SURVEY ON SALIENT OBJECT DETECTION

[1]Radha N. Dhameliya, [2]Mahasweta J. Joshi, [3]Dr. Mehfuza S. Holia
[1]P.G. Student, [2]Assistant Professor, [3]Assistant Professor
[1]Dept. of Computer Engineering,
[1] Birla Vishwakarma Mahavidyalaya,V.V. Nagar, Gujarat, Anand, Gujarat

Abstract: Detecting and segmenting Salient and co-salient object in an image and video has attracted a lot of interest in computer vision. Now a day's many applications have emerged and models have been proposed. We aim to supply a comprehensive analysis of recent add salient object detection and provides this field, completely different areas like generic scene segmentation, object proposal generation, strikingness for fixation prediction and co-saliency detection via looking deep and wide. Covering (20) publications, our survey (1) root, key concepts and task (2) core techniques and modeling trends, and (3) datasets that are used in the salient object detection and segmentation. We also suggest future research work.

***Index Terms* - Salient object detection, bottom-up saliency, explicit saliency, visual attention, a region of interest.**

I. INTRODUCTION

Humans are able to detect visually distinctive, so-called salient, scenic regions has been investigated by multiple disciplines such as cognitive psychology, neuroscience, and computer Vision [7]. There are some topic related to visual saliency include salient object detection [1], fixation prediction [2], [3], Object importance [4] - [6], scene clutter, video interestingness [8] - [11], etc. We only focus on salient object detection, some research area that has been developed in the past.

1.1 What is salient object detection?

"Salient object detection" or "Salient object segmentation" is commonly interpreted in computer vision as a process that includes two stages: 1) detecting the most salient object and 2) segmenting the accurate region of that object. Rarely, however, models explicitly distinguish between these two stages.

In general, it's united that smart salience detection a model ought to meet a minimum of the subsequent three criteria: 1) good detection: the chance of missing real salient regions and incorrectly marking the background as a salient region ought to be low, 2) high resolution: saliency maps should have high or full resolution to accurately find salient objects and retain original image information, and 3) computational efficiency: as front-ends to other complex processes, these models should detect salient regions quickly.

1.2 Situating Salient Object Detection

Salient object detection models typically aim to detect only the foremost salient objects in very scene and segment the complete extent of these objects. Fixation prediction models, on the other hand, typically try to predict where humans look, i.e., a small set of fixation points [12]. Since the two types of methods output a single continuous-valued saliency map, where a higher value in this map indicates that the corresponding image pixel is more likely to be attended, they can be used interchangeably. A strong correlation exists between fixation locations and salient objects. Further, humans often agree on which each other when asked to choose the most salient object in a scene [13]. These are illustrated in Fig. 1.

Unlike salient object detection and fixation prediction models, object proposal models aim at producing a small set, typically a few hundreds or thousands, of overlapping candidate object bounding boxes or region proposals. Object proposal generation and salient object detection are highly related. Saliency estimation is explicitly used as a cue in objectness methods [14].

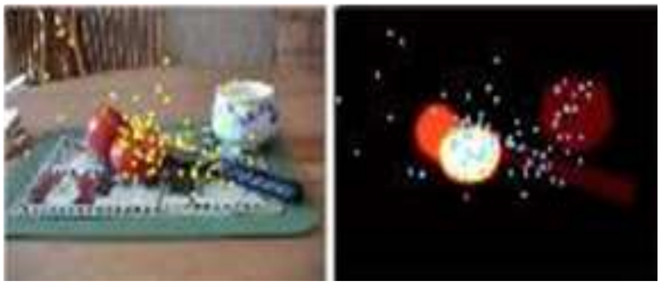


Fig. 1. An example image in Borji et al.'s experiment [13] along with annotated salient objects. Dots represent 3-second free-viewing fixations.

Image segmentation, semantic scene labeling or semantic segmentation, is one of the very well researched areas in computer vision. In distinction to salient object detection wherever the output could be a binary map, these models aim to assign a label, one out of many categories like a sky, road, and building, to each image pixel. Fig. 2 illustrates the difference among these themes.

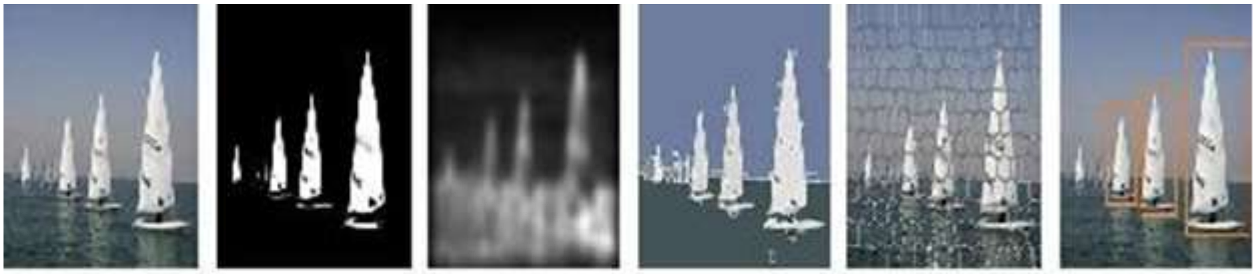


Fig. 2. Sample results produced by different models. From left to right: input image, salient object detection, fixation prediction, image segmentation (regions with various sizes), image segmentation (superpixels with comparable sizes), and object proposals (true positives).

II. SURVEY OF STATE OF THE ARTS

In this section review related works in 3 categories, including 1) salient object detection models, 2) applications, and 3) datasets.

2.1 Traditional Model

A large number of approaches have been proposed for detecting salient objects in images in the past two decades. Except for a few models which attempt to segment objects of interest (e.g. [15]), most of these approaches aim to identify the salient subsets1 from images first (i.e., compute a saliency map) and then integrate them to segment the entire salient object.

In general, classic approaches can be categorized in two different ways depending on the types operation or attributes they exploit.

1. **Block-based vs. region-based analysis:**

Two types of visual subsets have been utilized: blocks and regions, to detect salient objects. Blocks were primarily adopted by early approaches, while regions became popular with the introduction of super pixel algorithms.

2. **Intrinsic cues vs. extrinsic cues:**

A key step in detecting salient objects is to distinguish them from distracters. To extract various cues only from the input image itself to highlight targets and to suppress distracters (i.e., the intrinsic cues). However, intrinsic cues are often insufficient to distinguish targets and distracters especially when they share common visual attributes. To overcome this issue, they incorporate extrinsic cues such as user annotations, depth map, or statistical information of similar images to facilitate detecting salient objects in the image.

2.1.1. Block-based Models with Intrinsic Cues

In this section, we have a tendency to in the main review salient object detection models that utilize intrinsic cues extracted from blocks. Salient object detection is widely defined as capturing the uniqueness, distinctiveness, or rarity in a scene.

In early works, uniqueness was often computed as the pixel-wise center-surround contrast. The input image in a 2D space using the polar transformation of its features. Each region in the image is then mapped into a 1D linear subspace. Afterwards, the Generalized Principal Component Analysis (GPCA) [16] is used to estimate the linear subspaces without actually segmenting the image. Finally, salient regions are selected by measuring feature contrasts and geometric properties of regions. In Sec. 2.1.1 aim to detect salient objects based on pixels or patches where only intrinsic cues are utilized. These approaches usually suffer from two shortcomings: I) high-contrast edges usually stand out instead of the salient object, and II) the boundary of the salient object is not preserved well (especially when using large blocks). To overcome these problems, some strategies propose to compute salience supported regions. This offers two main advantages. First, the number of regions is far less than the number of blocks, which implies the potential to develop highly efficient and fast algorithms. Second, more informative features can be extracted from regions, leading to better performance. These region-based approaches will be discussed in the next subsection.

2.1.2. Region-based Models with Intrinsic Cues

Saliency models within the second subgroup adopt intrinsic cues extracted from image regions generated victimization ways like graph-based segmentation, mean-shift, and SLIC or Turbo pixels. Different from the block-based models, region-based models often segment an input image into regions aligned with intensity edges first and then compute a regional saliency map.

The regional saliency score is defined as the average saliency score of its contained pixels, defined in terms of multi-scale contrast. A set of rules to determine the background scores of each region based on observations from the background and salient regions. Saliency, defined as uniqueness in terms of **global regional contrast**. In [8], a region-based saliency algorithm is introduced by measuring the global contrast between the target regions with respect to all other image regions.

A **multi-scale local region contrast** based approach that calculates salience values across multiple segmentations for robustness purposes and combines these regional saliency values to get a pixel-wise salience map.

2.1.3. Models with Extrinsic Cues

Models in the third subgroup adopt the extrinsic cues to assist the detection of salient objects in images and videos. In addition to the visual cues observed from the single input image, the extrinsic cues can be derived from the ground truth annotations of the training images, similar images, the video sequences, a set of input images containing the common salient objects, depth maps, or light field images.

Salient object detection with similar images. With the provision of a progressively great deal of visual content on the web, salient object detection by leveraging the visually similar images to the input image has been studied in recent years. **Co-saliency object detection**. Instead of concentrating on computing saliency on a single image, co-salient object detection algorithms focus on discovering the common salient objects shared by multiple input images. That is, such objects can be the same object with different viewpoints or the objects of the same category sharing similar visual appearances. Note that the key characteristic of co-salient object detection algorithms is that their input is a set of images, while classical salient object detection models only need a single input image. Co-saliency detection is closely associated with the thought of image co-segmentation that aims to segment similar objects from multiple images. As stated in [17], three major differences exist between co-saliency and co-segmentation. First, co-saliency detection algorithms only focus on detecting the common salient objects while the similar but non-salient background might be also segmented out in co-segmentation approaches. Second, some co-segmentation methods, e.g., [18], need user input to guide the segmentation process in ambiguous situations. Third, salient object detection often serves as a pre-processing step, and thus more efficient algorithms are preferred than co-segmentation algorithms, especially over a large number of images.

2.1.4. Other Classic Model

In this section, we tend to review algorithms that aim to directly segment or localize salient objects with bounding boxes, and algorithms that are closely associated with saliency detection.

Localization models. Convert the binary segmentation map to bounding boxes. The final output is a set of rectangles around salient objects. Define saliency for a sliding window as its composition cost using the remaining image parts. Based on an over-segmentation of the image, the local maxima, which can efficiently be found among all sliding windows in a brute force manner, are assumed to correspond to salient objects.

Segmentation models. Segmenting salient objects are closely related to the figure-ground problem, which is essentially a binary classification problem trying to separate the salient object from the background. The complementary characteristics of imperfect saliency maps generated by different contrast-based saliency models. Specifically, two complementary saliency maps are first generated for each image, including a sketch-like map and an envelope-like map. The sketch-like map can accurately locate parts of the most salient object (i.e., a skeleton with high precision), while the envelope-like map can roughly cover the entire salient object (i.e., envelope with high recall). With these two maps, the reliable foreground and background regions can be detected from each image first to train a pixel classifier. By labeling all other pixels with this classifier, the salient object can be detected as a whole.

Supervised vs. unsupervised models. The majority of the prevailing learning-based works on saliency detection specialize in the supervised state of affairs, i.e., learning a salient object detector given a collection of coaching samples with ground-truth annotations. The aim here is to separate the salient elements of the background elements.

Salient object detection with depth. We live in real 3D environments where stereoscopic content provides additional depth cues for guiding visual attention and understanding the surroundings.

Salient object detection on a light field. The idea of using a light field for saliency detection. A light field, captured using a specifically designed camera e.g., Lytro, is essentially an array of images shot by a grid of cameras viewing the scene. The light field data offers two benefits for salient object detection: 1) it allows synthesizing a stack of images focusing at different depths, and 2) it provides an approximation of scene depth and occlusions.

2.1.5. Deep Learning Based Models

All the strategies that we have got reviewed to this point aim at detecting salient objects using heuristics. While hand-crafted features allow real-time detection performance, they suffer from several shortcomings that limit their ability in capturing salient objects in challenging scenarios.

Convolutional neural networks (CNN), as one of the most popular tools in machine learning, have been applied to many vision problems such as object recognition, semantic segmentation, and edge detection. Recently, it has been shown that CNN's, are also very effective when applied to salient object detection. Thanks to their multi-level and multi-scale features, CNN's are capable of accurately capturing the most salient regions without using any prior knowledge (e.g., segment-level information).

Basically, salient object detection models based on deep learning can be split into two main categories. The first category includes models that have used multi-layer perceptron (MLPs) for saliency detection. The second category includes models that are based on "Fully Convolutional Networks" (FCN-based).

1. CNN-based Models

One-dimensional (1D) convolution-based methods. As an early attempt, region-based approach to learning super pixel-wise feature representations. Their approach dramatically reduces the computational cost compared to pixel-wise CNN, meanwhile takes global context into consideration. However, representing a super pixel with the mean color is not informative enough. Further, the spatial structure of the image is difficult to be fully recovered using 1D convolution and pooling operations, leading to cluttered predictions, especially when the input image is a complex scene.

Leveraging local and global context. Wang et al. consider both local and global information for better detection of salient regions [19]. To this end, two sub networks are designed for local estimation and global search, respectively. A deep neural network (DNN-L) is first used to learn local patch features to determine the saliency value of each pixel, followed by a refinement operation which captures the high-level objectness. For global search, they train another deep neural network (DNN-G) to predict the saliency value of each salient region using a variety of global contrast features such as geometric information, global contrast features, etc.

2. FCN-based Models

Unlike CCN-based models that operate at the patch level, fully convolutional networks (FCNs) consider pixel level operations to overcome the problems caused by fully connected layers such as blurriness and inaccurate predictions near the boundaries of salient objects. Due to fascinating properties of FCNs, an excellent range FCN-based salient object detection models are introduced recently. Mainly 3 following advantages have been obtained in utilizing FCN-based models for saliency detection.

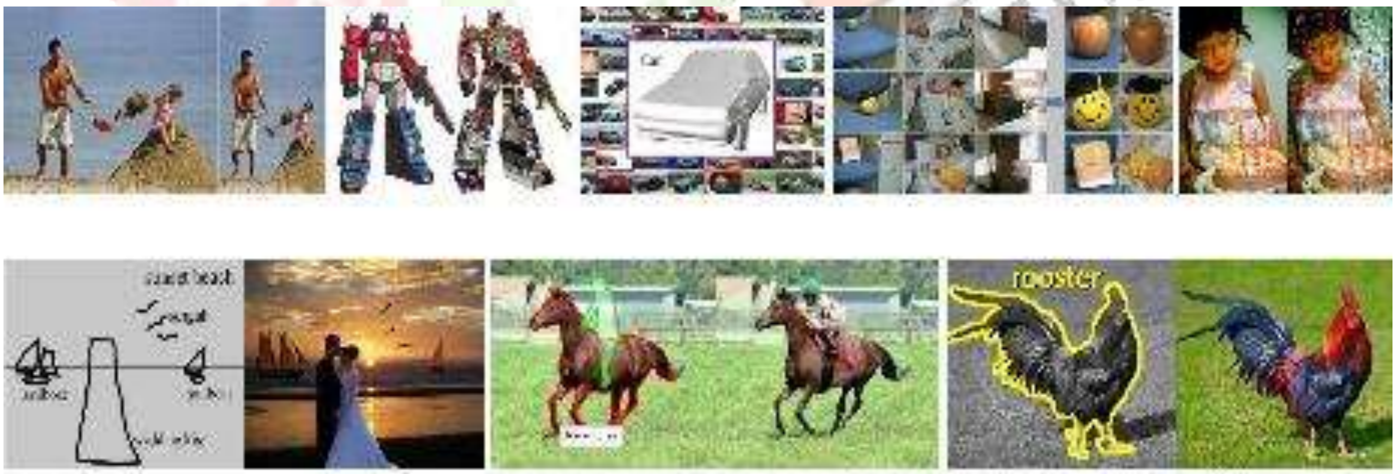
a. **Local vs. global**. FCN-based methods are capable of learning both local and global information internally. Lower layers tend to encode more detailed information such as edge and fine components, while deeper layers favor global and semantically meaningful information.

b. **Pre-training and fine-tuning.** The effectiveness of fine-tuning a pre-trained network has been demonstrated in many different applications. The network is typically pre-trained on the ImageNet dataset [20] for image classification. The learned features, more importantly, are able to capture high-level semantic knowledge on object categories, as the employed networks are pre-trained for a scene and object classification tasks.

c. **Versatile architectures**. A CNN architecture is formed by a stack of distinct layers that transform the input images into an output map through a differentiable function. The diversity of FCNs allows designers to design different structures that are appropriate for them.

2.2. Applications of salient object detection

The value of salient object detection models lies in their applications in many areas of computer vision, graphics, and robotics. Salient object detection models have been utilized for several applications such as object detection and recognition , image and video compression, video summarization, photo collage/media re-targeting/cropping/thumb-nailing, image quality assessment, image segmentation, content-based image retrieval and image collection browsing, image editing and manipulating, visual tracking, object discovery, and human-robot interaction. Fig. 3 shows example applications.


(a) Content aware resizing (b) Image collage (c) View selection (d) Unsupervised learning (e) Mosaic

(f) Image montage (g) Object manipulation (h) Semantic colorization

Fig. 3. Sample applications of salient object detection. Images are reproduced from corresponding references.

2.3. Salient object detection datasets

As more models have been proposed in the literature, more datasets have been introduced to further challenge saliency detection models. Early attempts aim to collect images with salient objects being annotated with bounding boxes (e.g., **MSRA-A** and **MSRA-B**), while later efforts annotate such salient objects with pixel-wise binary masks (e.g., **ASD** and **DUT- OMRON**).

Typically, images, which can be annotated with accurate masks, contain only limited objects (usually one) and simple background regions. On the contrary, recent attempts have been made to collect datasets with multiple objects in complex and cluttered backgrounds (e.g., [13]). A list of 22 salient object datasets including 20 image datasets and 2 video datasets is shown in Table 1. Notice that all images or video frames in these datasets are annotated with binary masks or rectangles.

Table 1. Overview of popular salient object datasets

Dataset	Year	Images	Obj	Ann	Resolution	Sbj	Eye	I/V
MSRA-A	2007	20K	~1	BB	400x300	3	-	I
MSRA-B	2007	5K	~1	BB	400x300	9	-	I
SED1	2007	100	1	PW	~300x225	3	-	I
SED2	2007	100	2	PW	~300x225	3	-	I
ASD	2009	1000	~1	PW	400x300	1	-	I
SOD	2010	300	~3	PW	481X321	7	-	I
ICoSeg	2010	643	~1	PW	~500X400	1	-	I
MSRA5K	2011	5K	~1	PW	400x300	1	-	I
Infrared	2011	900	~5	PW	1024X768	2	15	I
ImgSal	2013	235	~2	PW	640X480	19	50	I
CSSD	2013	200	~1	PW	~400X300	1	-	I
ECSSD	2013	1000	~1	PW	~400X300	1	-	I
MSRA10K	2013	10K	~1	PW	400X300	1	-	I
THUR15K	2013	15K	~1	PW	400X300	1	-	I
DUT-OMRON	2013	5172	~5	BB	400X400	5	5	I
Bruce-A	2013	120	~4	PW	681X511	70	20	I
Judd-A	2014	900	~5	PW	1024X768	2	15	I
PASCAL-S	2014	850	~5	PW	Variable	12	8	I
UCSB	2014	700	~5	PW	405X405	100	8	I
OSIE	2014	700	~5	PW	800X600	1	15	I
RSD	2009	62,356	Var	BB	Variable	23	-	V
STC	2011	4870	~1	BB	Variable	1	-	V

Table 1. Overview of popular salient object datasets. Top: image datasets, Bottom: video datasets. Obj = objects per Image; Ann = Annotation; Sbj = Subjects/Annotators; Eye = Eye tracking subjects; I/V = Image/Video.

III. FUTURE DIRECTION

Most benchmarks and saliency models discussed in this study deal with single images. Unfortunately, salient object detection on multiple input images, e.g., salient object detection on video sequences, co-salient object detection, and salient object detection over depth and light field images, are less explored. One reason behind this is the limited availability of benchmark datasets on these problems. For example, as mentioned in Sec. 4, there are only two publicly available benchmark datasets for video saliency (mostly cartoons and news). For these videos, only bounding boxes are provided for the key frames to roughly localize salient objects.

Existing saliency models do not split salient regions into objects. However, humans possess the potential of detecting salient objects at associate in nursing an instance level. Instance-level salience is helpful in many applications, like image editing and video compression.

IV. CONCLUSION

In this paper, we exhaustively review salient object detection literature with respect to its closely related areas. Detecting and segmenting salient objects is very useful. Objects in images automatically capture more attention than background stuff, such as grass, trees, and sky. Therefore, if we can detect salient or important objects first, then we can perform detailed reasoning and scene understanding at the next stage. Compared to traditional special-purpose object detectors, salience models are unit general, generally quick, and don’t would like significant annotation. These properties allow processing a large number of images at low cost. Although salient object detection and segmentation ways have created nice strides in recent years, an awfully sturdy salient object detection algorithmic rule that's able to generate high-quality results for nearly all images remains missing. Even for humans, what is the most salient object in the image, is sometimes a quite ambiguous question.

REFERENCES

[1] M. Cheng, N. J. Mitra, X. Huang, P. H.Torr, and S. Hu, "Global contrast based salient region detection," IEEE TPAMI, vol. 37, no. 3, pp. 569–582, 2015.

[2] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F.Durand, A. Oliva, and A. Torralba, "Mit saliency benchmark (2015)," 2015.

[3] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand, "Where should saliency models look next?" in European Conference on Computer Vision, 2016, pp. 809–824.

[4] M. Spain and P. Perona, "Measuring and predicting object importance," IJCV, vol. 91, no. 1, pp. 59–76, 2011.

[5] A. C. Berg, T. L. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos et al., "Understanding and predicting importance in images," in CVPR, 2012, pp. 3562–3569.

[6] B. M't Hart, H. C. Schmidt, C. Roth, and W. Einh¨auser, "Fixations on objects in natural scenes: dissociating importance from salience," Frontiers in psychology, vol. 4, 2013.

[7] Borji and L. Itti, "State-of-the-art in visual attention modeling," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 185–207, Jan. 2013.

[8] H. Katti, K. Y. Bin, T. S. Chua, and M. Kankanhalli, "Preattentive discrimination of interestingness in images," in IEEE ICME, 2008, pp. 1433– 1436.

[9] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Van Gool, "The interestingness of images," ICCV, 2013.

[10] S. Dhar, V. Ordonez, and T. L. Berg, "High-level describable attributes for predicting aesthetics and interestingness," in CVPR, 2011, pp. 1657–1664.

[11] Y.-G. Jiang, Y. Wang, R. Feng, X. Xue, Y. Zheng, and H. Yang, "Understanding and predicting interestingness of videos," AAAI, 2013.

[12] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," IEEE TPAMI, vol. 35, no. 1, pp. 185–207, 2013.

[13] A. Borji, D. N. Sihite, and L. Itti, "What stands out in a scene? a study of human explicit saliency judgment," Vision Research, vol. 91, pp. 62–77, 2013.

[14] B. Alex, T. Deselaers, and V. Ferrari, "What is an object?" in CVPR, 2010, pp. 73–80.

[15] G. Hua, Z. Liu, Z. Zhang, and Y. Wu, "Iterative local-global energy minimization for automatic extraction of objects of interest," IEEE TPAMI, vol. 28, no. 10, pp. 1701–1706, 2006.

[16] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (PCA)," IEEE transactions on pattern analysis and machine intelligence, vol. 27, no. 12, pp. 1945–1959, 2005.

[17] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," IEEE TIP, vol. 22, no. 10, pp. 3766–3778, 2013.

[18] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "iCoseg: Interactive co-segmentation with intelligent scribble guidance," in CVPR. IEEE, 2010, pp. 3169–3176.

[19] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3183– 3192.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein et al., "Imagenet large-scale visual recognition challenge," IJCV, vol. 115, no. 3, pp. 211–252, 2015.