Beyond Universal Saliency: Personalized Saliency Prediction with Multi-task CNN †

Yanyu Xu[1] , **Nianyi Li**[2,3], **Junru Wu**[1], **Jingyi Yu**[1,3], **and Shenghua Gao**[1*]
[1]ShanghaiTech University, Shanghai, China.
[2]University of Delaware, Newark, DE, USA.
[3] Plex-VR digital technology Co., Ltd.
{xuyy2, wujr1, yujy1, gaoshh}@shanghaitech.edu.cn, nianyi@udel.edu

Abstract

Saliency detection is a long standing problem in computer vision. Tremendous efforts have been focused on exploring a universal saliency model across users despite their differences in gender, race, age, *etc*. Yet recent psychology studies suggest that saliency is highly specific than universal: individuals exhibit heterogeneous gaze patterns when viewing an identical scene containing multiple salient objects.

In this paper, we first show that such heterogeneity is common and critical for reliable saliency prediction. Our study also produces the first database of personalized saliency maps (PSMs). We model PSM based on universal saliency map (USM) shared by different participants and adopt a multi-task CNN framework to estimate the discrepancy between PSM and USM. Comprehensive experiments demonstrate that our new PSM model and prediction scheme are effective and reliable.

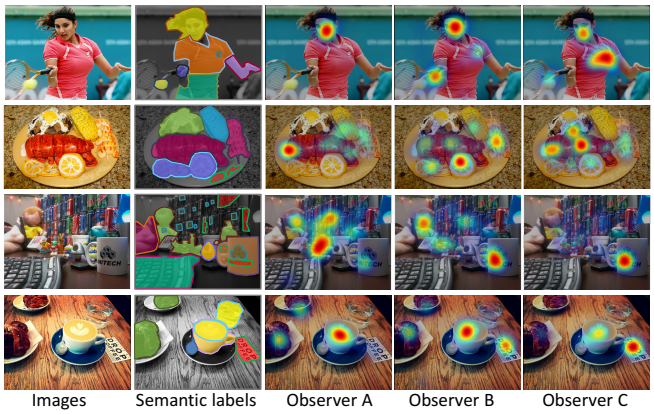

Figure 1: An illustration of PSM dataset. Our dataset provides both eye fixations of different subjects and semantic labels. Due to the large amount of objects in our dataset, for each image, we didn't fully segment it and only labelled objects that cover at least three gaze points from each individual. A notable difference between PSM and its predecessors is that each subjects looks 4 times on PSM data to derive solid fixation ground truth maps. Both commonality and distinctiveness exist for PSMs viewed by different participant. This motivates us to model PSM based on USM.

1 Introduction

Saliency refers to a component (object, pixel, person) in a scene that stands out relative to its neighbors and has been considered key to human perception and cognition. Traditional saliency detection techniques attempt to extract the most pertinent subset of the captured sensory data (RGB images or light fields) for predicting human visual attention. Applications are numerous, ranging from compression [Itti, 2004] to image re-targeting [Setlur *et al.*, 2005], and most recently to virtual reality and augmented reality [Chang *et al.*, 2016].

By far, nearly all previous approaches have focused on exploring a universal saliency model, i.e., to predict potential salient regions common to users while ignoring their differences in gender, race, age, personality, etc. Such universal solutions are beneficial in the sense they are able to capture all "potential" saliency regions. Yet they are insufficient in recognizing heterogeneity across individuals. Examples in Fig. 1 illustrate that while multiple objects are deemed highly salient within the same image (eg, *human face* (first row), *text* (last tow rows) and object of (*high color contrast*), different individuals have very different fixation preferences when viewing the image. For the rest of the paper, we use term *universal saliency* to describe salient regions that incur high fixations across all subjects and term *personalized saliency* to describe the heterogeneous ones.

Motivation. In fact, heterogeneity in saliency preference has been widely recognized in psychology: "Interestingness is highly subjective and there are individuals who did not consider any image interesting in some sequences" [Gygli *et al.*, 2013]. Therefore, once we know a person's personalized interestingness over each image (personalized saliency), we shall design tailored algorithms to cater to him/her needs. For example, in the application of image retargeting, the texts on the table in the fourth row in Fig. 1 should be pre-

*indicates corresponding author

†This work was supported by the Shanghai Pujiang Talent Program(No.15PJ1405700), and NSFC (No. 61502304).

served for observer B and C when resizing the image whereas such texts are less important for observer A. For applications in VR/AR, one can design data compression algorithms that personalized salient regions should be less compressed in order to both improve the users' experience and reduce the size of data in transmission. In addition, we can embed characters/logo/advertisement at those personalized salient regions for different individuals. Despite its importance, very little work has been carried out on studying such heterogeneity, partially due to the lack of suitable datasets and experiments. Further, the problem is inherently challenging as saliency variations across individuals are determined by multiple factors, e.g., gender, race, education, *etc.* , as well as the content of the image such as the color, location, size and type of objects.

In this paper, we present the first dataset of personalized saliency maps (PSMs) that consists of 1600 images viewed by 20 human subjects. To improve reliability, we ensure that each image is viewed by every subject for 4 times over about one week interval. We use the '*Eyegaze Edge*' eye tracker to track gaze and produce a total of 32,000 ($1,600 \times 20$) fixation maps. To correlate the acquired PSMs and the image contents, we manually segment each image into a collection of objects and semantically label them. Examples in Fig. 1 illustrate how fixations vary across three human subjects. Our annotated dataset provides fine-grained semantic analysis for studying saliency variations across individuals. For example, we observed that certain types of objects such as watches, belts would introduce more incongruity (possibly due to gender differences) whereas other types such as faces would lead to more coherent fixation maps, as shown in Table 2.

We further present a computational model towards this personalized saliency detection problem. Notice that saliency maps from different individual still share certain commonality via the USM. Hence, we model the PSM as a combination of USM and a residual map which is related to the identity and the image contents. We adopt a multi-task convolutional neural network (CNN) to identify the discrepancy between PSM and USM for each person, as shown in Fig. 4.

The contributions of our paper are two-fold: i) To our knowledge, it is the first work that specifically tackles the personalized saliency and we build the first dataset for personalized saliency detection; ii) We present a USM based PSM detection scheme and a multi-task CNN solution to estimate the discrepancy between PSM and USM. Experimental results demonstrate the effectiveness of our framework.

2 Related Work

Tremendous efforts on saliency detection have been focused on predicting universal saliency. For the scope of our work, we only discuss the most relevant ones. We refer the readers to [Borji *et al.*, 2014] for a comprehensive study on existing universal saliency detection schemes.

Universal Saliency Detection Benchmarks. There are a few widely used saliency object detection and fixation prediction datasets, in which each image is generally associated with a single ground truth saliency map, averaged across the fixation maps across the participates. To select images suitable for personalized saliency, we explore several popular eye fixation datasets. The MIT dataset [Judd *et al.*, 2009] contains 1,003 images viewed by 15 subjects. In addition, the PASCAL-S [Li *et al.*, 2014] dataset provide the ground truth for both eye fixation and object detection and consist of 850 images viewed by 8 subjects. The iSUN dataset [Xu *et al.*, 2015], a large scale dataset used for eye fixation prediction, contains 20,608 images from the SUN database. The images are completely annotated and are viewed by users. Finally, the SALICON dataset [Huang *et al.*, 2015] consists of 10,000 images with rich contextual information.

CNN Based Saliency Detection. It has been increasingly popular to use deep networks for saliency detection. Huang *et al.* [Huang *et al.*, 2015] propose to fine-tune CNNs pretrained for object recognition via a new objective function based on saliency evaluation metrics such as Normalized Scanpath Saliency (NSS), Similarity, or KL-Divergence,*etc.* Pan *et al.* [Pan *et al.*, 2016] propose to use a shallow convnet trained from scratch and fine-tune a deep convnet that trained for image classification on the ILSVRC-12 dataset. Liu *et al.* [Liu *et al.*, 2015] propose a multi-resolution CNNs that are trained from image regions centered on fixation and non-fixation locations at multi-scales. Srinivas *et al.* present a DeepFix [Kruthiventi *et al.*, 2015] network by using Location Biased Convolution filters to allow the network to exploit location dependent patterns. Kruthiventi *et al.* [Kruthiventi *et al.*, 2016] propose a unified framework to predict eye fixation and segment salient objects. All these approaches have focused on the universal saliency model and we show many merits of these techniques can also benefit personalized saliency.

3 PSM Dataset

We start with constructing a dataset suitable for personalized saliency analysis.

3.1 Data Collection

Clearly, the rule of thumb for preparing such a dataset is to choose images that yield distinctive fixation map among different persons. To do so, we first analyze existing datasets. A majority of existing eye fixation datasets provide the one-time gaze tracking results of each individual human subject. Specifically, we can correlate the level of agreement across different observers with respect to the number of object categories in the image. When an image contains few objects, we observe that a subject tends to fix his/her gaze at locations where objects that have specific semantic meanings, e.g., faces, text, signs [Judd *et al.*, 2009; Xu *et al.*, 2014]. These objects indeed attract more attention and hence are deemed more salient. However, when an image consists of multiple objects all with strong saliency as shown in Fig. 1, we observe a subject tends to diverge his/her attention. In fact, the subject focuses attention on objects that attract his/her most personally. We therefore deliberately choose 1,600 images with multiple semantic annotations to construct our dataset for PSM purpose. Among them, 1,100

images are chosen from existing saliency detection datasets including SALICON [Jiang *et al.*, 2015], ImageNet [Russakovsky *et al.*, 2015], iSUN [Xu *et al.*, 2015], OSIE[Xu *et al.*, 2014], PASCAL-S [Li *et al.*, 2014], 125 images are captured by ourselves, and 375 images are gathered from the Internet.

3.2 Ground Truth Annotation

To gather the ground truth, we have recruited 20 student participants (10 males, 10 females, aged between 20 and 24). All participants have normal or corrected-to-normal vision. In our setup, each observer sits about 40 inches in front of a 24-inches LCD monitor of a 1920×1080 resolution. All images are resized to the same resolution. We conduct all experiments in an empty and semi-dark room, with only one standby assistant. An eye tracker ('*Eyegaze Edge*' eye tracker) records their gazes as they view each image for 3 seconds. We partition 1,600 images into 34 sessions each containing 40 to 55 images. Each session lasts about 3 minutes followed by a half minute break. The eye tracker is re-calibrated at the beginning of each session. To ensure the veracity of the fixation map of each individual as well as to remove outliers, we have each image be viewed by each observer 4 times. We then combine the 4 saliency maps of the same image viewed by the same person, and use the result as the ground truth PSM of the observer. To obtain a continuous saliency map of an image from the raw data of eye tracker, we follow [Judd *et al.*, 2009] by smoothing the fixation locations via Gaussian blurs.

To further analyze the causes of saliency heterogeneity, we conduct the semantic segmentation for all 1,600 images via the open annotation tool LabelMe [Russell *et al.*, 2008]. Specifically, we annotate 26,100 objects of 242 classes in total and identify objects that attract more attention for each individual participant. To achieve this, we compare the fixation map with the mask of a specific object and use the result as the attention value of the corresponding object. We then average the result over all images that containing the same object, and use it to measure the interestingness of the object to a specific participant. In Fig. 2, we illustrate some representative objects and persons and show the distribution of the interestingness of various objects for a same participant. We observe that all participants exhibit a similar level of interestingness measure on faces where they exhibit different interestingness measures on various objects such as watch, bow tie, *et al.* This validates that it is necessary to choose images with multiple objects to build our PSM data.

3.3 Dataset Analysis

Why is each image viewed multiple times for ground-truth annotation? To validate whether it is necessity for a subject to view each image multiple times, we randomly sample 220 images, and each image is viewed by the same participant 10 times. The time interval for the same person to view the same image ranges from one day to one week because we want to get the short term memory of the person for the given image. We then calculate the differences of these saliency maps in terms of the commonly used metrics for saliency detection [Judd *et al.*, 2012]: CC, Similarity. We average these criteria for all persons and all images, and we show the results in Fig. 3. We observe that the saliency map obtained by viewing each image only once vs. multiple times exhibit significant differences. Further, the saliency map averaged over 4 or more times is closer to the long term result.

	Person 1	Person 4	Person 6	Person 7	Person 8
men bow tie	0.068388	0.046459	0.035015	0.07911	0.025138
women bow tie	0.014818	0.019792	0.078912	0.109666	0.004215
men hand watch	0.034834	0.034573	0.057979	0.036348	0.027059
women hand watch	0.035535	0.04356	0.041277	0.033336	0.022686
men face	0.025989	0.044911	0.04291	0.03387	0.03736
women face	0.027088	0.040768	0.043192	0.037849	0.035902

Figure 2: The distribution of the interestingness of various objects for a same participant. The value is calculated as follows: we sum values of the fixation map intersecting with the mask of a specific object, and divide it with the total of fixation maps over the whole image. Thus higher value indicates that the participant puts more attention on the object.

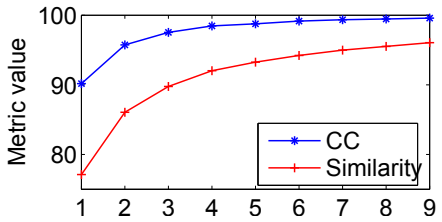

Figure 3: The point with $x = n$ measures the differences between ground truth saliency maps generated by viewing the same image n times and n+1 times. This figure shows that when $n \geq 4$, the ground truth saliency map generated by viewing the image n times has little difference with that generated by observing the image $n+1$ times. Thus viewing each image 4 times is enough to get a robust estimation of the PSM ground truth.

Heterogeneity among different datasets. To further illustrate that our proposed dataset is appropriate for personalized saliency detection task, we compare the inter-subject consistency, i.e., the agreement among different viewers, in our PSM dataset and other related datasets. Specifically, for each dataset, we first enumerate all possible subject-pairs, i.e., two different subjects, and then compute the average AUC scores across all pairs. Recall that our PSM dataset consists of images from different datasets, eg, MIT, OSIE, ImageNet, PASCAL-S, SALICON, iSUN *etc.* , and only MIT, OSIE, PASCAL-S are designed for saliency tasks*. Hence, we only compare the consistency scores among ours and above three datasets, and we show the results in Table 1. We observe that our dataset achieves the lowest inter-subject consistency values among all relative ones, indicating that the heterogeneity in our saliency maps are more severe than the others.

*Even though SALICON and iSUN are also saliency fixation datasets, the ground truth were annotated based on mouse-tracking and web camera respectively.

AUC judd scores			
Ours	MIT	OSIE	PASCAL-S
79.11	89.34	88.47	88.10

Table 1: Inter-subject consistency of different datasets. To compute the inter-subject consistency, we compute AUC judd for pair-wise saliency maps viewed by different observers for each image, then we average the results over all images. For fair comparison, the AUC judd of our method reported here is based on the saliency maps viewed by each observer once.

4 Approach

4.1 Problem Formulation

[Cornia *et al.*, 2016][Pan *et al.*, 2016] employed CNN in an end-to-end strategy to predict saliency map and now serves as the state-of-the-art. Intuitively, we can follow the same strategy for PSM prediction, *i.e.* training a separate CNN for each participant to map the RGB images to PSM. However, such strategy is neither scalable nor feasible for a number of reasons. Firstly, it needs a vast amount of training samples to learn a robust CNN for each participant. This requires subjects to view thousands of images with high concentration, which is hard and extremely time consuming. Secondly, training multiple CNNs for different subjects is computationally expensive and inefficient.

While each participant is unique in terms of their gender, race, age, personality, etc, resulting in their incongruity in saliency preference, different participants still share commonalities in their observed saliency maps because certain objects, such as faces and logos, always seem to attract the attention of all participants as shown in Fig. 1.

For this reason, instead of predicting the PSM directly, we set out to explore the difference map between USM and PSM. The discrepancy map $\Delta(P_n, I_i)$ for the given image I_i ($i = 1, \ldots, K$) of the n-th participant P_n ($n = 1, \ldots, N$) is of the form:

$$S_{PSM}(P_n, I_i) = S_{USM}(I_i) + \Delta(P_n, I_i) \tag{1}$$

where, $S_{PSM}(P_n, I_i)$ is the desired personalized saliency map and $S_{USM}(I_i)$ is the universal saliency map.

Note that the USMs by traditional saliency method entail the commonality in a saliency map observed by different participants. We convert the problem of predicting PSMs to estimating the discrepancy $\Delta(P_n, I_i)$ and we show it is much more efficient than directly estimating PSMs from RGB images as shown in . This is because that the universal saliency map $S_{USM}(I_i)$ itself already provides a rough estimation of the PSM, and predicting the discrepancy $\Delta(P_n, I_i)$ is actually easier than directly estimating the PSM from an RGB image. In addition, if we take the discrepancy $\Delta(P_n, I_i)$ as an error correction function, the PSM prediction problem can be therefore viewed as a regression task to correct the inaccurate input (USM), which can be implemented in high performance CNN scheme as shown in [Carreira *et al.*, 2015]. Given I_i and $S_{USM}(I_i)$, we propose a Multi-task CNN network to estimate $\Delta(P_n, I_i)$.

4.2 Multi-task CNN

Since $\Delta(P_n, I_i)$ is subject-dependent and at the same time dependant to the content of the input image, we construct a Multi-task CNN network to tackle it. The inputs of network are images with their corresponding universal saliency map and our goal is to estimate the discrepancy maps $\Delta(P_n, I_i)$ for n-th participants through n-th task. The network architecture of our Multi-task CNN is illustrated in Fig. 4.

Suppose we have N participants in total. We concatenate a 160×120 resolution RGB image with its USM from general saliency models and generate a $160 \times 120 \times 4$ cube as the input of the multi-task network. For image I_i, $\Delta(P_n, I_i)$ is the output of the n-th task corresponding to the discrepancy between PSM and USM for the n-th person. There are four convolutional layers shared by all participants after which the network is then split into N tasks which is exclusive for N participants. Each task has three convolutional layers followed by an ReLU activation function.

[Cornia *et al.*, 2016] and [Lee *et al.*, 2014] show that by adding the supervision in the middle layers, the features learned by CNN will be more discriminative and can boost the performance of an given task. Consequently, we set an additional Loss Layer on *conv5* and *conv6* layer of the n-th task to impose the middle layer supervision , which can help the prediction of $\Delta(P_n, I_i)$. For the n-th task, $f_\ell^n(S_{USM}(I_i), I_i) \in \mathbb{R}^{h_\ell \times w_\ell \times d_\ell} (\ell = 5, 6, 7)$ is the feature map after the ℓ-th convolutional layer (the first convolutional layer corresponds to the first exclusive convolutional layer, so ℓ starts from 5). For each feature map $f_\ell^n(S_{USM}(I_i), I_i)$, a 1×1 convolutional layer was employed to map it to $S_\ell(S_{USM}(I_i), I_i) \in \mathbb{R}^{h_\ell \times w_\ell \times 1}$, which is the target discrepancy. To make $S_\ell(S_{USM}(I_i), I_i)$ close to $\Delta_\ell(P_n, I_i)$, we set the objective function as:

$$\min \quad \sum_{\ell=5}^{7} \sum_{n=1}^{N} \sum_{i=1}^{K} \|S_k(S_{USM}(I_i), I_i) - \Delta_\ell(P_n, I_i)\|_F^2 \tag{2}$$

Then we use mini-batch based stochastic gradient descent to optimize all parameters in our Multi-task CNN.

Remarks: Compared with techniques that use separate CNNs to predict $\Delta(P_n, I_i)$ for different participants, our Multi-task CNN architecture has the two key advantages:

1. Previous approaches [Li *et al.*, 2016] [Zhang *et al.*, 2014] have shown that features extracted by the first several layers can be shared between multiple tasks. In a similar vein, we treat PSMs as some distinct but related regression tasks across different individuals. Different from the multi-task CNN for USM prediction [Li *et al.*, 2016], our network shares lots of parameters which reduces the number of parameters and the memory consumption. Therefore, we are able to train these shared parameters using all training samples from all participants.

2. Note that in our architecture, the first few layers are shared and trained by all participants. In the deployment stage, given any unrecorded observer, our model only requires training the last three layers. Thus such a multi-task framework makes the problem scalable for open set settings.

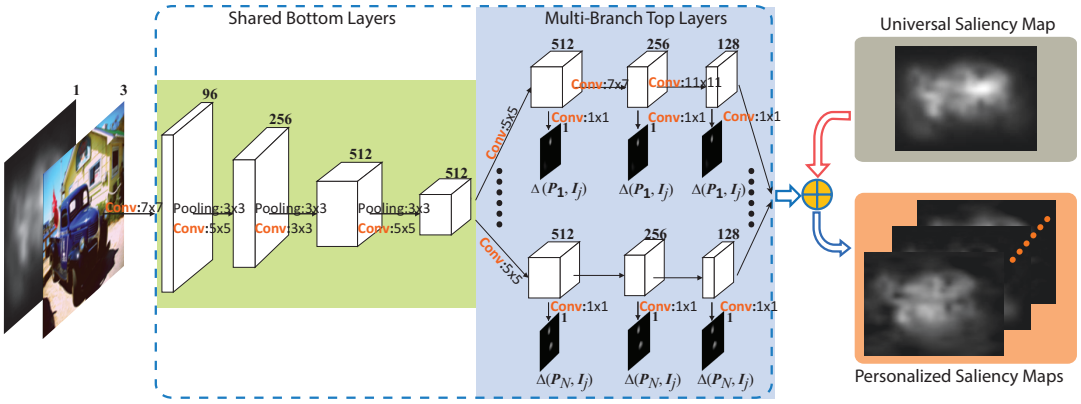

Figure 4: The pipeline of our Multi-task CNN based PSM prediction.

Methods	CC	Similarity	AUC judd
RGB based MultiConvNets	62.24	65.27	77.83
RGB based Multi-task CNN	64.68	66.28	79.98
LDS [Fang *et al.*, 2016]	65.73	63.34	82.96
LDS + MultiConvNets	70.71	75.65	83.69
LDS + Multi-task CNN	**72.19**	**76.07**	**84.97**
ML-Net [Cornia *et al.*, 2016]	41.35	51.30	71.80
ML-Net + MultiConvNets	65.35	79.42	81.70
ML-Net + Multi-task CNN	**67.53**	**80.17**	**83.45**
BMS [Zhang and Sclaroff, 2013]	59.59	71.36	80.26
BMS + MultiConvNets	68.68	79.66	83.79
BMS + Multi-task CNN	**70.33**	**80.41**	**85.03**
SalNet [Pan *et al.*, 2016]	72.66	74.18	84.67
SalNet + MultiConvNets	74.85	77.89	85.09
SalNet + Multi-task CNN	**76.28**	**79.08**	**85.94**

Table 2: The performance comparison of difference methods on our PSM dataset.

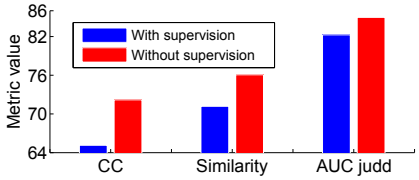

Figure 5: The effect of supervision on middle layers in our Multi-task CNN.

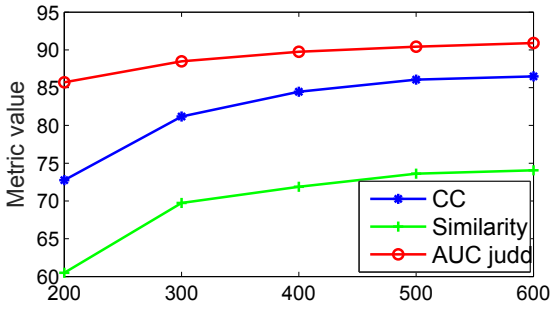

Figure 6: The effect of the number of training samples on the accuracy of PSM prediction.

5 Experiments

5.1 Experimental Setup

Parameters. We implement our solution on the CAFFE framework [Jia *et al.*, 2014]. We train our network with the following hyper-parameters setting: mini-batch size (40), learning rate (0.0003), momentum (0.9), weight decay (0.0005), and number of iterations (40,000). In our experiments, we randomly select 600 images ar training data, and use the rest 1,000 images for testing. To avoid over-fitting while improving model robustness, we augment the training data through left-right flip operations.

The parameters corresponding to the universal saliency map channel and 1×1 *conv* layers for middle layer supervision are initialized with 'xavier'. Using the initialization step in [Pan *et al.*, 2016] and [Kruthiventi *et al.*, 2016], we use the well-trained DeepNet model to initialize the corresponding parameters in our network. The network architecture of our Multi-task CNN is identical to that of DeepNet [Pan *et al.*, 2016] except that i) the parameters corresponding to tasks of different participants are different; ii) middle layer supervision is imposed by adding 1×1 *conv* layer after *conv5* and *conv6*; iii) a channel corresponding to USM is added in the input.

Baselines. Based on the performance of existing methods on the MIT saliency benchmark [Bylinskii *et al.*,] in terms of similarity, we choose LDS [Fang *et al.*, 2016], BMS [Zhang and Sclaroff, 2013], ML-Net [Cornia *et al.*, 2016], and SalNet [Pan *et al.*, 2016] to predict the universal saliency maps on our dataset. The first two methods are based on hand-crafted features, and the latter two are based on deep learning techniques. We use their code provided online to generate USMs.

To validate the effectiveness of our model, we have compared our scheme with several baseline algorithms:

- **RGB based MultiConvNets**: MultiConvNets are trained to predict $\Delta(P_n, I_i)$ for each participant independently, with RGB images as input.
- **RGB based Multi-task**: Multi-task CNN architecture is trained to predict $\Delta(P_n, I_i)$ for all participants simultaneously, with RGB images as input.

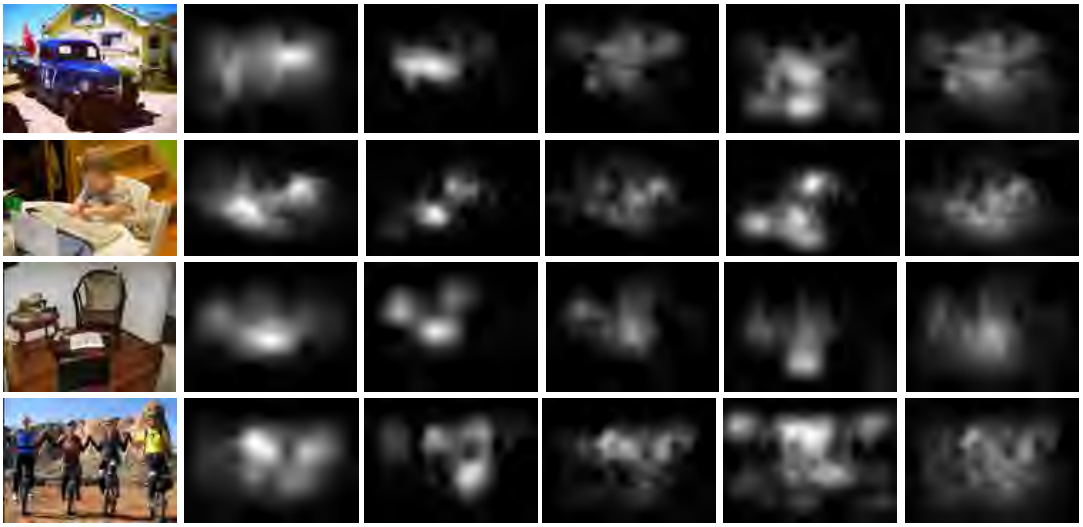

Figure 7: Some images, their ground truth PSM for different persons, and PSM predicted by our approach. The subscript indexes the ID of the participant.

- X**+MultiConvNets**: MultiConvNets are trained to predict $\Delta(P_n, I_i)$ for each participant independently, with RGB images and USM provided by method X as input, where X donates LDS, BMS, ML-Net, and SalNet respectively.

Notice that network architectures of the baseline ones are similar. The major differences are the number of input channels and whether the parameters are shared in the first few layers. For fair comparisons, we have employed the same strategies on data augmentation, middle layers supervision, and parameter initializations.

Measurements. We adopt the same evaluation metrics in [Liu *et al.*, 2015], [Pan *et al.*, 2016] and [Kruthiventi *et al.*, 2016] and choose CC, Similarity, and AUC [Judd *et al.*, 2012] to measure the differences between the predicted saliency map and ground truth.

5.2 Performance Evaluation

The performance of all methods are listed in Table 2. We also show some predicted saliency maps for different participants in Fig. 7. We observe that our solution achieves the best performance in locating the incongruity fixation among individuals. Furthermore, the discrepancy based personalized saliency detection methods consistently outperform directly predicting PSM from RGB images. This validates the effectiveness of our "error correction" strategy for personalized saliency detection. In addition, the multi-task CNN scheme shows higher performance for fixation prediction for individuals tasks than simply training a CNN for each individual.

The effect of supervision on middle layers Fig. 5 shows the accuracy gain from imposing supervision on middle layers in our Multi-task CNN. We observe that middle layer supervision is helpful for PSM prediction in line with previous findings [Lee *et al.*, 2014].

The effect of the number of training samples on the PSM prediction accuracy. Fig. 6 shows that increasing the number of training samples from 200 to 600 (the testing data are fixed) helps to improve the testing accuracy. However, training a more robust deep network requires large-scale training samples which would increase the time complexity tremendously.

6 Conclusion and Future Work

Our work demonstrates that heterogeneity in saliency maps cross individuals is common and critical for reliable saliency prediction, consistent with recent psychology studies showing that saliency is highly specific than universal. We have built the first PSM dataset and presented a framework to model such heterogeneity in terms of the discrepancy between PSM and USM. We have further presented a Multi-task CNN framework for the prediction of this discrepancy. To our knowledge, this is the first comprehensive study on personalized saliency and it is expected to stimulate significant future research.

In our data collection process, each participant needs to observe thousands of images on a single eye-tracker device, which is a bottleneck to increase both the number of images and participants. Clearly additional eye trackers will greatly improve the PSM collection process and can help build an even bigger dataset. Further, a key finding in our study is that personalized saliency is closely related to the observers' personal information (gender, race, major, *etc.*). If we obtain such information in prior, we can directly incorporate it into the PSM prediction to further improve the accuracy and efficiency.

References

[Borji *et al.*, 2014] Ali Borji, Ming Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Eprint Arxiv*, 16(7):3118, 2014.

[Bylinskii *et al.*,] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark.

[Carreira *et al.*, 2015] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. *arXiv preprint arXiv:1507.06550*, 2015.

[Chang *et al.*, 2016] Miko May Lee Chang, Soh Khim Ong, and Andrew Yeh Ching Nee. Automatic information positioning scheme in ar-assisted maintenance based on visual saliency. In *SAIENTO AVR*, pages 453–462. Springer, 2016.

[Cornia *et al.*, 2016] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A deep multi-level network for saliency prediction. *arXiv preprint arXiv:1609.01064*, 2016.

[Fang *et al.*, 2016] S. Fang, J. Li, Y. Tian, T. Huang, and X. Chen. Learning discriminative subspaces on random contrasts for image saliency analysis. *TNNLS*, 2016.

[Gygli *et al.*, 2013] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. The interestingness of images. In *ICCV*, pages 1633–1640, 2013.

[Huang *et al.*, 2015] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *ICCV*, pages 262–270, 2015.

[Itti, 2004] Laurent Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE TIP*, 13(10):1304–1318, 2004.

[Jia *et al.*, 2014] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[Jiang *et al.*, 2015] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *CVPR*, pages 1072–1080, 2015.

[Judd *et al.*, 2009] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *ICCV*, pages 2106–2113, 2009.

[Judd *et al.*, 2012] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012.

[Kruthiventi *et al.*, 2015] Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *arXiv preprint arXiv:1510.02927*, 2015.

[Kruthiventi *et al.*, 2016] Srinivas S. S. Kruthiventi, Vennela Gudisa, Jaley H. Dholakiya, and R. Venkatesh Babu. Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation. In *CVPR*, pages 5781–5790, 2016.

[Lee *et al.*, 2014] Chen Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. *Arxiv*, pages 562–570, 2014.

[Li *et al.*, 2014] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *CVPR*, pages 280–287, 2014.

[Li *et al.*, 2016] X. Li, L. Zhao, L. Wei, M. H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang. Deepsaliency: Multi-task deep neural network model for salient object detection. *TIP*, 25(8):3919–3930, 2016.

[Liu *et al.*, 2015] Nian Liu, Junwei Han, Dingwen Zhang, Shifeng Wen, and Tianming Liu. Predicting eye fixations using convolutional neural networks. In *CVPR*, pages 362–370, 2015.

[Pan *et al.*, 2016] Junting Pan, Elisa Sayrol, Xavier Giroinieto, Kevin Mcguinness, and Noel E. Oconnor. Shallow and deep convolutional networks for saliency prediction. In *CVPR*, pages 598–606, 2016.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015.

[Russell *et al.*, 2008] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 77(1-3):157–173, 2008.

[Setlur *et al.*, 2005] Vidya Setlur, Saeko Takagi, Ramesh Raskar, Michael Gleicher, and Bruce Gooch. Automatic image retargeting. In *MUM*, pages 59–68, 2005.

[Xu *et al.*, 2014] Juan Xu, Ming Jiang, Shuo Wang, Mohan S Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of vision*, 14(1):28–28, 2014.

[Xu *et al.*, 2015] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

[Zhang and Sclaroff, 2013] Jianming Zhang and Stan Sclaroff. Saliency detection: A boolean map approach. In *ICCV*, pages 153–160, 2013.

[Zhang *et al.*, 2014] Zhanpeng Zhang, Ping Luo, Change Loy Chen, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, pages 94–108, 2014.