Clustered Saliency Prediction

Rezvan Sherkati
rezvan.sherkati@mail.mcgill.ca
James J. Clark
james.clark1@mcgill.ca

Electrical and Computer Engineering Department,
McGill University
845 Sherbrooke St W,
Montreal, Canada

Abstract

We present a new method for image salience prediction, Clustered Saliency Prediction. This method divides subjects into clusters based on their personal features and their known saliency maps, and generates an image salience model conditioned on the cluster label. We test our approach on a public dataset of personalized saliency maps and cluster the subjects using selected importance weights for personal feature factors. We propose the Multi-Domain Saliency Translation model which uses image stimuli and universal saliency maps to predict saliency maps for each cluster. For obtaining universal saliency maps, we applied various state-of-the-art methods, DeepGaze IIE, ML-Net and SalGAN, and compared their effectiveness in our system. We show that our Clustered Saliency Prediction technique outperforms the universal saliency prediction models. Also, we demonstrate the effectiveness of our clustering method by comparing the results of Clustered Saliency Prediction using clusters obtained by our algorithm with some baseline methods. Finally, we propose an approach to assign new people to their most appropriate cluster and prove its usefulness in the experiments.

1 Introduction

When humans look at an image, they might fixate on some points (i.e. fixation points). Fixation points carry a lot of information, such as features and the important events happening in the image. They can also reflect on the personality traits of viewers. For this reason, predicting fixation points in an image, i.e. saliency prediction, has been an important research problem for decades. One of the early works in universal saliency prediction is [9] which started an era of several works in this area. With great improvements in the area of deep neural networks, there has been a lot of progress made in this subject ([2]). While there has been significant progress in the field of saliency prediction, there are still several flaws. An important concern is the difference of the saliency maps across different individuals or groups. Many factors such as personal features and biases might affect one's fixation points ([23]). Thus, it is essential to study saliency prediction from a personalized point of view, i.e. personalized saliency prediction.

In this paper, to study saliency prediction from a more personalized perspective, we introduce Clustered Saliency Prediction, which we define as the prediction of saliency maps for groups of similar subjects. We develop methods to first, group subjects based on some

personal features and their previous available saliency data (if any exists). Then, considering this clustering of subjects, we learn to predict the saliency maps of subjects.

The first part of our approach for predicting saliency is to find an appropriate clustering of the subjects. We propose the Subject Similarity Clustering (SSC) method, which builds a network using personalized saliency maps of subjects and their personal features. Then, we use the Louvain community detection algorithm on this network to find the clusters. The reason behind using a network structure is to take into account all the factors in relation to each other in a single entity. After putting subjects in appropriate clusters, we propose the Multi-Domain Saliency Translation (MDST) model for predicting saliency maps for each cluster. We show the improvements of Clustered Saliency Prediction over some existing universal saliency prediction methods in some experiments. We also propose a method to assign new people to the most appropriate cluster, using their personal features and any of their known saliency maps. We demonstrate this method's effectiveness in choosing the best cluster in our experiments.

There are multiple advantages to Clustered Saliency Prediction over individually personalized saliency prediction. First, by aggregating the saliency maps of similar subjects in clusters, we omit the unimportant noises and catch the main themes in the saliency patterns. Also, by putting subjects in clusters based on their saliency maps and personal features, we are actually assigning them to communities and we get better knowledge of the subjects which we can use for further applications, such as recommendation systems. Another advantage is that by putting subjects in clusters, their privacy would be further preserved, since it will be more difficult to associate an output with a specific user. This clustering approach was not used by prior works, and is a contribution of our paper.

In summary, our contributions in this paper are: 1) Proposing a new clustering method, Subject Similarity Clustering, using a network based structure and Louvain community detection algorithm, 2) Proposing the MDST model to convert universal saliency predictions obtained by previous methods, to saliency maps of the clusters, 3) Conducting experiments on a publicly available dataset containing personalized saliency maps, and demonstrating the superiority of our results to some existing saliency prediction methods, 4) Demonstrating the effectiveness of our clustering method in improving prediction of saliency maps, by comparing the results with some baseline cases, 5) Proposing a method to assign new people to their closest cluster, using their personal features and any of their available saliency maps and prove its usefulness.

2 Related work

Universal saliency prediction. Some of the first works in the area of saliency prediction are [9, 11, 21]. These early models were mostly based on extracting simple feature maps from the images. In [9], they have focused on extracting high-level image features such as faces, text elements, etc. and the extent to which they attract attention. Some of the first Deep Neural Network (DNN) based works on saliency prediction are eDN model ([22]) and DeepGaze I ([12]). In DeepGaze I, they show that deep convolutional networks that have been trained on computer vision tasks such as object detection, boost saliency prediction. In [13], authors introduce the DeepGaze II model for universal saliency prediction that uses transfer learning from the VGG-19 network to achieve a good performance. In a more recent work, authors in [18] proposed DeepGaze IIE as an improvement over DeepGaze II. In DeepGaze IIE, they replaced the VGG19 backbone with ResNet50 features ([6]), which provides a big

improvement on saliency prediction. Also, ML-Net ([5]) is a deep-learning based architecture for predicting universal saliency maps. This model uses multi-level features extracted from a Convolutional Neural Network (CNN) and it is end-to-end trainable. The model SalGAN ([20]), takes advantage of Generative Adversarial Networks (GAN), which consists of a VGG-16 based encoder-decoder model as the generator, and a discriminator.

Personalized saliency prediction. In [24], authors have produced a dataset of Personalized Saliency Maps (PSMs). We use this dataset in our experiments. In [24], they model PSMs based on Universal Saliency Maps (USMs) shared by different participants and adopt a multitask CNN framework to estimate the discrepancy between PSMs and USMs. In [25], which is an extended version of [24], they similarly decompose a PSM into a USM predictable by previous saliency detection models and a new discrepancy map across users that characterizes personalized saliency. Then, they present a new solution in addition to their previous work towards predicting such discrepancy maps. In [17], the authors develop Personalized Attention Network (PANet), which contains two streams of CNNs that share common feature extraction layers. One of the main limitations in personalized saliency prediction is the lack of large saliency datasets for each subject. In [19], the authors proposed few-shot personalized saliency prediction using a small amount of training data based on Adaptive Image Selection (AIS) considering object and visual attention. In a more recent work in [7], the proposed model is composed of universal saliency prediction and personalized gaze probability prediction modules and then the results of these two modules are integrated to generate a final saliency map.

3 Methods

In this section, we first describe the dataset and the evaluation metrics that we use in our experiments. After, we describe our clustering method. Then, we talk about the universal saliency prediction methods that we use and our proposed MDST model for obtaining the clusters' saliency maps. To easier understand the approach, we explain its usage on the PSM dataset from [24], but it can be adapted to other personalized saliency map datasets as well.

Datasets used in experiments. In order to test our approaches, we used the dataset collected in [24], which contains personalized saliency maps. The dataset consists of 1600 images with multiple semantic annotations, which were observed by 30 student participants (14 males, 16 females, aged between 20 and 25). The dataset also includes a survey to collect each observer's personal information. Specifically, the following information is collected - the observer's gender (1D), the preference to objects falling into the fashion category (ring, necklace, etc., 11D), the preference/disgust to colors (red, yellow, etc., 16D), the preference to different sports (football, etc., 11D), and the preference to objects falling into other categories (IT, plant, etc., 4D). You can see more details of the dataset in [24]. Subjects responded to each feature with 0 or 1. This yields a 43-dimensional vector for personal features of each subject in five categories: Gender, Fashion, Color, Sport and Other. We will refer to this dataset as the *PSM dataset*.

Evaluation metrics. For a review of many saliency evaluation metrics, see [3] and [16]. To evaluate the performance of our methods, we use Pearson's Correlation Coefficient (CC), Similarity (SIM), Normalized Scanpath Saliency (NSS) and Area under ROC Curve (AUC) Judd metrics.

3.1 Clustering

Here, we focus on creating a network structure based on subjects' information and clustering subjects using this network. To cluster the subjects, we create a weighted complex network of the subjects in the dataset using images seen by the subjects and their personal features. Then, we use a community detection method ([14]) on this weighted network, called the Louvain algorithm ([1]), for dividing the subjects into groups. We call this clustering algorithm *Subject Similarity Clustering (SSC)*. Now, we describe the SSC algorithm.

First, we initiate a network called *Subject Similarity Network (SSN)* with 30 nodes (N_1, ..., N_{30}), corresponding to subjects ($P_1, \ldots, P_{30}$) in the PSM dataset respectively. We denote the weight of the edge between node u and node v as $W(u,v)$, initialized to zero. We denote the personalized saliency map for person P_i and image x as $S_{PSM}(P_i,x)$, which are normalized to have pixel values in [0,1]. All saliency maps are resized to the same size, with resolution R. For personal features of subjects, we denote the feature vector of person P with $F(P)$ and the subset of feature vector for feature category C with $F_C(P)$, e.g. $F_{Gender}(P)$. Also, for feature category C, $|C|$ denotes the number of features in this category, e.g. $|Gender| = 1$. For feature categories we consider the feature weights W_{Gender}, $W_{Fashion}$, W_{Color}, W_{Sport} and W_{Other}. Now, for each pair of nodes N_i and N_j in SSN (order of the nodes does not matter, since SSN is undirected), we add an edge between them with the weight:

$$W(N_i,N_j) = \frac{m_{i,j}}{\sum_{x\in I_{i,j}} \frac{\|S_{PSM}(P_i,x) - S_{PSM}(P_j,x)\|_1 + 1}{R}} + \sum_{C\in FCATS} \frac{W_C \times |C|}{\|F_C(P_i) - F_C(P_j)\|_1 + 1}, \quad (1)$$

where $m_{i,j}$ is the number of elements of $I_{i,j}$ which is the set of common stimuli images of subjects N_i and N_j, also $FCATS = \{Gender, Fashion, Color, Sport, Other\}$. Using the above algorithm for constructing SSN, we get a weighted network in which the weight of the link between two subjects is an indicator of the similarity of their personal features and personalized saliency maps. For clustering the subjects, we chose the Louvain algorithm, since this algorithm works well with weighted network. The Louvain algorithm is a heuristic algorithm which aims at maximizing modularity in the process of detection of communities. One of the nice features of this algorithm is that it does not require the number of communities or the size of them, before execution. By applying the Louvain community detection algorithm to SSN, we obtain the clusters. As we will see in Section 4.1, we will experiment different values for $W_{Gender}, W_{Fashion}, W_{Color}, W_{Sport}$ and W_{Other} and choose the case which yields the clustering with the **highest modularity** in the Louvain algorithm.

3.2 Universal saliency prediction

For universal saliency prediction, we evaluated various state-of-the-art methods, such as DeepGaze IIE ([18]), ML-Net ([6]) and SalGAN ([20]). As we will see in Section 3.3, we use the USMs obtained by these methods as part of the inputs in training the MDST model.

3.3 Multi-Domain Saliency Translation

In order to obtain the saliency maps for each cluster, we propose the multi-domain saliency translation (MDST) model, based on Conditional Generative Adversarial Networks (cGAN). This model is adapted from the Pix2Pix model by Isola *et al.* [8]. Here, we have a cluster-mapping network inspired by the class network in [15] which takes the cluster label as the

input and generates a point in the class space. This network is comprised of an embedding layer and followed by 4 fully connected layers. The Equalized Learning Rate technique ([10]) is used for the fully connected layers. The output of the cluster-mapping network for each cluster label is concatenated to the input image and its USM along the channel dimension and fed to the generator which is identical to the U-Net generator of the Pix2Pix model. Also, to the discriminator, which is identical to the discriminator of Pix2Pix model, we feed the concatenation of the input image and its USM and the result generated by the generator, concatenated along the channel dimension. So, this model outputs the Clustered Saliency map for a viewer in cluster c and image x, given input image x, USM of x using a method of choice and cluster label (c) as inputs. The objective of our network, similarly to the one for Pix2Pix model, is as below:

$$G^\star = arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda \mathcal{L}_{L1(G)}, \tag{2}$$

where G is the generator, D is the discriminator and λ is the weight for L1 loss function. We replace the loss function for cGAN, which is a cross-entropy objective in Pix2Pix model, with a Mean Squared Error (MSE) loss. This modification improves the performance based on the results of some experiments. Here we use $\lambda = 100$ in Equation (2). We train the model for 200 epochs, using Adam optimizer with initial learning rate of 0.0002 for the first 100 epochs and linearly decaying learning rate to 0 in the remaining 100 epochs. For the Adam optimizer, the momentum parameters are $\beta_1 = 0.5$ and $\beta_2 = 0.999$ and we use weight decay of 0.00001, to help prevent overfitting.

Training approach for MDST model. Suppose that after clustering subjects using the SSC algorithm, we have n clusters $C_1,\dots,C_n$. Assume that in the PSM dataset we have personalized saliency maps for a set of stimuli images I. We denote the USM of an image x obtained using method M by $S^M_{USM}(x)$ and define $\{S^M_{USM}(x') \,|\, x' \in I\} = S^M_{USM}(I)$. For a cluster C_i and image $x \in I$, $S_{PSM}(C_i,x) = \frac{\sum_{P\in C_i} S_{PSM}(P,x)}{|C_i|}$, where $|C_i|$ is the number of subjects in C_i. MDST is trained by setting image x, $S^M_{USM}(x)$ and label of C_i as input and $S_{PSM}(C_i,x)$ as the target image, for all stimuli images $x \in I$ and all clusters $C_i \in \{C_1,\dots,C_n\}$.

3.4 Prediction of saliency maps for a new subject

Having a dataset D that contains personalized saliency data and personal information of some subjects, for a new subject A, we want to predict the A's saliency map for an image stimulus. First, using the SSC algorithm, we put all the subjects of the dataset D into clusters, $\{C_1,\dots,C_n\}$, and train MDST model, as explained in Section 3.3. Now we want to determine which cluster in dataset D subject A belongs to, given A's personal features and A's available saliency maps. Assume that from A, we have the vector $F(A) = [f^A_1, f^A_2, \cdots, f^A_m]$ of values for the set of personal features $\{f_1, f_2, \cdots, f_m\}$, and a set $\{S_{PSM}(A,x) \,|\, x \in I\} = S_{PSM}(A,I)$ for the set of images, I, such that all $x \in I$ exists in the stimuli images set of dataset D. We normalize all the saliency maps such that their pixel values are in [0,1] and resize them to the same size, with resolution R. For each subject P in dataset D, as the feature vector we only consider the values of the features in $\{f_1, \cdots, f_m\}$, and show this vector by $F(P) = [f^P_1, \cdots, f^P_m]$. For each cluster C_i, we denote $F(C_i)$ as the element-wise average of all $F(P)$ for $P \in C_i$. For feature categories, we consider the same feature weights used for SSC algorithm (e.g. W_{Gender}, $W_{Fashion}$, W_{Color}, W_{Sport} and W_{Other} for the PSM dataset) as in Section 3.1. We denote the subset of features of $F(P)$ and $F(C_i)$ which are in feature category C by $F_C(P)$ and $F_C(C_i)$.

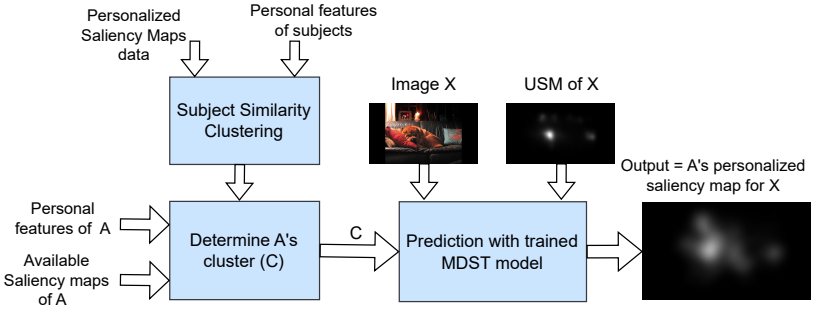

Figure 1: The pipeline of our Clustered Saliency Prediction model, which is combined from clustering of the subjects and the MDST model.

We define a closeness measure between subject A and each cluster C_i, called $SalClose(A,C_i)$, computed as below.

$$SalClose(A,C_i) = \frac{m_{A,C_i}}{\sum_{x \in I_{A,C_i}} \frac{\|S_{PSM}(A,x) - S_{PSM}(C_i,x)\|_1 + 1}{R}} + \sum_{C \in FCATS} \frac{W_C \times |C|}{\|F_C(C_i) - F_C(A)\|_1 + 1}, \quad (3)$$

where m_{A,C_i} is the number of elements of I_{A,C_i} which is the set of common stimuli images between A and cluster C_i and $FCATS$ is the set of feature categories of dataset D, e.g. $\{Gender, Fashion, Color, Sport, Other\}$ for PSM dataset. We assign A to the cluster C_{ch} such that $SalClose(A,C_{ch})$ is the maximum among all the clusters. To predict the saliency map of A for an image stimulus x_s, we input x_s, $S^M_{USM}(x_s)$ (M is a chosen USM prediction model) and label of cluster C_{ch} to MDST model. We consider the output as the Clustered Saliency Prediction for A. The overall pipeline of this process is shown in Figure 1.

4 Experimental results

4.1 Clustering

For each of the 6 random splits of train/validation/test sets with proportions of 64%, 16% and 20%, for feature weights W_{Gender}, $W_{Fashion}$, W_{Color}, W_{Sport} and W_{Other}, we examine each of the values in {1, 2, 4, 8} (1024 total cases) in SSC algorithm in Section 3.1 using train set images of the split and pick the case which yields the clustering with highest modularity. The average of the obtained feature weights over the 6 splits are $W_{Gender} = 3.5$, $W_{Fashion} = 8$, $W_{Color} = 1$, $W_{Sport} = 1$ and $W_{Other} = 1$. These settings for each split give 3 clusters which we use in the experiments in Section 4.2. For the case with $W_{Gender} = 0$, $W_{Fashion} = 0$, $W_{Color} = 0$, $W_{Sport} = 0$ and $W_{Other} = 0$, we get one cluster with all the subjects in it.

4.2 Multi-Domain Saliency Translation model

We run 3 different set of experiments, where in each set we use one of DeepGaze IIE, ML-Net and SalGAN, to obtain USMs, as a part of the input for MDST model. We use data

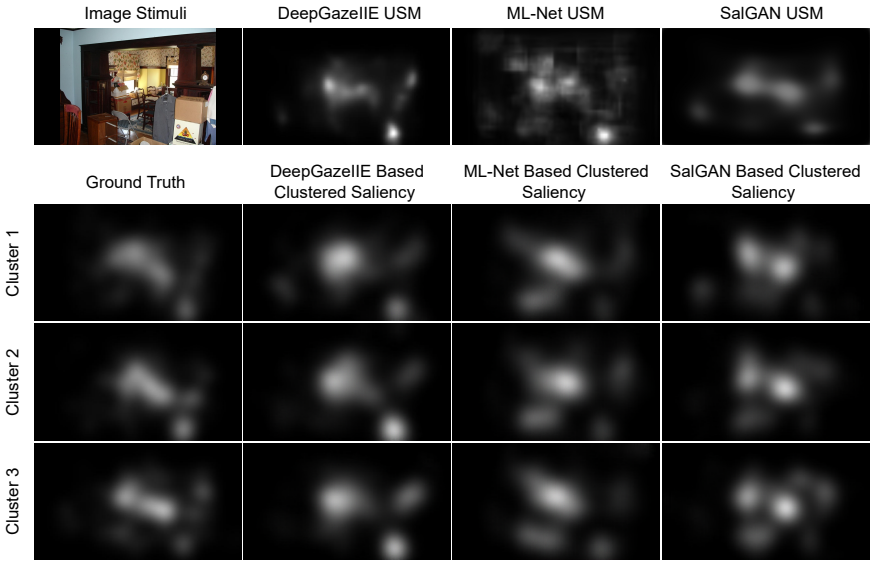

Figure 2: Clustered Saliency Prediction for the depicted stimulus image for obtained clusters from Section 4.1, using different USM prediction methods as the base for MDST model.

Prediction method	CC	SIM	AUC	NSS
DeepGaze IIE based Clustered	**0.7418**	**0.6369**	0.8862	2.2518
DeepGaze IIE	0.6768	0.5949	**0.8972**	**2.6413**
ML-Net based Clustered	0.7115	0.6145	0.8765	2.1360
ML-Net	0.6504	0.5701	0.8729	2.2585
SalGAN based Clustered	0.6938	0.6026	0.8735	2.0772
SalGAN	0.6606	0.5816	0.8757	2.0182

Table 1: Mean performance of our Clustered Saliency Prediction for all subjects in PSM dataset and comparison to 3 USM prediction methods. Higher score in each metric is better.

augmentation techniques, *i.e.* resize to 286×286 pixels and then random crop to 256×256 pixels and also random horizontal flip for the input images and USMs and PSMs. The cluster label is mapped to a code of size 256×16 by an embedding layer and the output is fed into 4 consecutive fully connected layers with the same input and output sizes, 256×16. The output of this is duplicated 4 times, concatenated to each other and resized to shape (256, 256). In all the experiments, we train MDST for 200 epochs (as also explained in Section 3.3) and batch size of 16, using the clustering obtained in Section 4.1. The evaluation results are found in Table 1, where the results for each metric and each experiment are the average of scores for all subjects in the PSM dataset on the test set over 6 random splits of images into train/validation/test sets with proportions of 64%, 16% and 20% (same splits as in Section 4.1). In Figure 2, we see the USMs of a random image using each of USM prediction models, and the clustered predictions based on these USMs.

As in Table 1, the results for CC and SIM metrics are higher in the Clustered Saliency

Clustering		CC	SIM	AUC Judd	NSS
SSC	Most populated cluster	**0.7573**	**0.6483**	**0.8933**	**2.3376**
	Average of all clusters	0.7418	0.6369	0.8862	2.2518
One cluster		0.7422	0.6368	0.8876	2.2519
3 Random clusters		0.7416	0.6368	0.8864	2.2487
30 clusters		0.7274	0.6295	0.8687	2.2736

Table 2: Performance of DeepGaze IIE based MDST network on different ways of clustering, averaged over 6 random splits. "SSC" in the 1st row stands for Subject Similarity Clustering (same results as in the 1st row of Table 1). Higher score for each metric is better.

Predictions, than the results of universal saliency predictions. Also for NSS metric, SalGAN based Clustered Saliency Prediction shows superiority to SalGAN. For AUC Judd metric, the performance of ML-Net based Clustered Saliency Prediction is better than ML-Net. All of these show the improvement of our model over universal saliency prediction models. The reason for higher performance in AUC Judd and NSS for some cases in the universal saliency model is that the MDST model is trained to convert input images, USMs and cluster labels to saliency maps of the clusters and it does not take into account the unprocessed fixation points data, while AUC Judd and NSS use fixation points as ground truths. Considering this, AUC Judd and NSS might not be good metrics for our evaluations. As we see in Figure 2, in the USMs obtained by DeepGaze IIE, the salient parts are more sharp than the salient parts in the Clustered Saliency Predictions for each cluster. Since NSS penalizes false positives at fixation locations, a more sharp saliency map which has higher peaks, might gain a higher NSS than a saliency map with more spread-out salient regions.

To prove the positive effect of the clustering by SSC algorithm on the saliency prediction, we evaluate the results of DeepGaze IIE based MDST network on three baseline cases: 1) having only one cluster comprised of all the subjects, 2) having 3 random clusters, where each person is assigned randomly to one of them, based on uniform distribution, 3) having 30 clusters, where each subject is in a separate cluster. Based on Table 2, where the same 6 random splits as in Table 1 were used, for DeepGaze IIE based Clustered Saliency using SSC clustering approach the performance is higher for all the metrics than baseline cases 2 and 3 and is higher in some metrics than baseline case 1. The most populated cluster of each split over 6 random splits, all have the same 10 members in common, plus a few different members for each split. As we see in Table 2, the average performance for these most populated clusters of all the splits is much improved compared to all other cases. Considering this observation, to get an even further improved performance, for individuals outside of the most-populated cluster we can also use the MDST model trained on the average PSMs of all the individuals (similar to baseline case 1) to predict saliency maps.

4.3 Saliency prediction for new subjects and comparison with other methods

Similar to [25], from the PSM dataset we randomly pick 20 subjects whose PSMs are used for clustering and training MDST. The other 10 subjects are used as new targets. In open-set evaluation, the model trained on the 20 subjects is tested on PSMs in the test set of the 10 other subjects. In closed-set evaluation, the trained model is tested on PSMs in the test set of the 20 subjects. Here we have used 5 random splits, where in each split proportions for

images in test, validation and train sets are 20%, 16% and 64% respectively. Using the SSC algorithm in Section 3.1, from all the combination of weights {1, 2, 4, 8} for each feature category, across all 5 splits the chosen feature weights for Gender, Color, Sport, Fashion and Other which yield the clustering with highest modularity are 8, 1, 1, 8, 1, respectively. In each split this gives 2 clusters. Then, we train MDST and evaluate the results (Table 3). For the 10 test subjects for open-set evaluation, over all 5 splits, we average the scores when assigning each new person to their chosen cluster by our algorithm in Section 3.4 and compare to the average of all the scores when assigning each new person to each of the non-chosen clusters. As in Table 3, the average of scores for assigning the new subjects to chosen clusters for Clustered Saliency Prediction is higher than the average of scores for assigning the new subjects to non-chosen clusters. This proves the effectiveness of our algorithm in assigning the new subjects to the most appropriate cluster for saliency prediction.

In [25], they predict personalized saliency maps from USMs using multi-task CNN architecture and CNN with Person-specific Information Encoded Filters (CNN-PIEF). They evaluate their approach under closed-set and open-set settings, by randomly choosing 20 subjects to train their models and test on the remaining subjects. Since we do not know the exact settings of their experiments, comparison with their results under the same circumstances is not possible. However if we overlook this, we have higher performance in CC, AUC Judd and NSS for both DeepGaze IIE based and ML-Net based Clustered Saliency Prediction (Table 3). In [19], they also chose 10 random viewers from the PSM dataset as new targets and used the other 20 subjects for training. Once again, since the evaluation settings of this paper is not the same as ours, we cannot accurately compare their results to ours. Their performance appears higher than ours (Table 3), but they are solving a different problem. Keep in mind that our method predicts the average saliency for each cluster, not each person. It is important to keep in mind that our clustered salience approach has other advantages compared to the competing methods mentioned in this section. Most significantly, our method associates individuals with clusters of other viewers. This association can be used to infer additional information for the individual based on any information known for the cluster. For example, it may be known what the favorite music is for members of the cluster, which could be applied to provide recommendations to the new individual. We do not investigate leveraging the cluster associations in this paper, but leave it for future work.

5 Discussion and Conclusion

From our experiments, we conclude that our Clustered Saliency Prediction method improves on standard universal saliency prediction methods. Based on Table 1, the performance of DeepGaze IIE based Clustered Saliency Prediction is higher than ML-Net and SalGAN based Clustered Saliency Predictions. This is logical, since DeepGaze IIE has a better performance comparing to two other universal saliency models. One of the advantages of our Clustered Saliency Prediction model is that the USMs that we use as the base can be obtained with any universal saliency model. So as universal saliency models improve, we can upgrade our model by using a better performing universal saliency model base.

Also, as we discussed in Section 4.2, our SSC approach divides the subjects into clusters, where each cluster contains subjects with more similar saliency patterns. We also see that some features categories such as fashion, appear to correlate more with the saliency of subjects. Moreover, the clusters found in our approach, can be used in future works for further applications, such as recommendation systems, advertising, etc.

Methods		CC	SIM	AUC Judd	NSS
[25], closed-set	ML-Net based CNN-PIEF	0.6368	0.8095	0.8365	1.5105
	ML-Net based Multi-task CNN	0.6463	0.8077	0.8414	1.4960
[25], open-set	ML-Net based CNN-PIEF	0.6450	0.8166	0.8559	1.6879
	ML-Net based Multi-task CNN	0.6117	0.7946	0.8534	1.5490
[19]	Few-shot PSM pred.	0.7845	0.6557	-	-
Our method, closed-set	ML-Net based Clustered	0.7107	0.6167	0.8725	2.1057
	DeepGaze IIE based Clustered	0.7417	0.6398	0.8819	2.2181
Our method, open-set	ML-Net based Clustered	0.7030	0.5981	0.8852	2.2019
	ML-Net based Non-Chosen Clustered	0.6976	0.5954	0.8842	2.1876
	DeepGaze IIE based Clustered	0.7336	0.6216	0.8945	2.3157
	DeepGaze IIE based Non-Chosen Clustered	0.7274	0.6184	0.8936	2.3004

Table 3: Comparison of our methods under closed-set and open-set evaluation settings with other approaches. The rows "ML-Net/DeepGaze IIE based Non-Chosen Clustered" have the average for ML-Net/DeepGaze IIE based Clustered Saliency Predictions using MDST, when assigning the new subjects to the clusters **not chosen** by our algorithm in Section 3.4.

6 Acknowledgement

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) and funding from the Ministère de l'Économie, de l'Innovation et de l'Énergie of Québec.

References

[1] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[2] Ali Borji. Saliency prediction in the deep learning era: Successes and limitations. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[3] 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.

[4] Moran Cerf, E Paxon Frady, and Christof Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of vision*, 9(12):10–10, 2009.

[5] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A deep multi-level network for saliency prediction. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3488–3493. IEEE, 2016.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[7] Tomoki Ishikawa and Takahiro Yakoh. Saliency prediction based on object recognition and gaze analysis. *Electronics and Communications in Japan*, 104(2):e12303, 2021.

[8] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[9] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[10] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hk99zCeAb.

[11] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987.

[12] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze I: boosting saliency prediction with feature maps trained on imagenet. *CoRR*, abs/1411.1045, 2014.

[13] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. Deepgaze II: reading fixations from deep features trained on object recognition. *ArXiv*, abs/1610.01563, 2016.

[14] Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: A comparative analysis. *Phys. Rev. E*, 80:056117, Nov 2009. doi: 10.1103/PhysRevE.80.056117. URL https://link.aps.org/doi/10.1103/PhysRevE.80.056117.

[15] Héctor Laria, Yaxing Wang, Joost van de Weijer, and Bogdan Raducanu. Transferring unconditional to conditional gans with hyper-modulation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3839–3848, 2022. doi: 10.1109/CVPRW56347.2022.00429.

[16] Olivier Le Meur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods*, 45(1):251–266, 2013.

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

[18] Akis Linardos, Matthias Kümmerer, Ori Press, and Matthias Bethge. Deepgaze IIE: calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 12899–12908. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01268. URL https://doi.org/10.1109/ICCV48922.2021.01268.

[19] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Few-shot personalized saliency prediction based on adaptive image selection considering object and visual attention. *Sensors*, 20(8):2170, 2020.

[20] Junting Pan, Cristian Canton-Ferrer, Kevin McGuinness, Noel E. O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giró-i-Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *CoRR*, abs/1701.01081, 2017. URL http://arxiv.org/abs/1701.01081.

[21] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[22] Eleonora Vig, Michael Dorr, and David Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2798–2805, 2014.

[23] David W.-L. Wu, Walter F. Bischof, Nicola C. Anderson, Tanya Jakobsen, and Alan Kingstone. The influence of personality on social attention. *Personality and Individual Differences*, 60:25–29, 2014. ISSN 0191-8869. doi: https://doi.org/10.1016/j.paid.2013.11.017. URL https://www.sciencedirect.com/science/article/pii/S0191886913013755.

[24] Yanyu Xu, Nianyi Li, Junru Wu, Jingyi Yu, and Shenghua Gao. Beyond universal saliency: Personalized saliency prediction with multi-task cnn. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3887–3893, 2017. doi: 10.24963/ijcai.2017/543. URL https://doi.org/10.24963/ijcai.2017/543.

[25] 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.