Personalized Visual Attention Modelling

Miroslav Čulík*
Supervised by: Miroslav Laco†

Faculty of Informatics and Information Technologies
Slovak University of Technology, Bratislava

Abstract

The vast majority of models predicting visual attention are not designed to take observer-specific information into account. However, some models that use these features could be useful in predicting personalized attention, what can provide space for categorizing observers based on their visual attention properties. We introduce two variants of personalized models built upon the generalized model of Convolutional Neural Network (CNN). Predictions of these two personalized variants were evaluated and compared with the predictions of the generalized model. Our observations indicate that better results were produced by the model predicting personalized saliency maps for a particular observer.

Keywords: visual attention, saliency, personalized visual attention modelling

1 Introduction

In recent decades, much attention in the field of computer vision has been paid to examination of the human visual attention, especially in defining the generalized concepts of human visual attention and modelling it. Those models are beneficial in capturing common patterns in visual attention across all the individuals but lack identifying the individual variations specific for each subject.

Apart from bottom-up and top-down factors, there are also individual factors, which can influence the selection of visual stimuli of a given observer. For instance, the human eye movements are found to be idiosyncratic, which means they are more consistent within an individual than between individuals [1]. Other heterogeneous factors in human visual perception that can be considered are identity, age, and the sensitivity to distracting stimuli [11].

Models predicting personalized visual attention can take these factors into account in various ways. Unlike the models that predict generalized visual attention, personalized models of visual attention address the problem of visual attention prediction for a particular observer with their specifics included in the process of prediction personalized saliency maps.

*xculikm@stuba.sk
†miroslav.laco@stuba.sk

The topic of personalized visual attention emerged in a period when methods of deep learning demonstrated the best results in similar topic of generalized visual attention. Inspired by this, researchers designing personalized visual attention models have chosen the path of deep learning and neural networks. For instance, convolutional neural networks are used as a traditional approach in the prediction of personalized saliency maps.

Many convolutional networks use already trained layers from other networks, which were trained on similar tasks. In practice, it means adopting the weights of particular layers or whole network to be part of a network that will be re-trained on a similar task [15]. This concept is also called transfer learning.

In this study, our goal is to develop personalized models predicting personalized saliency maps for a given observer and show that these models can achieve better results, compared to predictions for a given observer made by a generalized model. In Section 2, we review the most successful existing approaches in the prediction task of personalized saliency maps. In Section 3, we describe the proposed approach in the prediction of personalized saliency maps using the concept of transfer learning. Based on that, we provide a discussion on gained results in Section 4, which allow us to use this concept in application area (as stated in Section 6).

2 Related Work

An example of partial transfer learning can be found in the work of Li and Chen [11], where the authors used pre-trained convolutional and inception layers from networks VGG-16 [16] and GoogLeNet [17]. These layers were used to extract deep features representing global information. They also employed robust multi-task learning (RMTL) framework [6] designed to incorporate the individuality features to predict the outlying values of those individuals who were not related to the majority in viewing behaviour.

Also, Xu et al. [19] developed a multi-task framework with CNN for personalized saliency prediction for multiple subjects. Using four shared convolutional layers, the multi-task section treats the prediction of personalized saliency maps for different subjects as separate tasks. Moreover, Xu et al. [19] proposed another approach in the

personalized saliency map prediction when their network consisted of seven convolutional layers while the output of the fourth convolutional layer was convolved with person-specific information encoded into filters.

Lin and Hui [12] predicted the personalized saliency maps with a developed two-stream convolutional network. One stream predicts saliency for the input image with the use of VGG-16 [16] without the last fully-connected layer, which was used as a multi-scale feature extractor. Also, the outputs from the customized layers of Single Shot multi-box Detector (SSD) [13] are concatenated with multi-scale features from the adopted VGG-16 model in the process of saliency detection. The SSD module was also used in the second stream with the responsibility for the training of subject preference vectors. In creating the personalized saliency map, outputs of both streams were merged, blurred by Gaussian filter, and center saliency prior was added.

Yu and Clark [20] developed the Generative Adversarial Networks (GAN) combining the advantages of Conditional GAN [14] and StackGAN [21]. The network consists of a generator, which tries to predict the personalized saliency map for a given subject with the specific information encoded (user-specific information were two binary classes of age and language group) and the discriminator which has to distinguish between the real personalized saliency map (ground-truth personalized saliency map) from a fake personalized saliency map (predicted personalized saliency map).

3 Proposed Method

In order to predict personalized visual attention as a prediction task of personalized saliency maps for a particular user, we introduce our proposed method in the modelling personalized visual attention with a machine learning approach, specifically with deep neural networks. Based on results from generalized saliency benchmarks [4] [9], we observe that models of deep neural networks achieve far better results than classical models with hand-crafted features. Deep neural networks can extract features from image data in multiple scales and with numerous tunable parameters in their layers, which helps to secure the generalizability of these models [2].

With these facts in mind, we need to choose datasets suitable for these tasks, corresponding architecture, and also to specify reasonable methodology of training. We decided to reuse the existing model of Convolutional Neural Network developed by Kroner et al. [8], that predicts generalized saliency maps and with the usage of transfer learning re-train this model to predict personalized saliency maps for each participant as a separate task.

With this proposed method, our ultimate goal is to achieve better results with our predicted personalized saliency maps against ground-truth personalized saliency maps for every participant than with predicted generalized saliency maps. Using popular saliency metrics [5], we compare the model predicting generalized saliency maps with each model predicting personalized saliency maps for a given participant.

3.1 Used Datasets

In our experiment, we choose two datasets for the prediction of generalized saliency maps and one dataset for the prediction of personalized saliency maps for a particular subject. The chosen datasets are:

1. SALICON dataset [7]
2. CAT2000 dataset [3]
3. Personalized Saliency Dataset (PSD) [19]

SALICON dataset [7], is used in the initial phase of the training of a model, which predicts generalized saliency maps. Despite the disadvantages of the mouse-tracking data collection (lower inter-participant congruency or less accurate ground-truth generalized saliency maps compared with the eye-tracking data and ground-truth generalized saliency maps created from these data [18]), these data can be used in the training process of generalized saliency models, which can lead to slightly worse results when compared to models trained on eye-tracking data [18]. Nevertheless, this dataset contains 15000 training samples and belongs to the largest datasets in the field. Even if the mouse-tracking data cannot fully replace standard eye-tracking data, they can serve as a coarse estimation of eye movements in a generalized saliency prediction task. Since we want to preserve the original training procedure as the authors of the re-used model that predicted generalized saliency maps [8] we use this dataset in the initial training step of the generalized model.

CAT2000 dataset [3], is used in the re-training phase after the model predicting generalized saliency maps was initially trained on the SALICON dataset. We choose the CAT2000 dataset because it is frequently used in the task of generalized saliency maps prediction and is part of both generalized saliency benchmarks [4] [9]. Moreover, it contains a sufficient number of input stimuli (4000) separated into 20 different categories. As these data were collected from standard eye-trackers, they provide a decent opportunity to fine-tune this re-used model on standard eye-tracking data.

Personalized Saliency Dataset (PSD) [19], is used in another re-training of a fine-tuned model that can predict generalized saliency maps. In order to predict the personalized saliency maps of a given subject using Convolutional Neural Network, we need enough eye-tracking data for each participant. The PSD dataset contains 1600 personalized saliency maps for each of 30 subjects collected from standard eye-trackers. The majority of input stimuli from the PSD dataset [19] contains multiple objects captured in different environments, which can contribute to

the detection of personalized aspects of visual attention of particular subjects. To our knowledge, there is no larger dataset in the context of the number of subjects and the number of ground-truth personalized saliency maps per participant. This leads us to the decision that this dataset is our first choice in personalized saliency modelling.

3.2 Proposed Model

We address the problem of personalized visual attention prediction as supervised learning based on Convolutional Neural Network (CNN) and transfer learning. In addition, all the models which predict personalized saliency maps mentioned in Section 2 used deep learning concepts in their solutions with notable results.

At first, we train the CNN on a generalized visual attention prediction task, represented as a prediction of the generalized saliency map. After fine-tuning, this generalized CNN model serves as a base for a personalized visual attention prediction task, represented as a prediction of a personalized saliency map for a particular subject. For each participant, we re-train fine-tuned generalized CNN model to be able to predict personalized saliency maps for a given participant. For this purpose, we experiment with two variations of input:

1. Input image with generalized saliency map (depicted in Fig. 1)

2. Input image only (same architecture as in Fig. 1, but only RGB image is taken as input)

In **the first variant**, we decide to predict personalized saliency maps with the prediction of a generalized saliency map for a given image, because we consider a generalized saliency map as a rough estimation of a personalized saliency map. This idea was presented in the work of Xu et al. [19] and might be helpful in the process of prediction of personalized saliency map using our method. In **the second variant** without a generalized saliency map in the input, we want to observe how much a generalized saliency map affects the process of a prediction personalized saliency map for each participant. Finally, we want to compare both variants of input with respect to each participant.

However, if we want to predict personalized saliency maps for a new participant not included in the training personalized dataset, we have to collect the eye-tracking data when hundreds of stimuli were first viewed. Then we can preprocess these data to get continuous ground-truth and valid binary fixations maps containing the participant's fixation locations. Finally, we can advance to the step of re-training model predicting generalized saliency maps, to predict personalized saliency maps for the new participant. This process itself is very time and resource-consuming, which leads to a search for alternatives in such a use case.

Architecture

We used the design and implementation of the encoder-decoder Convolutional Neural Network from Kroner et. al [8]. On one hand, we chose this model because it achieves competitive results in both generalized saliency benchmarks [4] [9] on multiple metrics. Moreover, the model employs convolutional layers from VGG-16 [16] trained on object and scene classification datasets which serve as a feature extractor. The network itself is computationally efficient in comparison to other well-performing models and can be trained in limited conditions. Additionally, the authors [8] designed the architecture of the model to work with "multi-scale information and global context based on semantic feature representations" [8] providing significantly better results.

On the other hand, in more complex scenes with multiple objects, the network fails to detect salient regions when human faces blend with other objects or background or in cases where small text or low-level feature contrast is present. These disadvantages can be particularly improved by using convolutional layers of the better performing object classification model in the feature extraction process [8].

The network consists of three modules:

1. *Encoder* - includes 18 layers, where 10 classical convolutional layers are without dilation rate and 3 convolutional layers have dilation rate of 2 and 5 max-pooling layers. Fig. 1 illustrates this part from the beginning of the network to the ASPP module.

2. *ASPP (Atrous Spatial Pyramid Pooling) module* - this module uses five branches in parallel with the different setting of convolutional layers. Three convolutional layers have dilation rates of 4, 8, and 12; one classic convolutional layer was without dilation rate, and one classic convolutional layer followed by an upsampling layer. The outputs of those five parallel branches are concatenated and fed into the convolutional layer, which is the last one in this module (in Fig. 1, this module is highlighted in yellow and also contains a convolutional layer right behind it).

3. *Decoder* - includes seven layers: four convolutional layers used to prevent the checkerboard effect and three upsampling layers realized by the bilinear upsampling method to increase spatial information (in Fig. 1 starting from the left upsampling layer to the end of the network).

Kullback-Leibler Divergence is used in the adopted network [8] is used as an error (loss) function and ReLU (Rectified Linear Unit) as an activation function in every layer, except the last one. The learning algorithm used is mini-batch stochastic gradient descent with Adam Optimizer. Our intention is to add several metrics that would be able to evaluate the current progress of the network during every epoch.

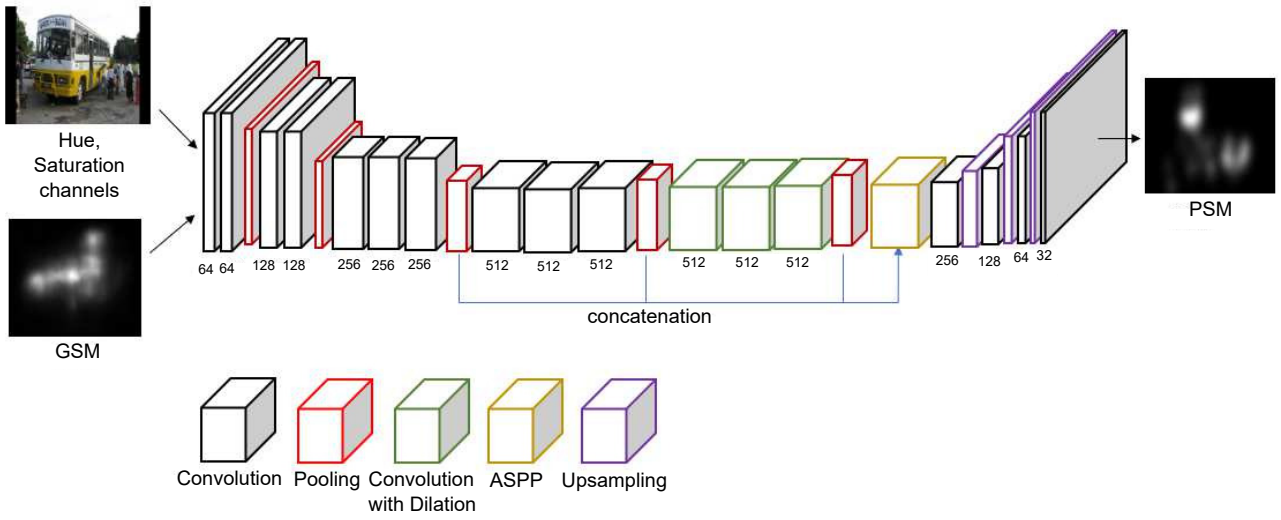

Figure 1: Graphical representation of the inputs used in the prediction of personalized saliency maps. Input image is transformed from RGB into HSV colour space and only Hue and Saturation channels are taken to be concatenated with gray-scale generalized saliency map.

3.3 Proposed Methodology of Training

In the training process of the model that predicts personalized saliency maps for a particular participant, we define three main steps:

1. Initial training of the generalized model on SALICON dataset [7]

2. Re-training and fine-tuning of the generalized model on CAT2000 dataset [3]

3. Re-training of the personalized model on the PSD dataset [19]

The previously described encoder-decoder CNN [8] is adopted and generalized saliency maps are predicted for input stimuli with this network. As the training of this network requires **initial training on the SALICON dataset [7]**, we have to proceed with this initial training. Input data are split in a ratio 67:33 (train set:valid set) as suggested by the authors [8]. Weights from pre-trained convolutional layers of VGG-16 [16] are used as the initial weights.

Using the network trained initially on the Salicon dataset [7], we continue in **training with data from the CAT2000 dataset [3]**, where input images are used as input and predicted generalized saliency maps are expected as output. Those predictions are confronted with a ground-truth generalized saliency map for a given input image. In training, the data split ratio is 80:20 (train set:valid set) as suggested by the authors [8]. In order to fine-tune the best performing model, manual hyperparameter tuning is done and the weights of the best model are saved for the third step.

Restoring the checkpoint of network weights gaining the best results from the previous step, we move to the **prediction of personalized saliency maps** for each of 30 subjects. To fine-tune every personalized model for each participant, we use the Random Search technique in order to find the best combination of hyper-parameters. We consider learning rate and batch size as two hyper-parameters that need to be found. Additionally, we want to intelligently stop the training process of the personalized model by adding two early stopping rules checked after each epoch of training. The first stopping rule is used when more than two epochs pass and minimal validation error is not gained in the last three epochs. The second early stopping rule is used when more than five epochs passed and the absolute value of the difference between validation error and train error from the last epoch is higher than our defined threshold value of 0.08. This threshold value is defined by the rule of thumb and helps in over-fitting prevention of trained personalized model. During the step of re-training the generalized model for the prediction personalized saliency maps for given subject, we propose two variants, which are marked as variant 1 and variant 2 in Fig. 2.

Firstly, input images taken from Personalized Saliency Dataset (PSD) [19] are converted from RGB into HSV colour space. Only H (Hue) and S (Saturation) channels are taken from the converted input image and these are concatenated with a gray-scale generalized saliency map

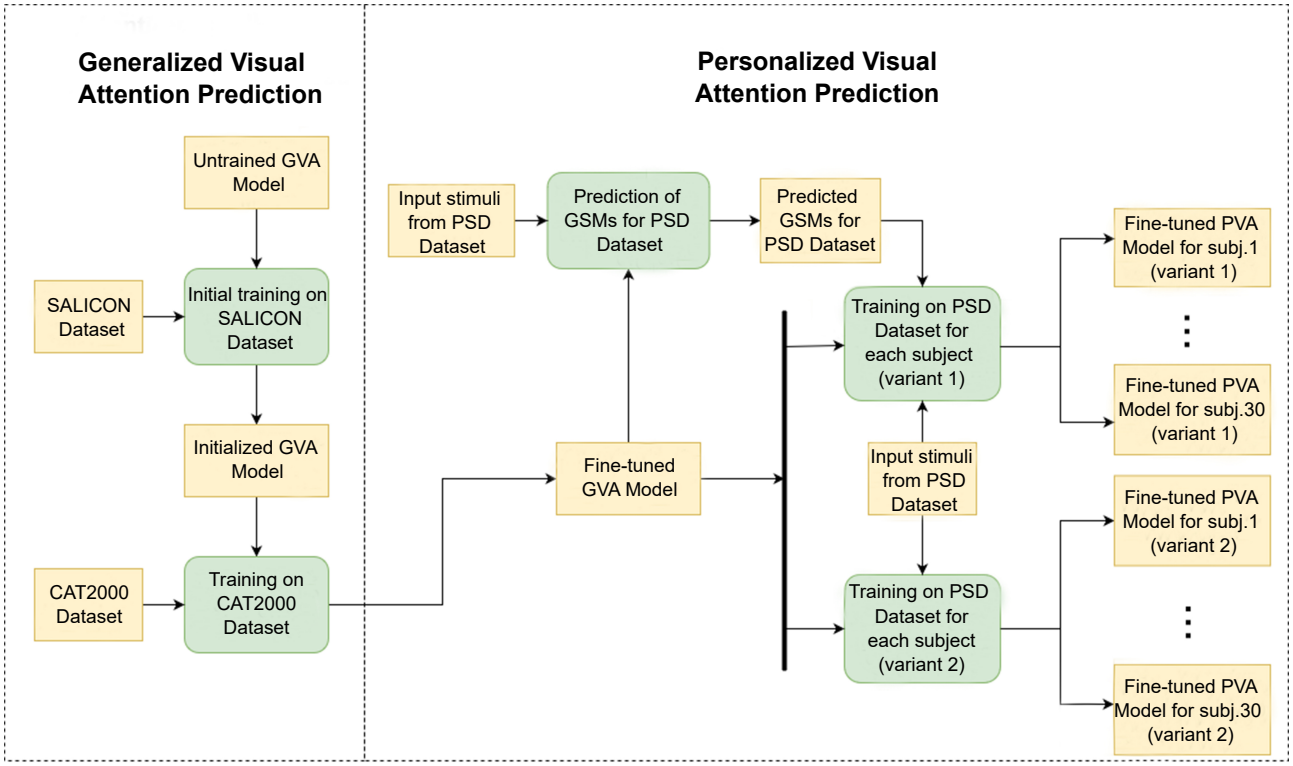

Figure 2: A proposed flow of training describes the whole training process from initial training on the SALICON dataset [7] followed by re-training the generalized visual attention (GVA) model on the CAT2000 dataset [3]. This fine-tuned generalized model predicts generalized saliency maps (GSMs) for input stimuli from the PSD dataset [19], which are used together as the first variant of input in re-training of the fine-tuned generalized model to predict personalized saliency maps for a given subject using personalized visual attention (PVA) model. In the second variant, only input stimulus is used for re-training of this generalized model.

for a given input image (as described in Fig. 1). These generalized saliency maps are generated by the fine-tuned network from previous generalized saliency maps prediction task, but this time for input images from the PSD dataset [19].

Secondly, as an alternative option, only three-channel RGB images are used as input. In both above-mentioned cases, a personalized saliency map for a particular subject is expected as the network output and confronted with ground-truth personalized saliency map of this subject. Process in this step is repeated for each participant with both variations of input which results in training two models for each of 30 participants present in the PSD dataset [19]. Data in this training step are split in ratio 80:20:10 (train set:valid set:test set). In both variants, we decide to have three-channel input because we want to preserve trained weights from all the layers of the fine-tuned generalized model.

4 Results

After we fine-tuned the model that predicts generalized saliency maps, we predicted generalized saliency maps for all the stimuli from the PSD dataset [19], which was used during the whole process of personalized visual attention prediction.

4.1 Quantitative results

In order to predict personalized saliency maps for each of 30 observers, these predicted generalized saliency maps were used as part of the input together with input stimuli in the first variant (marked as variant 1 in Fig. 2). In the second variant (marked as variant 2 in Fig. 2), we used only input stimuli as the input to the model. The predicted personalized saliency maps were confronted with ground-truth saliency maps of the given observer and given input stimulus. After fine-tuning of each personalized model, we predicted 160 personalized saliency maps from the test set, which were compared to corresponding ground-truth personalized saliency maps for a given participant using selected metrics. Depending on the type of metric, we used discrete ground-truth binary fixation maps for location-based metrics (AUC metrics, NSS and IG) and continuous ground-truth maps for distribution-based metrics (SIM, CC, and KLD) [5]. Although we used nine metrics in total, we considered AUC Judd metric (calculated

for multiple different thresholds as number of pixels that overlap with ground-truth binary fixation map and total number of incorrectly non-fixated pixels) and Information Gain over generalized saliency map (IG GSM; describes how much information in bits brings predicted personalized saliency map over the generalized saliency map on fixation locations [10]) as the most important metrics for our study.

Predicted generalized saliency maps for input stimuli from the PSD dataset [19] also served as a prediction of average observer's visual attention and their test set predictions were also used in the evaluation. As we can see in Table 1, median scores across all 30 participants show that our first variant of the personalized model (with generalized saliency map as the input) achieves the best results in seven out of nine selected metrics. The two remaining best results in Normalized Saliency Scan-path (NSS) and Information Gain over center-prior map were achieved by generalized model.

Table 1: Results of models predictions compared with ground-truth personalized saliency maps for given participants. Values of each metric are calculated as the median of results for 30 participants from the PSD dataset [19].

Metric	GVA model [8]	PVA model variant 1	PVA model variant 2
AUC Judd(↑)	0.891	**0.896**	0.88
AUC Borji(↑)	0.829	**0.853**	0.834
AUC Shuffled(↑)	0.681	**0.71**	0.676
NSS (↑)	**2.432**	2.417	2.128
SIM (↑)	0.612	**0.653**	0.618
CC (↑)	0.703	**0.765**	0.697
KLD (↓)	0.591	**0.513**	0.58
IG Center Prior(↑)	**1.279**	1.25	1.094
IG GSM(↑)	0	**0.012**	-0.153

Worth noting is the fact, that the median score across all observers from the PSD dataset [19] is higher in all metrics in favour of the first variant of our personalized saliency model which contains a generalized saliency map in the input, while the second variant of the personalized saliency model does not contain this generalized map. As we can see in Table 2, the second variant of our personalized model gained the best score only once (AUC Borji metric) with only one observer, while all other best scores were split between the first variant of our personalized model and the generalized model. Except NSS and both IG metrics, the first variant of a personalized model achieved the best results with the remaining metrics in at least 23 out of 30 observers, which is more than 76% observers for each of those metrics.

In NSS and both IG metrics, the first variant of the personalized model achieved the best score for 16 (NSS) and 17 (both IG metrics) observers respectively, while the generalized model achieved the best results for the remaining 14 and 13 observers respectively. However, the median score of these three metrics across all the participants (see Table 1) showed a higher score of NSS and IG over center-prior baseline map. In the case of NSS, which is very sensitive to false positives, we interpret this phenomenon as the inability of the personalized model to predict fewer false positives, because of the big impact of generalized saliency map in the first variant of personalized predictions (as pointed out in the last example in the subsection 4.2). Similarly to NSS, scores of IG over center-prior baseline map also indicate that the first variant of personalized models relies heavily on generalized saliency maps used in this variant as part of the input.

Table 2: Number of best resulting participant samples for each model throughout all participants from the PSD dataset [19] and metrics. In metric AUC-Judd we evaluated the results of three participants as equal.

Metric	GVA model [8]	PVA model variant 1	PVA model variant 2
AUC Judd(↑)	10	**23**	0
AUC Borji(↑)	2	**27**	1
AUC Shuffled(↑)	5	**25**	1
NSS (↑)	14	**16**	0
SIM (↑)	1	**29**	0
CC (↑)	1	**29**	0
KLD (↓)	3	**27**	0
IG Center Prior(↑)	13	**17**	0
IG GSM(↑)	13	**17**	0

4.2 Qualitative results

As shown in the bottom part of Fig. 3, both personalized variants successfully detected human faces, which are considered the most salient in general. Unlike the predicted generalized saliency map, both personalized predictions did not mark the region in the centre of the stimulus as salient. However, the first variant of the personalized saliency map copies the pattern of emphasizing human faces in the background (as it is in a generalized saliency map), while the ground-truth personalized saliency map rates faces in the foreground as more salient.

Comparing both variants of personalized saliency maps, we can observe that the second variant of personalized saliency maps produces more biased predictions. This could have been caused by the non-existence of coarse estimation as it was in the case of the first personalized variant. From the bottom part of Fig. 3, we can observe, that light background and multiple amounts of objects in the scene resulted in very biased predictions in both personalized variants. Generalized prediction in this scenario managed to indicate salient objects, but only in the central region of the image.

If we compare personalized predictions of successful first variant between different observers, who have different visual behaviour, we can observe that this personalized variant is very dependent on the similarity of predicted generalized saliency map and ground-truth personalized

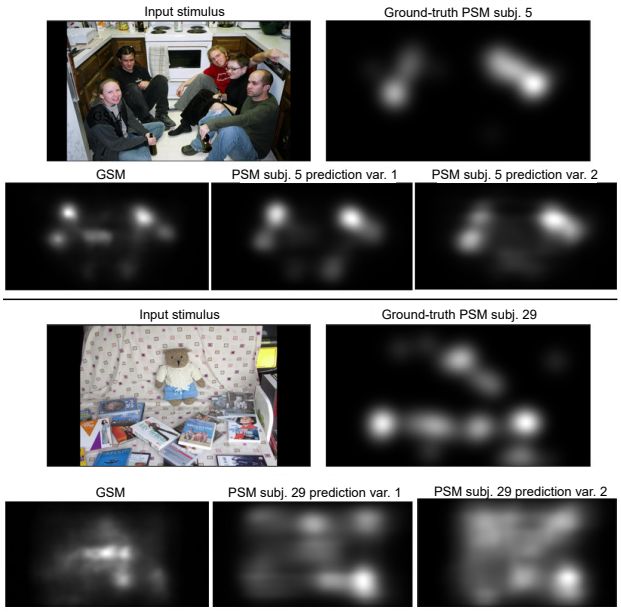

Figure 3: Comparison of both variants of predicted Personalized Saliency Maps (PSM) using our proposed PVA models for observers labeled 5 and 29 with input stimuli and predicted Generalized Saliency Maps (GSM) using GVA model [8].

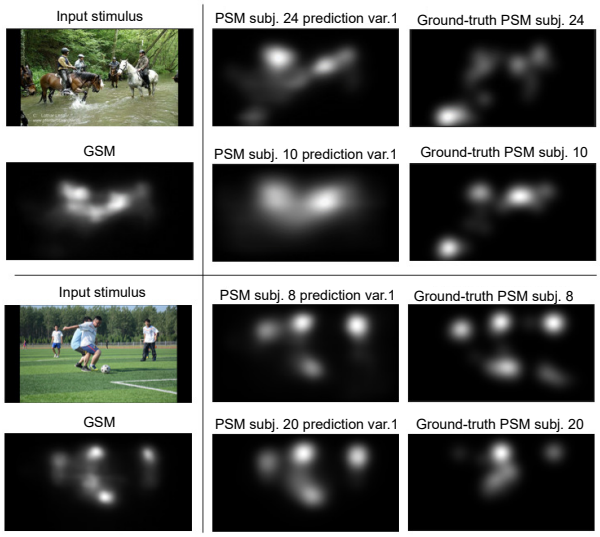

Figure 4: Comparison of predicted Personalized Saliency Maps (PSM) for observers 24, 10, 8, and 20 with input stimuli and predicted Generalized Saliency Maps (GSM).

saliency map, as illustrated in Fig. 4. In the top part, we can see that subject 24 focused more on light photo caption than on horse riders. While subject 10 also registered this photo caption, he paid more attention to the horse rider on the right side of the input stimulus. However, predictions of personalized saliency maps for these subjects were assessed as salient only when objects were present in the central region and left out light-coloured photo caption.

In the bottom part of Fig. 4, we can see the example of two subjects who observed the same image in different ways. While visual attention of subject 8 was attracted by human faces, ball, and lines on the ground, subject 20 observed the player's face located in the foreground and body posture and face of a player on the right. The predicted personalized saliency maps are, however, almost identical, as the generalized saliency map forced both personalized models to mark faces and the ball as the salient regions. We interpret this result as the inability of the personalized model to leave out the regions, which are marked as salient by generalized saliency map, but the subject considers these regions as indistinct.

5 Conclusion

We designed the method of re-training of the Convolutional Neural Network (the Encoder-Decoder type) originally predicting generalized visual attention to predict personalized visual attention using the concept of transfer learning. Regarding personalized visual attention, we experimented with two variants of input, one with a predicted generalized saliency map for a given input image and the other without it. In order to find out, whether our personalized models achieve better results for a given participant, we performed evaluation on a set of 30 participants. We observed the improved results in seven out of nine selected metrics for one variant of our personalized approach, which indicates, that our models producing personalized saliency maps can better capture salient regions specific to the individual subject.

We also found out, that our better performing personalized approach is highly dependent on the generalized saliency map incorporated in the input. This leads to different results across observers with different visual behaviour. If mentioned generalized saliency map for given image shares similarities with a personalized ground-truth saliency map for a given observer, then a personalized model that predicts personalized saliency maps can outperform the generalized model in the context of personalized saliency prediction. However, if this generalized saliency map compared to the personalized ground-truth saliency map contains significant dissimilarities (the visual patterns of a given observer are different from patterns of the majority of observers), it often leads to inaccurate detection of the salient region for a given observer. These inaccuracies result in worse predictions than predictions produced by the generalized model but are still better than results of the personalized model without this generalized saliency map as part of the input.

6 Future Work

The field of medical diagnostics, concretely diagnostics of early stages of Alzheimer's disease, creates demand for personalized visual attention models. In response to this demand and with achieved results, we plan to develop our own deep-learning model designed to classify individuals into two groups (patients with Alzheimer's disease and healthy people) based on their ground-truth eye-tracking data. For further evaluation, we aim to retrain generalized model [8] on the aggregated maps of all the participants from the PSD dataset [19] and compare its generalized predictions with the personalized saliency maps for these participants separately. Post-processing methods for predicted personalized saliency maps are also considered.

References

[1] Nicola Anderson, Fraser Anderson, Alan Kingstone, and Walter Bischof. A comparison of scanpath comparison methods. *Behavior research methods*, 47(4):1377–1392, 2015.

[2] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Computational Visual Media*, 5(2):117–150, 2019.

[3] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*, 2015.

[4] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark. http://saliency.mit.edu/.

[5] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018.

[6] Jianhui Chen, Jiayu Zhou, and Jieping Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 42–50, 2011.

[7] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015.

[8] Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. Contextual encoder-decoder network for visual saliency prediction. *Neural Networks*, 2020.

[9] Matthias Kümmerer, Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit/tübingen saliency benchmark. https://saliency.tuebingen.ai/.

[10] Matthias Kümmerer, Thomas Wallis, and Matthias Bethge. How close are we to understanding image-based saliency? *arXiv preprint arXiv:1409.7686*, 2014.

[11] Aoqi Li and Zhenzhong Chen. Personalized visual saliency: Individuality affects image perception. *IEEE Access*, 6:16099–16109, 2018.

[12] Sikun Lin and Pan Hui. Where's your focus: Personalized attention. *arXiv preprint arXiv:1802.07931*, 2018.

[13] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[14] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[15] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[18] Hamed R Tavakoli, Fawad Ahmed, Ali Borji, and Jorma Laaksonen. Saliency revisited: Analysis of mouse movements versus fixations. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 1774–1782, 2017.

[19] Yanyu Xu, Shenghua Gao, Junru Wu, Nianyi Li, and Jingyi Yu. Personalized saliency and its prediction. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2975–2989, 2018.

[20] Bingqing Yu and James J Clark. Personalization of saliency estimation. *arXiv preprint arXiv:1711.08000*, 2017.

[21] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.