7470	0.2296
MC-sc	1.0485	0.6256	0.8323	0.7171	0.2622
MC-mc	1.0695	0.6421	0.8987	0.7196	0.2446
fine-tuned DSS	1.1151	0.6869	0.3090	0.7026	0.1999
DSS	1.3986	0.6753	3.9898	0.6202	0.2040
eDN	1.4311	0.5056	0.4342	0.6555	0.3030
oriIF	1.4937	0.2594	0.7835	0.5789	0.3165
GC	1.5386	0.4167	2.9459	0.5482	0.2945
SalNet	1.6752	0.2338	0.5351	0.5963	0.3172
DNEF	1.8629	0.3113	1.6273	0.5344	0.3164
RCC	1.8809	0.4891	8.7734	0.4434	0.2759
Mr-CNN	2.1830	0.1372	0.9566	0.5489	0.3554

C. Semantic Collage Evaluations

We evaluate the proposed semantic collage algorithm with the state-of-the-art saliency based importance map methods, i.e., MC [16] (including MC-mc and MC-sc), GC [23], RCC [50], importance map used in the original IF (oriIF) [10], DSS [27], eDN [51], and DNEF [52]), Mr-CNN [53] and the SalNet [54]. In addition, we binarize the *S-Retarget* training dataset and fine-tune the two models in MC [16], which we denote as fine-tuned MC-mc and fine-tuned MC-sc respectively.

All evaluation results are presented in Table VI. Overall, the proposed semantic collage algorithm performs favorably on the *S-Retarget* dataset using all metrics. Figure 11 shows the semantic collage generated by the proposed SP-DIR algorithm and the other three importance maps using a typical indoor image. The MC-sc [16] and fine tuning MC-sc methods are able to distinguish foreground with background objects. The SalNet [54] scheme generates weak location information of salient objects and does not perform well. The semantic collage generated by the SP-DIR algorithm is close to the ground truth. With the extracted foreground and background semantic components by the proposed SP-DIR algorithm, the semantic collage (Figure 11(c)) contains all semantic information well.

D. Evaluation on the S-Retarget Dataset

We combine our importance map with the Importance Filtering carrier for the quality results and speed, and compare our SP-DIR (with map and carrier) with the state-of-the-art retargeting schemes including the SOAT [11], Seam Carving [12], Multi-Operator [4], Warp [1], OSS [2], and AAD [3] schemes. We note that the Multi-Op method is based on our implementation with the same parameters as [4]. For the other baseline methods, we use the released codes. The importance maps generated by the baseline methods are used in the evaluations on the *S-Retarget* dataset.

For each input, 7 target images are generated by the proposed SP-DIR algorithm and baseline retargeting methods. Table VII shows that the proposed algorithm performs well against all the baselines. For example, the number 2, 985(255) means the results by the proposed algorithm are preferred 2, 985 times while those generated by other methods are favored 255 times.

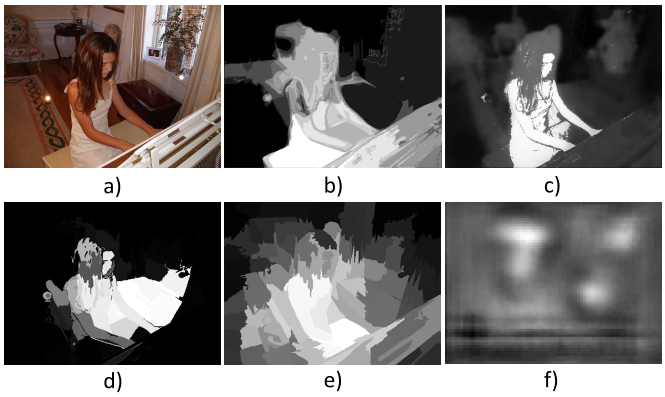


Fig. 11. Comparison of importance maps. (a) original image. (b) ground truth. (c) SP-DIR. (d) MC-sc. (e) fine-tuned MC-sc. (f) SalNet.

TABLE VII

COMPARISONS WITH STATE-OF-THE-ART RETARGETING SYSTEMS

	SOAT	ISC	Multi-Op	Warp	ADD	OSS
Our	2985(255)	2890(350)	2648(592)	3072(168)	2652(588)	2878(362)

Sample retargeted images by all evaluated methods are presented in Figure 12. The first row shows the results for single person subset. The retargeted images by the baseline methods do not capture the essence of the input image well. In the second example, the proposed semantic collage method preserves the semantics well without significant distortion. Similar observations can be observed in the single object and multiple objects results (3rd and 4th rows). Similarly for images containing multiple objects and outdoor scenes, the proposed SP-DIR algorithm performs favorably against the other state-of-the-art methods in which the important objects are preserved well in the retargeted results.

E. Evaluation on the RetargetMe Dataset

We apply the proposed model to the RetargeMe dataset [40] for evaluation. Note that in this experiments, no ground truth annotation of the semantic collage is available. We feed 7 importance maps (our semantic collage and 6 other importance maps, i.e., MC [16], GC [23], RCC [50], importance map used in the original IF [10], eDN [51], and DNEF [52]) into 3 carriers (AAD [3], Multi-Op [4], IF [10]). Table VIII shows the results where the rows are carriers and columns are types of importance maps. Overall, the proposed semantic collage method performs well against all the baseline importance maps. For example, when AAD is used as carrier, our semantic collage (with 204 votes) is favored more than the GC method (with 36 votes).

Qualitative results are shown in Figure 13. For the image on the first row, the retarget image generated by the GC method (3rd column) significantly distorts the players while the image generated by the RCC scheme (6th column) misses

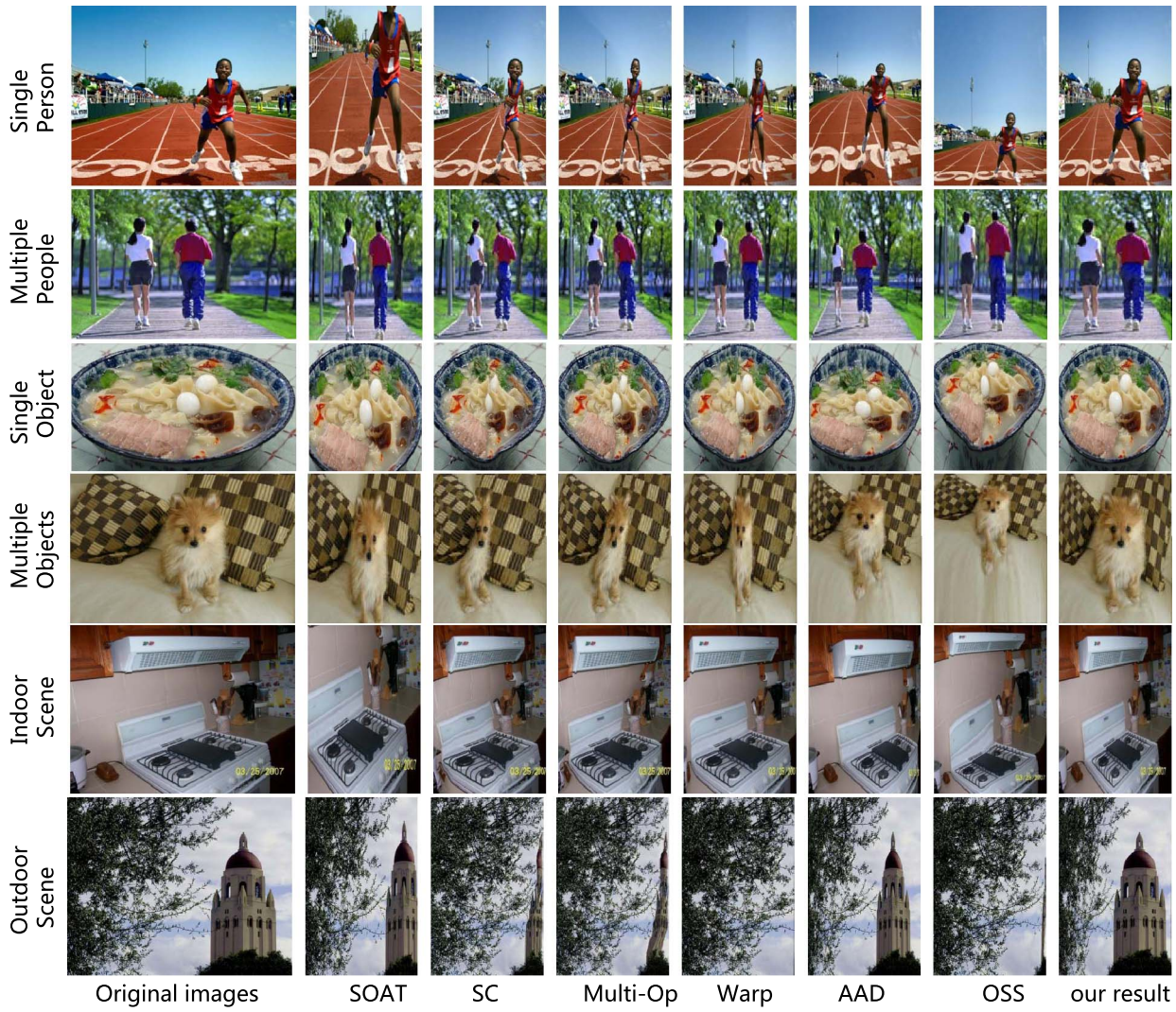


Fig. 12. Comparisons with SOAT, ISC, Multi-operator, Warp, AAD, OSS on the *S-Retarget* dataset.

TABLE VIII
COMPARISONS BETWEEN SP-DIR AND 6 BASELINE MAPS WHEN PAIRED WITH 3 CARRIERS

Map / Carrier	eDN	GC	oriIF	DNEF	RCC	MC
AAD	190(50)	204(36)	179(61)	207(33)	200(40)	190(50)
Multi-Op	222(18)	213(27)	218(22)	228(12)	189(51)	204(36)
IF	215(25)	200(40)	195(45)	213(27)	194(46)	201(39)

the ball. The retargeted image (last column) by the proposed algorithm contains all the important objects and contexts (players, actions and scenes). The image in the second row shows a typical outdoor scene with a brick house. The result by the eDN method (1st column) deforms the brick house significantly while the retargeted image by the proposed algorithm (last column) retains most essential contents. In the third image, the proposed algorithm performs well without distorting the subject.

F. Comparison Between S-Retarget and ECSSD

To further show the value of the new S-Retarget dataset, we compare the retarget results generated by the models trained on different datasets. Besides the proposed dataset, we also consider the ECSSD dataset [48]. To make fair comparisons, we use the following experimental settings. ECSSD is split into a training set and a test set with 900 and 100 images respectively. We also select 100 images from the test set of the S-Retarget dataset. The selected 200 images from both datasets (100 from each one) form an unbiased test set. Our SP-DIR model is trained both on the S-Retarget and ECSSD datasets, and then tested on the new unbiased test set. For clarification, we use ⟨training dataset⟩⟨salience method⟩ to denote different training dataset and salience method settings. Besides our SP-DIR method, we also test with a state-of-the-art saliency method, i.e., MC method [16]. Therefore, there are totally 4 different experiment settings including: $\langle ECSSD\rangle\langle SP\text{-}DIR\rangle$, $\langle S\text{-}Retarget\rangle\langle SP\text{-}DIR\rangle$, $\langle ECSSD\rangle\langle MC\rangle$, $\langle S\text{-}Retarget\rangle\langle MC\rangle$. In all experiments, we use IF as the retargeting method. The aim of this experiment is to study for a specific method SP-DIR or MC, how different training datasets affect the retargeted results.

The evaluation results on retargeted images using Amazon Mechanic Turks are shown in Table IX.

There are 200 images for testing. For each saliency method, we have 200 pairs of retargeted results obtained by training

Fig. 13. Results on the *RetargetMe* dataset by 3 retargeting methods (AAD, Multi-Op, and IF) and 7 importance maps (eDN, GC, oriIF, DNEF, RCC, MC, and our method).

TABLE IX

COMPARISONS BETWEEN THE RETARGETED RESULTS BY S-RETARGET AND ECSSD USING TWO SALIENCY METHODS (SP-DIR AND MC)

Salience + retarget method \ Training Datasets	S-Retarget	ECSSD
SP-DIR	382	218
MC	361	239

on both S-Retarget and ECSSD. Given each image pair, we invited 3 AMT workers to choose the better one. The results are reported in Table IX. Specifically, if the retargeted result generated by $\langle S\text{-}Retarget\rangle$ $\langle SP\text{-}DIR\rangle$ is better than the one by $\langle ECSSD\rangle$ $\langle SP\text{-}DIR\rangle$, then we add one vote at the Row (SP-DIR) and Col (S-Retarget) of the Table IX. From the table, we can see that for both SP-DIR and MC, the retargeted results from the models trained on S-Retarget are better than those trained on ECSSD. It indicates that the new proposed S-Retarget brings improvements on the retargeted results.

VI. CONCLUSIONS

In this paper, we propose a deep image retargeting algorithm that preserves the semantic meaning of the original image. A semantic collage that represents the semantic meaning carried by each pixel is generated in two steps. First, multiple individual semantic components, i.e., including foreground, contexts and background, are extracted by the state-of-the-art deep understanding modules. Second, all semantic component maps are combined via a classification guided fusion network to generate the semantic collage. The network first classifies the image as object or scene-oriented one. Different classes of images have their respective fusion parameters. The semantic collage is fed into the carrier to generate the target image. Our future work include exploring image caption methods [55] for better retargeting and related problems. In addition, we plan to integrate the PixelCNN [56] and GAN [57]–[59] modules to the proposed algorithm for retargeting tasks.

REFERENCES

[1] L. Wolf, M. Guttmann, and D. Cohen-Or, "Non-homogeneous content-driven video-retargeting," in *Proc. ICCV*, Oct. 2007, pp. 1–6.
[2] 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, 2008, Art. no. 118.
[3] D. Panozzo, O. Weber, and O. Sorkine, "Robust image retargeting via axis-aligned deformation," *EUROGRAPHICS*, vol. 31, no. 2, pp. 1–8, 2012.
[4] M. Rubinstein, A. Shamir, and S. Avidan, "Multi-operator media retargeting," *ACM Trans. Graph.*, vol. 28, no. 3, 2009, Art. no. 23.
[5] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, 2015, pp. 3431–3440.
[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.
[7] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. CVPR*, Jun. 2014, pp. 1717–1724.
[8] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. NIPS*, 2014, pp. 487–495.
[9] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 271–280.
[10] Y. Ding, J. Xiao, and J. Yu, "Importance filtering for image retargeting," in *Proc. CVPR*, Jun. 2011, pp. 89–96.
[11] J. Sun and H. Ling, "Scale and object aware image thumbnailing," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 135–153, 2013.
[12] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," *ACM Trans. Graph.*, vol. 27, no. 3, 2008, Art. no. 16.
[13] G.-X. Zhang, M.-M. Cheng, S.-M. Hu, and R. R. Martin, "A shape-preserving approach to image resizing," *Comput. Graph. Forum*, vol. 28, no. 7, pp. 1897–1906, 2009.

[14] J. Luo, "Subject content-based intelligent cropping of digital photos," in *Proc. ICME*, Jul. 2007, pp. 2218–2221.

[15] L. Marchesotti, C. Cifarelli, and G. Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in *Proc. ICCV*, Sep./Oct. 2009, pp. 2232–2239.

[16] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proc. CVPR*, Jun. 2015, pp. 1265–1274.

[17] J. Chen, G. Bai, S. Liang, and Z. Li, "Automatic image cropping: A computational complexity study," in *Proc. CVPR*, Jun. 2016, pp. 507–515.

[18] J. Yan, S. Lin, S. B. Kang, and X. Tang, "Change-based image cropping with exclusion and compositional features," *Int. J. Comput. Vis.*, vol. 114, no. 1, pp. 74–87, 2015.

[19] J. Yan, S. Lin, S. B. Kang, and X. Tang, "Learning the change for automatic image cropping," in *Proc. CVPR*, Jun. 2013, pp. 971–978.

[20] Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 80–106, Jul. 2017.

[21] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. (2016). "Photo aesthetics ranking network with attributes and content adaptation." [Online]. Available: https://arxiv.org/abs/1606.01621

[22] X. Lu, Z. Lin, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *Proc. ICCV*, Dec. 2015, pp. 990–998.

[23] 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.

[24] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.

[25] J. Huang, H. Chen, B. Wang, and S. Lin, "Automatic thumbnail generation based on visual representativeness and foreground recognizability," in *Proc. ICCV*, Dec. 2015, pp. 253–261.

[26] S. D. Jain, B. Xiong, and K. Grauman, "Pixel objectness," in *Proc. CVPR*, 2017, pp. 1–18.

[27] Q. Hou, M.-M. Cheng, X. W. Hu, A. Borji, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," in *Proc. CVPR*, 2017, pp. 5300–5309.

[28] S. A. Esmaeili, B. Singh, and L. S. Davis, "Fast-AT: Fast automatic thumbnail generation using deep neural networks," in *Proc. CVPR*, Jul. 2017, pp. 1–9.

[29] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. ICCV*, 2017, pp. 1–10.

[30] D. Cho, J. Park, T.-H. Oh, Y.-W. Tai, and I.-S. Kweon, "Weakly- and self-supervised learning for content-aware deep image retargeting," in *Proc. ICCV*, 2017, pp. 4568–4577.

[31] R. Mottaghi *et al.*, "The role of context for object detection and semantic segmentation in the wild," in *Proc. CVPR*, Jun. 2014, pp. 891–898.

[32] R. Hu, P. Dollár, K. He, T. Darrell, and R. Girshick. (2017). "Learning to segment every thing." [Online]. Available: https://arxiv.org/abs/1711.10370

[33] Q. Hou, P. K. Dokania, D. Massiceti, Y. Wei, M. M. Cheng, and P. Torr, "Bottom-up top-down cues for weakly-supervised semantic segmentation," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*. Venice, Italy: Springer, 2017.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. CVPR*, 2014, pp. 1–14.

[35] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. CVPR*, 2013, pp. 1–8.

[36] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. CVPR*, 2015, pp. 91–99.

[37] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene CNNs," in *Proc. CVPR*, 2014, pp. 1–12.

[38] S. Zheng *et al.*, "Conditional random fields as recurrent neural networks," in *Proc. ICCV*, 2015, pp. 1529–1537.

[39] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. CVPR*, 2016, pp. 1–9.

[40] 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.

[41] A. Mansfield, P. Gehler, L. Van Gool, and C. Rother, "Visibility maps for improving seam carving," in *Proc. ECCV*, 2012, pp. 131–144.

[42] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[43] J. Xiao, K. Ehinger, J. Hays, A. Torralba, and A. Oliva, "SUN database: Exploring a large collection of scene categories," *Int. J. Comput. Vis.*, vol. 119, no. 1, pp. 3–22, 2014.

[44] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. ICCV*, vol. 2, Jul. 2001, pp. 416–423.

[45] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[46] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 705–718.

[47] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *Proc. CVPR*, Jun. 2007, pp. 1–8.

[48] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. CVPR*, Jun. 2013, pp. 1155–1162.

[49] Z. Bylinskii *et al.*, "MIT saliency benchmark," Accessed: May 10, 2016. [Online]. Available: http://saliency.mit.edu/

[50] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *Proc. ICCV*, Dec. 2013, pp. 1529–1536.

[51] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. CVPR*, Jun. 2014, pp. 2798–2805.

[52] C. Shen, M. Song, and Q. Zhao, "Learning high-level concepts by training a deep network on eye fixations," in *Proc. NIPS*, 2012, pp. 1–9.

[53] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, "Predicting eye fixations using convolutional neural networks," in *Proc. CVPR*, Jun. 2015, pp. 362–370.

[54] J. Pan, K. McGuinness, E. Sayrol, N. O'Connor, and X. Gir-I-Nieto, "Shallow and deep convolutional networks for saliency prediction," in *Proc. CVPR*, 2016, pp. 598–606.

[55] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. CVPR*, 2015, pp. 3128–3137.

[56] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proc. CVPR*, 2016, pp. 1–11.

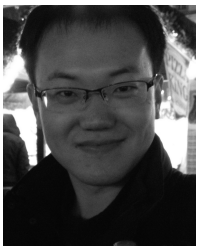
[57] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. CVPR*, 2015, pp. 1–16.

[58] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[59] M. Mirza and S. Osindero, "Conditional generative adversarial nets," in *Proc. CVPR*, 2014, pp. 1–7.

Si Liu received the Ph.D. degree from the Institute of Automation, CAS. She was a Research Fellow with the Learning and Vision Research Group, National University of Singapore. She is currently an Associate Professor with the School of Computer Science and Engineering, Beihang University. Her research interests include object categorization, object detection, image parsing, and human pose estimation.

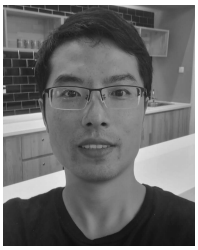
Zhen Wei received the B.S. degree in computer science and technology from the Yingcai Honors School, University of Electronic Science and Technology of China, Chengdu, China. He is currently pursuing the master's degree with the Institute of Information Engineering, Chinese Academy of Sciences.



Yao Sun received the Ph.D. degree from the Academy of the Mathematics and Systems Science, Chinese Academy of Sciences. He is currently an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences.

Xinyu Ou received the Ph.D. degree in computer science and technology from the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, in 2017. He was a Visiting Doctor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing. He is currently an Associate Professor with the College of Communication and Information Engineering, Yunnan Open University, Kunming. His research interests include deep learning, image retrieval, and object detection and recognition.

Junyu Lin is currently the Director Assistant of the Laboratory of Cyberspace Technology, Institute of Information Engineering, Chinese Academy of Sciences. He has more than 50 publications on *Peer-to-Peer Networking and Applications* and *Journal of Software* and IEEE conferences and journals. He is a member of the CCF YOCSEF Academic Committee and the CCF TCAPP Standing Committee. He is also a member of the CCF Council.



Bin Liu received the master's degree from the Department of Automation, Shanghai Jiao Tong University. He has a wide-ranging expertise in machine learning, computer vision, and high-performance computing. His recent research interests include use of deep learning methodology for modeling, classification, detection, and tracking; face recognition; and content-based image retrieval.

Ming-Hsuan Yang (SM'17) received the Ph.D. degree in computer science from the University of Illinois at Urbana–Champaign in 2000. He is currently a Professor in electrical engineering and computer science with the University of California at Merced, Merced. He is a Senior Member of the ACM. He received the Google Faculty Award in 2009 and the Distinguished Early Career Research Award and the Distinguished Research Award from the UC Merced Senate in 2011 and 2015, respectively. He was a recipient of the Faculty Early Career Development (CAREER) Award from the National Science Foundation in 2012.