A COMPREHENSIVE STUDY ON IMAGE RETARGETING IN COMPOSING SEMANTIC COLLAGE

PEDDI RENUKA

Assistant professor, Department of ECE, Indur Institute of Engineering and Technology,

Siddipet,Telangana, India.

Email id: renoostar@gmail.com

Abstract—

Certifiable applications could profit by the capacity to automatically re-focus on an image to various viewpoint proportions and resolutions, while protecting its visually and semantically critical substance. Nonetheless, not all images can be similarly all around prepared that way. There is an expanding necessity for productive image re-focusing on systems to adjust the substance to different types of advanced media. With fast development of versatile correspondences and dynamic site page formats, one regularly needs to resize the media substance to adjust to the coveted show sizes. For different formats of pages and ordinarily little sizes of handheld compact gadgets, the significance in the first image content gets jumbled in the wake of resizing it with the approach of uniform scaling. Along these lines, there happens a requirement for resizing the images in a substance mindful way which can automatically dispose of unimportant data from the image and present the remarkable highlights with more size. In this work, we present the idea of image retarget capacity to depict how well a particular image can be dealt with by content-mindful image retargeting The semantic component maps are incorporated by a classification guided fusion network. In particular, the profound network characterizes the first image as protest or scene-arranged, and melds the semantic component maps as indicated by classification comes about. The network yield alluded to as the semantic collage with an indistinguishable size from the first image, is then bolstered into any current advancement technique to create the objective image. Broad trials are completed on the Retarget Me dataset and S-Retarget database created in this work. Trial comes about show the benefits of the proposed calculation over the cutting edge image retargeting techniques.

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

I. INTRODUCTION

Image retargeting is a generally examined issue that intends to show a unique image of self-assertive size on an objective gadget with various determinations by trimming and resizing. Considering a source image is basically a transporter of visual data, we characterize the image retargeting issue as a task to produce the objective image that jam the semantic data of the first image. Fig. 1-1 that scaling operation renders the boat in the image too thin thereby obfuscating the contents of the boat; cropping results in clipping of the right part of the boat; content aware image resizing preserves the whole structure of the boat and provides a better visually pleasant retargeted image. The initial three target images are less educational as certain semantic components are absent. The last target image is the special case that jellies each of the four semantic components. Existing retargeting strategies [1], [2], [3], [4] work in light of significance delineate shows pixel-wise significance. To create an objective image in Figure 1 that jelly semantics well, the pixels comparing to semantic components, e.g., kid and ball ought to have higher weights in the significance guide to such an extent that these are protected in the objective image. As it were, significance delineates to protect semantics of the first image well.

Fig. 1-1 (From left to right) Original Image, Image Retargeted using scaling, Image retargeted using cropping, and Image retargeted using content aware image resizing. The retargeting is attempted to make the width 40% of the original width with no change in the original image height. Here, we have used our novel algorithm (presented in Chapter 4) for content aware resizing. [Image Courtesy - http://www.flickr.com/photos/ayushbhandari/1524259093/ (Sailors at St. Malo)]

It presently moves toward becoming clearer that substance mindful image resizing expects to retarget the images in the way where the critically unmistakable locales are held with slightest conceivable twisting, while the lesser huge areas are misshaped more. It may show up from Fig. 1-1 that one may likewise achieve the substance mindful retargeting result with physically movable editing and scaling; nonetheless, such an activity isn't generally conceivable given unbounded number of image structures and in such cases, a nonexclusive calculation for content mindful resizing is constantly helpful. This idea will turn out to be more obvious in ensuing sections where we dive into the theme.

With focusing on images on little show gadgets at the top of the priority list, some underlying endeavors were made for image and video retargeting on little show gadgets. Specialists in [2] - [9] have displayed different methods for formulating versatile image and video resizing strategies for portable/handheld gadgets. These techniques have focussed more on the image and video perusing on handheld gadgets with automatic zooming and panning in light of the perceived/client indicated area of intrigue.

It is significant here, that the utilization of the terms visual saliency, critical locales (areas of enthusiasm), being utilized as a part of the setting of substance mindful image retargeting are kind of conceptual terms and don't frame the meaning of some very much characterized set, since for various images, their definition from the eyewitnesses' perspective may change. For example, one may contend that the content composed on the vessel is huger than the pontoon itself subject to shifted applications. This very thought has now given the idea of substance mindful image resizing a novel application. Content mindful resizing techniques currently likewise discover their utilization when an image should be resized for an inserted object's/area's assurance. With proficient image preparing programming, for example, Adobe Photoshop CS5 [10], one regularly endeavors to secure or evacuate some client characterized region in an image. This requires an adjustment of the image retargeting calculation close by since the whole issue presently gets included with a client characterized imperative, and subsequently can be generally thought of as resizing diverse districts of the image obliged by non-rectangular or discretionary question limits, while additionally safeguarding image's visual soundness with a conceivably unique however important semantics.

The image retargeting issue begins with changing over an advanced image I of measurements m x n (m lines and n sections) to an objective image I' of measurements m' x n' (m' lines and n' segments). The measurements given here indicate the quantity of lines and sections in the image just, and we are not extremely worried about the RGB or the

dark scale nature of the image while retargeting. From a more specialized and a programmatic perspective, while a dark scale image will have measurements of m x n, a RGB image will have the relating measurements of m x n x 3; the factor 3 being there for the red (R), green (G), blue (B) components of the image. Be that as it may, for a substance mindful image retargeting issue, we are required to scale a dark scale image into a dim scale one and a RGB image into a RGB one, with change in the quantity of lines and sections of the image as it were.

Fig. 2 Formulation of the Image Retargeting Problem. A desired targeted size is generally (unless the approach is implemented as an interactive one) defined before input image is processed.

Fig. 2demonstrates diagrammatically the detailing of the image retargeting issue in a bland way. Here, client communication is discretionary since the image retargeting issue is frequently intended to be automatic; in any case, given the boundless number of image structures joined with the different applications, discretionary client cooperation includes greater adaptability. The term client association ordinarily has the accompanying implications.

A client cooperation towards the meaning of the significance delineate basically the particular by the client of the districts he/she sees as essential for their application or something else. Such a connection can either be through particularly specifying the inspecting focuses for the coveted imperative areas or through the determination of some flexible parameters gave by the interface. The movable parameters regularly are given in a way with the goal that their qualities draw an exchange off between the for the most part distinguishable striking locales and the for the most part recognizable non-remarkable areas of the image.

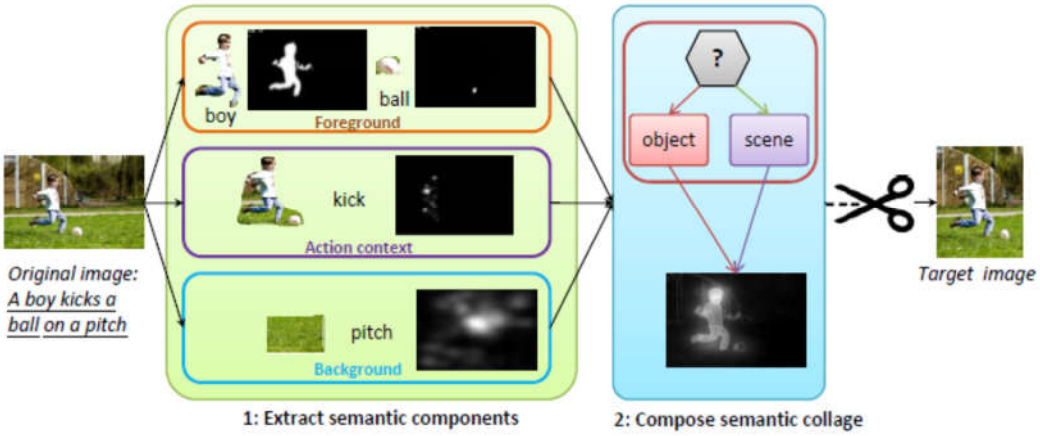

Fig. 3. Main steps of the SP-DIR algorithm. The semantic meaning of the original image is: a boy kicks a ball on a pitch. Three semantic components including boy, ball, kick and pitch are extracted first. These are fused via a classification guided fusion network to generate a semantic collage, which is fed into the carrier to render the target image.

The proposed semantics protecting profound image retargeting (SP-DIR) calculation comprises of two principle steps (see Figure 3).

Removing semantic components. Three semantic compo-nents including forefront, activity setting and foundation are separated from an image. For instance, in the image of Figure 3, the kid and ball are closer view components, kick and pitch have a place with the activity setting, and the rest is foundation. Semantic components are extricated by utilizing the phase of-the-workmanship modules in light of profound learning. Closer view articles can be confined through image parsing [5] and classification [6]. The activity setting can be confined by the classification strategy [7], and foundation districts are distinguished by the scene classification [8] technique. Each semantic component is spoken to by a guide.

Creating semantic collage. In spite of the fact that the best in class modules are utilized, semantic components may not be removed well in an image. In this way, we join all the semantic compo-nent maps by means of a classification guided fusion network to gen-erate the semantic collage. As protest and scene images have distinctive properties [8], [9], the fusion network initially characterizes an image into two sorts. The semantic component maps are then melded by the relating sub-network in light of the predetermined class. As opposed to existing techniques, we misuse the semantic collages in view of three characterized components for image retargeting. The produced semantic collage is bolstered into a bearer strategy, e.g., [3], [4], [10], to create the objective image.

II. RELATED WORK

Various image retargeting strategies have been created including the scale and protest mindful thumbnailing (SOAT) [11], crease cutting (ISC) [12], multi-administrator [4], twist [1], improved scale-and-stretch (OSS) [2], shape-safeguarding [13] and any current bearer (AAD) [3] plans.

A. Traditional Image Retargeting

Early image retargeting strategies are produced in light of saliency location that models the human eye obsession process [14], [10], [15], [16], [17]. As these base up strategies are driven by low-level visual prompts, edges and corners in images are identified as opposed to semantic districts. In spite of the fact that the thumb-nailing strategy utilizes comparative images in a commented on dataset to build a saliency outline editing [15], this task-driven approach does not adventure or save abnormal state visual semantics. Interestingly, the proposed SP-DIR calculation can better protect semantic implications for image retargeting. Other retargeting strategies [18], [19], [20] edit images to enhance visual nature of photos [21], [22]. Be that as it may, these plans don't expressly protect visual semantics, which may dispose of essential substance for visual quality and esthetics.

B. Semantic-Based Image Retargeting

[10] Consolidate confront recognition results to register significance maps for image retargeting. Goferman et al. [24] propose setting mindful saliency which recognizes the essential parts of the scene. Then again, Huang et al. [25] show a thumbnail age plot in light of model expressiveness and conspicuousness of frontal area protests in the wake of editing. Conversely, the proposed SP-DIR calculation considers semantics from items to scenes at multiple scales and also activity settings.

In [26], Jain et al. propose a conclusion to-end profound completely convolu-tional network for frontal area protest division and apply it to image retargeting. Hou et al. [27] propose a saliency strategy by acquainting short associations with the skip-layer structures inside the HED design. The Fast-AT conspire [28] is as of late created for producing thumbnail images in view of profound neural networks. Fan et al. [29] demonstrate that abnormal state semantic comprehension is fundamental for saliency assessment and propose the structure-measure to assess the forefront maps. In [30] Cho et al. propose a feebly and self-regulated profound CNN for content-mindful image retargeting. This network creates a retargeted image specifically from an info image and an objective viewpoint proportion by learning a semantic guide (consideration delineate. We take note of that these profound learning based techniques foresee paired maps though the SP-DIR calculation predicts delicate probabilities. Moreover, these techniques oper-ates on

the understood presumption that each image contains one striking item. Interestingly, we show that the proposed SP-DIR calculation can create target images containing multiple little questions in various scenes.

III. COMPOSING SEMANTIC COLLAGE

In this segment, we introduce the SP-DIR calculation which extricates semantic components and makes a semantic col-lage for image retargeting. Every collage is bolstered into a transporter to produce the objective image by evacuating or mutilating less essential pixels. The semantic collage can be joined with any transporter, i.e., AAD [3], multi-administrator [4] and significance separating (IF) [10].

A. Semantic Component

The semantic components including closer view, activity con-content and foundation are removed to depict an image for retargeting.

1) **Semantic Foreground Components**: The striking articles in an image are considered as the semantic frontal area components. For instance, the image in Figure 2 contains two principle closer view components, i.e., kid and ball. We utilize the best in class image parsing and classification modules to find closer view components.

Image parsing. We apply the pre-prepared completely convolution network [5] to parse each info image into 59 basic classes characterized in the Pascal-Context [31] dataset. The 59 classes, however still constrained, incorporate basic questions that every now and again happen as a rule images. We utilize every one of the 59 parsing confidence maps where each semantic component delineates meant by Mp. As appeared in Figure 3, the semantic component maps feature the articles, i.e., individual and building, admirably.

To begin with, for solidness we utilize 59 classes characterized in the Pascal-Context dataset [31] to show the viability of the proposed calculation. While constrained, they incorporate common questions that habitually happen when all is said in done images. Second, a few bigger semantic division datasets are discharged as of late. For instance, the ADE20K dataset contains 150 question and stuff classes with different explanations of scenes, objects, parts of articles, and sometimes even parts of parts. Third, it requires broad manual naming work to reach out to an extensive number of classes, i.e., 3000 classifications. One attainable approach is to fall back on the feebly directed semantic division techniques where bouncing box [32] or image level comments [33] are accessible.

Image classification. We utilize the VGG-16 network [34] pre-prepared on the ILSVRC-2012 dataset to foresee a name dispersion more than 1; 000 protest classifications in an image. As every classification is completed on the image level, a significance delineate acquired by means of a back proliferation go from the VGG network yield [35]. The semantic component delineate by the classification yield utilizing 1-channel image is meant by Mc. Figure 4 demonstrates the help of the fundamental articles, e.g., signal and flying creatures, can be visualized in spite of the fact that they involve little regions in the first image. The significance maps got from classification are reciprocal to those instigated from image parsing since more classifications (1; 000 versus 59) are considered.

2) **Action Context**: We consider the activity setting suradjusting the closer view objects for image retargeting. On the off chance that there is no activity in the scene, all pixels of the comparing semantic component outline near zero.

Activity acknowledgment. Figure 2 demonstrates an image where a kid kicks a ball. The activity setting in this scene is the kicking activity between two articles (i.e., kid and ball). We prepare a profound model to characterize 10 fine-grained activities in a route like the strategy by Oquab et al. [7]. The activity acknowledgment process is completed on the identified the jumping box encompassing a human by the Faster R-CNN technique [36], and the mistake back

spread is limited inside the bouncing box. Two agent cases of playing instruments are appeared in Figure 5. In the two cases, the activity settings with every single included protest are featured in the second section. The info images overlaid with the activity settings are appeared in the third segment. Given an image, the semantic component outline from the activity setting is meant as Ma.

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

Scene classification. Figure 2 demonstrates a scene containing a pitch. Scene classification gives holistic comprehension of an image. We utilize the profound model [8], which is prepared on the Places dataset with 2:5 million images to order 205 classes. The semantic component delineate from scene classification Ms is acquired comparatively as Mc. As appeared

Fig. 4. Semantic component maps constructed from scene classification. On each row, the input image, semantic component map from scene classification and overlaid image are shown

in Figure4, two images are anticipated as kitchen and island separately. The acquired Ms features the most representative subjects that can clarify the scene names. In the kitchen scene, the hearth is featured. In the island scene, the stone and encompassing water are found. These outcomes concur with the work by Zhou et al. [37] which demonstrates that question locators gained from the preparation procedure are in charge of scene classification. In this manner, the semantic component outline from scene classification feature districts of distinguished items. Interestingly, insignificant items, for example, lights on the roof, are overlooked.

B. Semantic Collage

The semantic components presented in Section III-A have a few confinements. To start with, in spite of the fact that the best in class profound modules are utilized, the semantic component maps may not be precise. For instance, the location module are probably going to produce false positives or negatives. Second, the setting data between various semantic components is missing. For instance, in Figure 2, the spatial connection amongst kid and ball is absent in the individual semantic component maps. To address these issues, we propose a classification guided fusion network to incorporate all component maps. While the significance maps have been utilized as a part of the current image retargeting strategies, we underline the semantic collage in this work viably protects semantics and coordinates multiple semantic component maps in view of various prompts.

1) **Classification-guided Fusion Network:** It has been re-gently demonstrated that question and also scene images have drastically extraordinary properties and ought to be freely treated for the classification [8] or esthetic assessment [9] tasks. Spurred by these perceptions, our network expressly groups an image as either protest situated or scene-arranged, and combined by independent weights.The contributions of the network are 62-channel semantic component maps. The activity, scene, classification maps all have 1 channel, while the division delineate 59 channels.

Likewise, the original images are additionally utilized as a part of the CRF-RNN [38] modules. The linked semantic component maps are bolstered into a 128 1 convlayer, trailed by two sub-networks. To begin with, the classification sub-network predicts the image as either scene-situated or protest arranged. It contains a worldwide normal pooling layer and a completely associated layer for expectation. Second, the relapse sub-network does convolutions in a path like the origin network [39]. The CRF-RNN modules are then added to smooth expectations. We combine both convolutional maps and CRF-RNN yield by pixel-wise normal pooling to relapse the objective inside the scope of [0,1]. The scene and protest arranged maps from the relapse sub-network are weighted by the classification results to create the semantic collage. Note that in the relapse sub-network, protest situated and scene-arranged image have isolate conv layers and CRF-RNN layers.

The semantic collage Mg is gotten by

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

Where $M = \{M_p, M_c, M_3, M_a\}$ is the concatenation of all semantic component maps to be fused and contains 62 channels. In the above equation, $r_o(.)$ and $r_s(.)$ 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 whetherc() 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 pa-rameters related to the classification sub-network are updated. The loss function L1 at this stage is a weighted multinomial logistic loss:

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

where !i 2 f0; 1g is ground truth classification label, !^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 function L2 is

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

Where $W_I \in \{0,1\}$ and 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 L2.

IV. S-RETARGET DATASET

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, and 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 is almost the same in the three sets.

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 of an image. In the second stage, the annotator's 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 super pixels and compactness factors. Each annotator then assigns a value to each image segment where a higher score corresponds to high relevance.

CONCLUSIONS

In this paper, we propose a profound image retargeting calculation that jellies the semantic significance of the first image. A semantic collage that speaks to the semantic significance conveyed by every pixel is produced in two stages. To begin with, multiple individual semantic components, i.e., including forefront, settings and foundation, are extricated by the cutting edge profound understanding modules. Second, all semantic component maps are joined by means of a classification guided fusion network to create the semantic collage. The network initially characterizes the image as question or scene-arranged one. Diverse classes of images have their separate fusion parameters. The semantic collage is encouraged into the bearer to create the objective image. Our future work incorporate investigating image inscription techniques [55] for better retargeting and related issues. What's more, we intend to incorporate the PixelCNN [56] and GAN [57], [58], [59] modules to the proposed calculation for retargeting tasks.

REFERENCES

[1] L. Wolf, M. Guttmann, and D. Cohen-Or, "Non-homogeneous content-driven video-retargeting," in ICCV, 2007.

[2] [2]Y.-S. Wang, C.-L. Tai, O. Sorkine, and T.-Y. Lee, "Optimized scale-and-stretch for image resizing," ACM TOG, 2008. 1, 2, 7

[3]D. Panozzo, O. Weber, and O. Sorkine, "Robust image retargeting via axis-aligned deformation," in EUROGRAPHICS, 2012. 1, 2, 3, 7, 8

[4]M. Rubinstein, A. Shamir, and S. Avidan, "Multi-operator media retar-geting," ACM TOG, 2009. 1, 2, 3, 7, 8

[5]J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," CoRR, vol. abs/1411.4038, 2014. 1, 3

[6]A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in NIPS, 2012. 1

[7]M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in CVPR, 2014. 1, 3

[8]B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in NIPS, 2014. 1, 3, 4

[9]S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in MM, 2010. 1, 4

[10]Y. Ding, J. Xiao, and J. Yu, "Importance filtering for image retargeting," in CVPR, 2011. 1, 2, 3, 7, 8

[11]J. Sun and H. Ling, "Scale and object aware image thumbnailing," IJCV, vol. 104, no. 2, pp. 135–153, 2013. 2, 7

[12]M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," in ACM TOG, 2008. 2, 7

[13]G.-X. Zhang, M.-M. Cheng, S.-M. Hu, and R. R. Martin, "A shape-preserving approach to image resizing," Computer Graphics Forum, 2009. 2

[14]J. Luo, "Subject content-based intelligent cropping of digital photos," in ICME, 2007. 2

[15]L. Marchesotti, C. Cifarelli, and G. Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in ICCV, 2009.2

[16]R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in CVPR, 2015. 2, 6, 7, 8

[17]J. Chen, G. Bai, S. Liang, and Z. Li, "Automatic image cropping: A computational complexity study," in CVPR, 2016. 2

[18]J. Yan, S. Lin, S. B. Kang, and X. Tang, "Change-based image cropping with exclusion and compositional features," IJCV, 2015. 2

[19 "Learning the change for automatic image cropping," in CVPR, 2013, pp. 971–978. 2

[20]Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: An experimental survey," CoRR, vol. abs/1610.00838, 2016. 2

[21]S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, Photo AestheticsRanking Network with Attributes and Content Adaptation. Springer International Publishing, 2016. 2

[22]X. Lu, Z. Lin, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in ICCV, 2015. 2

[23]M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," TPAMI, 2015. 2, 7, 8

[24]S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," TPAMI, 2012. 2

[25]J. Huang, H. Chen, B. Wang, and S. Lin, "Automatic thumbnail genera-tion based on visual representativeness and foreground recognizability," in ICCV, 2015. 2

[26]S. D. Jain, B. Xiong, and K. Grauman, "Pixel objectness," CoRR, vol. abs/1701.05349, 2017. 2

[27]Q. Hou, M. M. Cheng, X. W. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," CVPR, 2017.2, 7

[28]S. A. Esmaeili, B. Singh, and L. S. Davis, "Fast-at: Fast auto-matic thumbnail generation using deep neural networks," CoRR, vol. abs/1612.04811, 2016. 2

[29]D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A New Way to Evaluate Foreground Maps," in ICCV, 2017. 2

[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," 2017. 2

[31]R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, Urtasun et al., "The role of context for object detection and semantic segmentation in the wild," in CVPR, 2014. 3

[32]R. Hu, P. Dollr, K. He, T. Darrell, and R. Girshick, "Learning to segment every thing," arXiv:1711.10370, 2017. 3

[33]Q. Hou, P. K. Dokania, D. Massiceti, Y. Wei, M. M. Cheng, and Torr, "Bottom-up top-down cues for weakly-supervised semantic segmentation," EMMCVPR, 2017. 3

[34]K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," CoRR, vol. abs/1409.1556, 2014. 3

[35]K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," CoRR, vol. abs/1312.6034, 2013. 3

[36]S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," CoRR, vol.

abs/1506.01497, 2015. 3

`[37]B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," CoRR, vol. abs/1412.6856, 2014. 4

[38] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du,C. Huang, and P. Torr, "Conditional random fields as recurrent neuralnetworks," in ICCV, 2015. 4

[39]C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in CVPR, 2016. 4

[40]M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir, "A comparative study of image retargeting," in ACM TOG, 2010. 4, 6, 8

[41]A. Mansfield, P. Gehler, L. Van Gool, and C. Rother, "Visibility maps for improving seam carving," ECCV 2010 Workshops, 2012. 4

[42]M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisser-man, "The pascal visual object classes (voc) challenge," IJCV, 2010. 5

[43]J. Xiao, K. Ehinger, J. Hays, A. Torralba, and A. Oliva, "Sun database: Exploring a large collection of scene categories," IJCV, 2014. 5

[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 ICCV, vol. 2, 2001, pp. 416–423. 5

[45]R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," TPAMI, 2012. 5

[46]A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in ECCV, 2008. 5

[47]T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in CVPR, 2007. 6

[48]Q. Yan, L. Xu, J. Shi, andJ. Jia, "Hierarchical saliency detection," in CVPR, 2013. 6, 8

[49]Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, “MIT saliency benchmark,” http://saliency.mit.edu/. 6

[50]M.-M. Cheng, J. Warrell, L. Wen-Yan, S. Zheng, V. Vineet, and N. Crook, “Efficient salient region detection with soft image abstraction,” in ICCV, 2013. 7, 8

[51]E. Vig, M. Dorr, and D. Cox, “Large-scale optimization of hierarchical features for saliency prediction in natural images,” in CVPR, 2014. 7, 8

[52]C. Shen, M. Song, and Q. Zhao, “Learning high-level concepts by train-ing a deep network on eye fixations,” DLUFL Workshop, in conjunction with NIPS, 2012. 7, 8

[53]N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, “Predicting eye fixations using convolutional neural networks,” in CVPR, 2015. 7

[54]J. Pan, K. McGuinness, E. Sayrol, N. E. O’Connor, and X. Giro´ i Nieto, “Shallow and deep convolutional networks for saliency prediction,” CoRR, vol. abs/1603.00845, 2016. 7

[55]A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in CVPR, 2015. 9

[56]A. V. D. Oord, N. Kalchbrenner, and K. Kavukcuoglu, “Pixel recurrent neural networks,” ICML, 2016. 9

[57]A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” CoRR, vol. abs/1511.06434, 2015. 9

[58]I. J. Goodfellow, J. Pougetabadie, M. Mirza, B. Xu, D. Wardefarley, S. Ozair, A. Courville, Y. Bengio, Z. Ghahramani, and M. Welling, “Generative adversarial nets,” NIPS, 2014. 9

[59]M. Mirza and S. Osindero, “Conditional generative adversarial nets,” CoRR, vol. abs/1411.1784, 2014. 9