sensors

MDPI

Article

Few-Shot Personalized Saliency Prediction Based on Adaptive Image Selection Considering Object and Visual Attention[†]

Yuya Moroto [1,*], Keisuke Maeda [2,*], Takahiro Ogawa [3] and Miki Haseyama [3]

[1] Graduate School of Information Science and Technology, Hokkaido University, N-14, W-9, Kita-ku, Sapporo, Hokkaido 060-0814, Japan

[2] Office of Institutional Research, Hokkaido University, N-8, W-5, Kita-ku, Sapporo, Hokkaido 060-0808, Japan

[3] Faculty of Information Science and Technology, Hokkaido University, N-14, W-9, Kita-ku, Sapporo, Hokkaido 060-0814, Japan; ogawa@lmd.ist.hokudai.ac.jp (T.O.); miki@ist.hokudai.ac.jp (M.H.)

[*] Correspondence: moroto@lmd.ist.hokudai.ac.jp (Y.M.); maeda@lmd.ist.hokudai.ac.jp (K.M.)

[†] This paper is an extended version of our paper published in: Moroto, Y.; Maeda, K.; Ogawa, T.; Haseyama, M. User-Specific Visual Attention Estimation Based on Visual Similarity and Spatial Information in Images. In the Proceedings of the IEEE International Conference on Consumer Electronics—Taiwan (IEEE 2019 ICCE-TW), Ilan, Taiwan, 20–22 May 2019.

Received: 15 March 2020; Accepted: 9 April 2020; Published: 11 April 2020

Abstract: A few-shot personalized saliency prediction based on adaptive image selection considering object and visual attention is presented in this paper. Since general methods predicting personalized saliency maps (PSMs) need a large number of training images, the establishment of a theory using a small number of training images is needed. To tackle this problem, although finding persons who have visual attention similar to that of a target person is effective, all persons have to commonly gaze at many images. Thus, it becomes difficult and unrealistic when considering their burden. On the other hand, this paper introduces a novel adaptive image selection (AIS) scheme that focuses on the relationship between human visual attention and objects in images. AIS focuses on both a diversity of objects in images and a variance of PSMs for the objects. Specifically, AIS selects images so that selected images have various kinds of objects to maintain their diversity. Moreover, AIS guarantees the high variance of PSMs for persons since it represents the regions that many persons commonly gaze at or do not gaze at. The proposed method enables selecting similar users from a small number of images by selecting images that have high diversities and variances. This is the technical contribution of this paper. Experimental results show the effectiveness of our personalized saliency prediction including the new image selection scheme.

Keywords: personalized saliency map; adaptive image selection; multi-task CNN; object detection

1. Introduction

Many researchers have attempted to predict a saliency map that indicates image components that are more attractable than their neighbors [1–4]. Since a saliency map reflects human visual attention, it has been expected to contribute to image processing tasks including image re-targeting [5,6], image compression [7,8], and image enhancement [9,10]. The purpose of those studies is prediction of instinctual human visual attention, that is, the common regions of images to humans. Such a saliency map is called a Universal Saliency Map (USM). However, visual attention can differ between persons if individual backgrounds are taken into account [11–13]. In fact, since it has been reported that each person views regions of images reflecting personalized interests [14–16], the prediction of person-specific visual attention, which is called a Personalized Saliency Map (PSM), has been needed [17,18].

In order to accurately predict PSMs, Xu et al. constructed a PSM dataset and proposed a PSM prediction method [14,19]. Their PSM dataset includes a large number of images and their corresponding gaze data obtained from many persons. To the best of our knowledge, their PSM dataset is the first dataset focusing on PSM prediction. Xu's method, which is based on a multi-task Convolutional Neural Network (multi-task CNN) [20], needs a large amount of training data for PSM prediction. Thus, for predicting a PSM by this method for a new person not included in the PSM dataset, a large amount of gaze data, which involve a heavy burden, must be obtained for retraining the multi-task CNN. Furthermore, in a real-world situation, PSM prediction of new images not included in the PSM dataset is necessary. Thus, a PSM prediction method without such large-scale data acquisition is desirable. Our previous study revealed that the use of gaze data obtained from similar persons, who view regions in images similar to those viewed by the target person, is effective for PSM prediction [21]. Since the previous study assumes that similar persons have already gazed at the new image, the actual gaze data of similar persons can be utilized. However, since the new image is not always gazed by similar persons, a method that can estimate a PSM of the target person based on predicted PSMs calculated from similar persons is a much more practicable approach. From the above discussions, construction of such a method is a challenging but indispensable task to be addressed.

For predicting a PSM for the new target person, he/she needs to view several images to search for similar persons. Before this procedure, we need to select images from the PSM dataset for calculating person similarities between the target person and those included in the PSM dataset. However, if the selected images are visually similar to each other, the calculated person similarities are not reliable. In order to realize robust PSM prediction with reduction in the number of selected images, an adaptive image selection scheme solving the above problem is necessary. Specifically, we focus on the following two aspects: 1) diversity of images and 2) variance of PSMs. Since the PSM dataset consists of images that have high diversity, we should also select images with maintenance of their diversity. Moreover, the variance of PSMs for persons included in the PSM dataset should be high since the regions in images that many persons commonly gaze at or do not gaze at can be represented by a USM. Thus, by introducing an adaptive image selection scheme focusing on the above two aspects, it is expected that PSM prediction for the new target person is realized with high accuracy.

This paper presents a few-shot PSM prediction (FPSP) method using a small amount of training data based on adaptive image selection (AIS) considering object and visual attention. Figure 1 shows an illustration of the problem we try to tackle. First, we construct and train a multi-task CNN from the PSM dataset for predicting PSMs of persons included in the PSM dataset [20]. Next, the person similarity is calculated by using selected images included in the PSM dataset. These images are chosen by AIS focusing on the diversity of images and the variance of PSMs. For guaranteeing the high diversity of the selected images, AIS focuses on the kinds of objects included in the training images in the PSM dataset by using a deep learning-based object detection method. Then, objects that have high variances of PSMs are detected, and then we can adaptively select images including such objects, which are shown in the orange area in Figure 1. Finally, FPSP of a target image for the new target person is realized on the basis of the person similarity and PSMs predicted by the multi-task CNN trained for the persons in the PSM dataset. Consequently, FPSP based on AIS for the new target person can be realized from a small amount of training data with high accuracy.

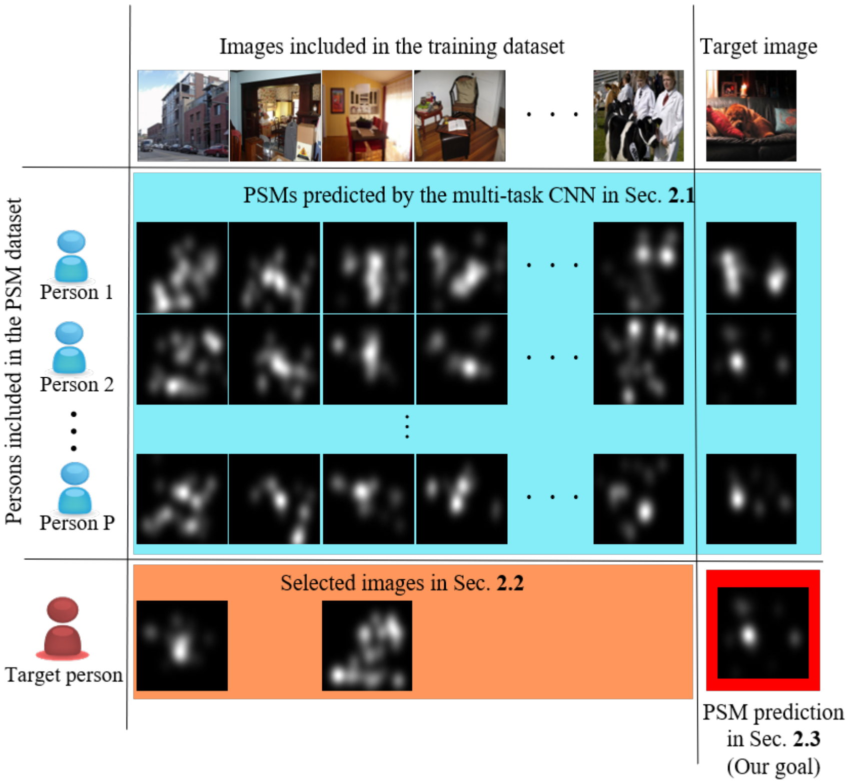
Images included in the training dataset
Target image
PSMs predicted by the multi-task CNN in Sec. **2.1**
Persons included in the PSM dataset
Person 1
Person 2
Person P
Selected images in Sec. **2.2**
Target person
PSM prediction in Sec. **2.3** (Our goal)

Figure 1. The problem setting of our study. The purpose of our study is Personalized Saliency Map (PSM) prediction of a target person for images not included in the training dataset. For predicting a PSM for a target image, the target person needs to view only some images, which have been viewed by persons included in the training PSM dataset.

It should be noted that this paper is an extended version of [22]. Specifically, we enable the novel PSM prediction of the target person from those predicted from similar persons based on the multi-task CNN. Furthermore, we newly introduce the AIS into the above PSM prediction approach.

2. Few-shot PSM Prediction Based on Adaptive Image Selection

In this section, we explain variables used in this section shown in Table 1 and our proposed method shown in Figure 2. In our method, the multi-task CNN is trained from the PSM dataset for saliency prediction of persons included in the PSM dataset (See Section 2.1). Then, we chose images based on AIS from the PSM dataset (See Section 2.2), and the target person needs to view only the selected images for his/her PSM prediction. Finally, we predict the target person's saliency map for the new target image by using the predicted saliency maps of similar persons in the PSM dataset (See Section 2.3).

Table 1. The list of variables used in Section 2.

Section 2.1	
$\boldsymbol{X}_n$	nth image in the training data
$\boldsymbol{X}^{\text{tgt}}$	Target image
$\boldsymbol{S}^{\text{USM}}(\boldsymbol{X}_n)$	Universal Saliency Map (USM) of image $\boldsymbol{X}_n$
$\boldsymbol{S}^{\text{PSM}}(p, \boldsymbol{X}_n)$	PSM of image $\boldsymbol{X}_n$ for person p
$\boldsymbol{S}^{\text{out}}(p, \boldsymbol{X}^{\text{tgt}})$	PSM predicrted by multi-task Convolutional Neural Network (CNN) for image $\boldsymbol{X}^{\text{tgt}}$ and person p
P	Number of person
N	Number of images
$\boldsymbol{\Delta}(p, \boldsymbol{X}_n)$	Difference map between USM $\boldsymbol{S}^{\text{USM}}(\boldsymbol{X}_n)$ and PSM $\boldsymbol{S}^{\text{PSM}}(p, \boldsymbol{X}_n)$
$\hat{\boldsymbol{\Delta}}_l(p, \boldsymbol{X}_n, \boldsymbol{S}^{\text{USM}}(\boldsymbol{X}_n))$	Difference map calculated for image $\boldsymbol{X}_n$ and person p
n	Index of images
p	Index of persons
l	Index of decoding layers
d_1	Height of image
d_2	Width of image
d_3	Number of color channels
Section 2.2	
$\boldsymbol{O}_{(n,m)}$	mth object including nth image
$\boldsymbol{S}^{\text{PSM}}(p, \boldsymbol{O}_{(n,m)})$	PSM of object $\boldsymbol{O}_{(n,m)}$
$\bar{\boldsymbol{S}}^{\text{PSM}}(\boldsymbol{O}_{(n,m)})$	Average PSM of object $\boldsymbol{O}_{(n,m)}$
$d^w_{(n,m)}$	Width of mth object included in nth image
$d^h_{(n,m)}$	Height of mth object included in nth image
$v_{(n,m)}$	Variance of mth object including nth image
$\bar{v}_n$	Average of $v_{(n,m)}$
M	Kinds of objects in all images included in PSM dataset
C	Number of selected images
m	Index of objects
j	Index of width of pixel location
k	Index of height of pixel location
c	Index of selected images
Section 2.3	
β^p	Similarity score between a target person and person p
p^{new}	Target person
τ	Threshold value for person similarity
a^p	Selection coefficient for person similarity
w^p	Person similarity between a target person and person p
$\boldsymbol{S}^{\text{FPSP}}(p^{\text{new}}, \boldsymbol{X}^{\text{tgt}})$	PSM predicted by Few-shot Personalized Saliency Prediction (FPSP)

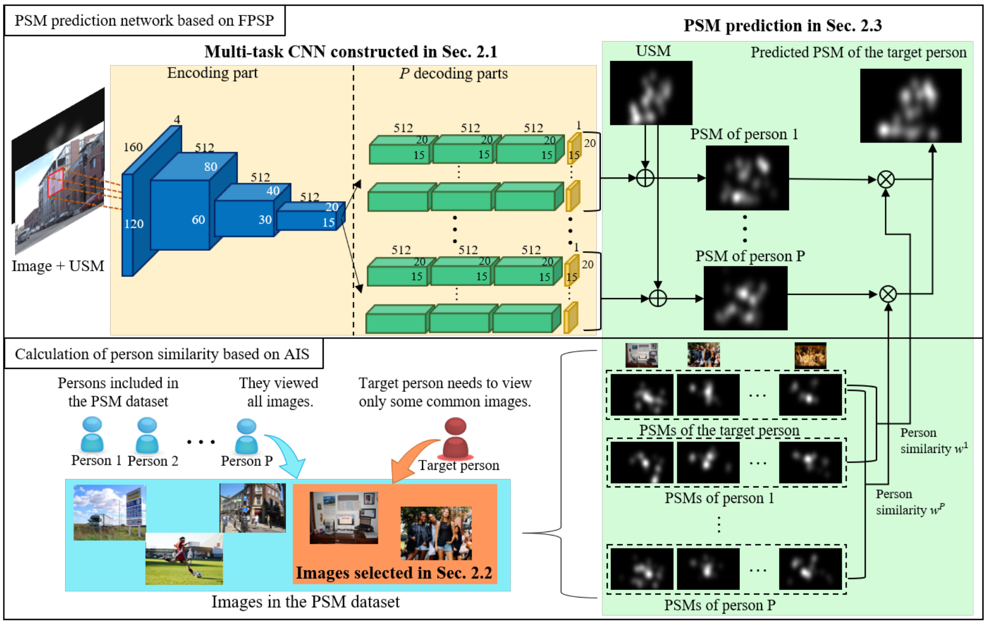
PSM prediction network based on FPSP
PSM prediction in Sec. 2.3
Multi-task CNN constructed in Sec. 2.1
Encoding part
P decoding parts
Image + USM
160
120
4
512
80
60
512
40
30
512
20
15
1
20
15
USM
Predicted PSM of the target person
PSM of person 1
PSM of person P
Calculation of person similarity based on AIS
Persons included in the PSM dataset
They viewed all images.
Target person needs to view only some common images.
Person 1
Person 2
Person P
Target person
Images selected in Sec. 2.2
Images in the PSM dataset
PSMs of the target person
PSMs of person 1
PSMs of person P
Person similarity w^1
Person similarity w^P

Figure 2. Overview of FPSP and Adaptive Image Selection (AIS). The upper row shows the pipeline of FPSP and the lower row shows the method for the calculation of person similarities based on AIS.

The rest of this paper is organized as follows. In Section 2, FPSP including the AIS scheme is explained in detail. In Section 3, the effectiveness of our proposed method is shown from experimental results. Finally, in Section 4, we conclude this paper.

2.1. Construction of a Multi-Task CNN for PSM Prediction

In this subsection, we explain the construction of a multi-task CNN for PSM prediction. This multi-task CNN is constructed for calculating P PSMs, where P is the number of persons, who are those included in the PSM dataset [20]. In the proposed method, the input data including images $\boldsymbol{X}_n \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ ($n = 1, 2, \ldots, N$; N being the number of training images, $d_1 \times d_2$ being the number of pixels, and d_3 being the number of color channels) and USMs $\boldsymbol{S}^{\mathrm{USM}}(\boldsymbol{X}_n) \in \mathbb{R}^{d_1 \times d_2}$ are used for training the multi-task CNN. The USM means the area which many persons gaze at. In our method, the USM $\boldsymbol{S}^{\mathrm{USM}}(\boldsymbol{X}_n)$ can be obtained by an arbitrary method, and it is not our contribution. Thus, it is shown in Section 3. Given PSMs $\boldsymbol{S}^{\mathrm{PSM}}(p, \boldsymbol{X}_n) \in \mathbb{R}^{d_1 \times d_2}$ for P persons, where $\boldsymbol{S}^{\mathrm{PSM}}(p, \boldsymbol{X}_n)$ is obtained from pth person's gaze data for image $\boldsymbol{X}_n$ and included in the PSM dataset [20], we calculate a difference map $\boldsymbol{\Delta}(p, \boldsymbol{X}_n)$ between the USM and the PSM of each person as $\boldsymbol{\Delta}(p, \boldsymbol{X}_n) = \boldsymbol{S}^{\mathrm{PSM}}(p, \boldsymbol{X}_n) - \boldsymbol{S}^{\mathrm{USM}}(\boldsymbol{X}_n)$ by following [14]. The multi-task CNN has one encoding part and P decoding parts consisting of three layers. Its output layer provides P results of $\boldsymbol{\Delta}(p, \boldsymbol{X}_n)$ $(p = 1, 2, ..., P)$. The detail of multi-task CNN is shown in Figure 3. Moreover, we train the multi-task CNN by minimizing the following loss function:

$$\sum_{l=1}^{3} \sum_{p=1}^{P} \sum_{n=1}^{N} ||\hat{\boldsymbol{\Delta}}_l(p, \boldsymbol{X}_n, \boldsymbol{S}^{\mathrm{USM}}(\boldsymbol{X}_n)) - \boldsymbol{\Delta}(p, \boldsymbol{X}_n)||_F^2, \tag{1}$$

where $\hat{\boldsymbol{\Delta}}_l(\cdot)$ is a difference map calculation function that applies a 1×1 convolution layer to the outputs obtained from lth decoding layer, and $|| \cdot ||_F^2$ means the operator of the two-order Frobenius norm. Given a new target image $\boldsymbol{X}^{\mathrm{tgt}}$, by using the above trained network, we predict the PSM of person p as follows:

$$\boldsymbol{S}^{\mathrm{out}}(p, \boldsymbol{X}^{\mathrm{tgt}}) = \hat{\boldsymbol{\Delta}}_3(p, \boldsymbol{X}^{\mathrm{tgt}}, \boldsymbol{S}^{\mathrm{USM}}(\boldsymbol{X}^{\mathrm{tgt}})) + \boldsymbol{S}^{\mathrm{USM}}(\boldsymbol{X}^{\mathrm{tgt}}). \tag{2}$$

Therefore, PSMs of multiple persons can be predicted by the single model based on the multi-task CNN.

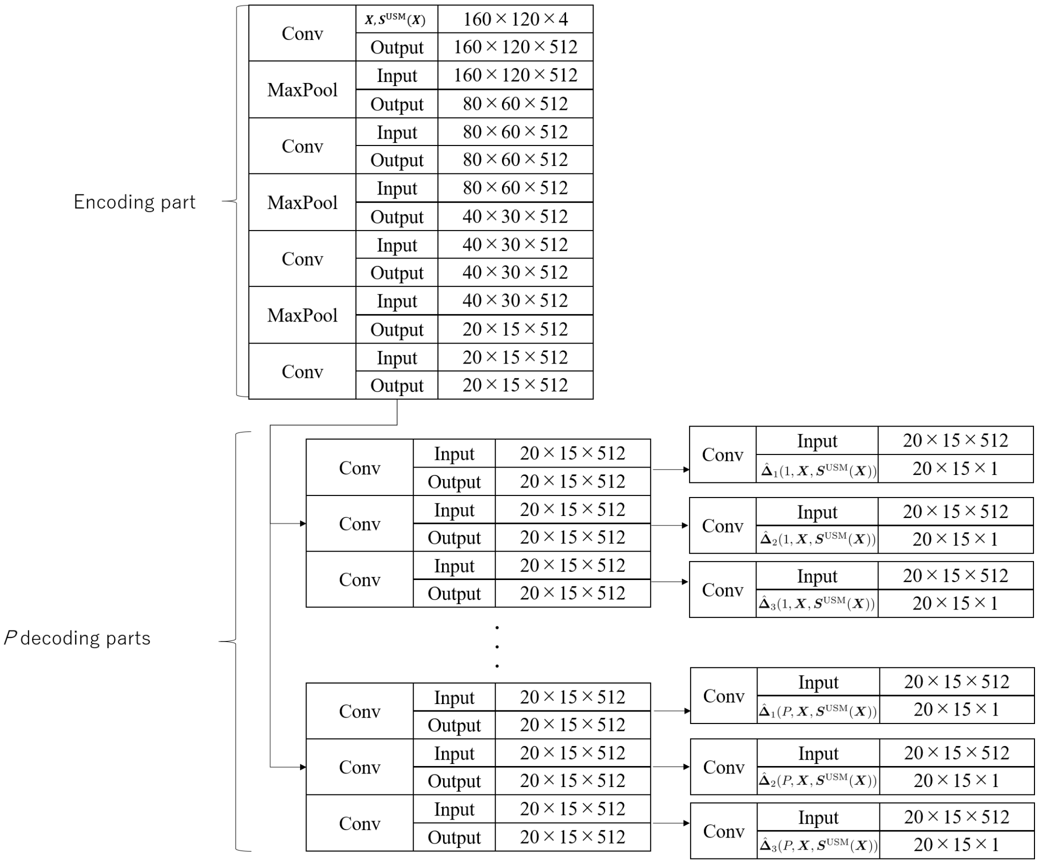

Figure 3. The details of the multi-task CNN used in our method. In this figure, "Conv" and "MaxPool" mean applying a convolution and maxpooling layer to each input data, respectively.

2.2. Adaptive Image Selection for Reduction of Viewed Images

In this subsection, we explain the AIS scheme for the reduction of images that the target person views for predicting his/her PSMs. Given the new person p^{new} not included in the PSM dataset, the multi-task CNN cannot learn the new person's PSM since the target person does not gaze at all of the images in the PSM dataset. Therefore, from the target person, we obtain some seed PSMs for images in the PSM dataset [20]. Note that the number of images viewed by the target person should be small for reducing his/her burden. Thus, the influence of one image on training is large, and the diversity of images significantly depends on its selection scheme. Although the PSM dataset has diversity of images, selected images do not necessarily have high diversity. We therefore propose a novel image selection method for maintaining the diversity of images with consideration of the kinds of objects in images and the variance of PSMs as shown in Figure 4. For maximizing the kinds of objects included in the selected images, we apply YOLO-v3 [23], which is one of the novel object detection methods, to the images included in the PSM dataset. Moreover, we use gaze data of persons included in the PSM dataset in order to consider the variance of PSMs. In AIS, we select images based on the detected objects and their PSMs. Specifically, we select objects that have high variances of PSMs since objects that have low variances are expected to be represented by a USM. Finally, we select images that include many kinds of objects having a high variance of PSMs. First, we detect objects $\boldsymbol{O}_{(n,m)}$ ($m = 1, \ldots, M$; M being the kinds of objects in all images) in images by using YOLO-v3 [23]. Detected objects are represented by the bounding box in which sizes are $d^h_{(n,m)} \times d^w_{(n,m)}$. Then, we calculate the object variance $v_{(n,m)}$ as follows:

$$v_{(n,m)} = \frac{1}{d^h_{(n,m)} \times d^w_{(n,m)}} \sum_{j=1}^{d^h_{(n,m)}} \sum_{k=1}^{d^w_{(n,m)}} \frac{1}{P} \sum_{p=1}^{P} \left\{ \boldsymbol{S}^{\mathrm{PSM}}(p, \boldsymbol{O}_{(n,m)})_{(j,k)} - \bar{\boldsymbol{S}}^{\mathrm{PSM}}(\boldsymbol{O}_{(n,m)})_{(j,k)} \right\}^2, \tag{3}$$

$$\bar{\boldsymbol{S}}^{\mathrm{PSM}}(\boldsymbol{O}_{(n,m)})_{(j,k)} = \frac{1}{P} \sum_{p=1}^{P} \boldsymbol{S}^{\mathrm{PSM}}(p, \boldsymbol{O}_{(n,m)})_{(j,k)}, \tag{4}$$

where $\boldsymbol{S}^{\mathrm{PSM}}(p, \boldsymbol{O}_{(n,m)})$ is the PSM of person p for the object $\boldsymbol{O}_{(n,m)}$, and (j,k) means the pixel location. Note that we treat $v_{(n,m)} = 0$ if image $\boldsymbol{X}_n$ does not include mth object, and we adopt the largest $v_{(n,m)}$ if image $\boldsymbol{X}_n$ includes mth objects. For selecting images, we calculate the sum of variances, $\bar{v}_n$, of PSMs for each image as follows:

$$\bar{v}_n = \sum_{m=1}^{M} v_{(n,m)}. \tag{5}$$

Finally, we select C images that have the highest values in Equation (5) from the PSM dataset. It is known that human visual attention depends on objects, and our method that explicitly uses this relationship is a simple but useful for maintaining the diversity and the variance. In particular, AIS adopts YOLO-v3 that has achieved remarkable high performance in the field of recent object recognition and enables for calculating the variance of visual attention pixel-wise. Thus, AIS that focuses on the combination use of the object detection and the visual attention is effective.

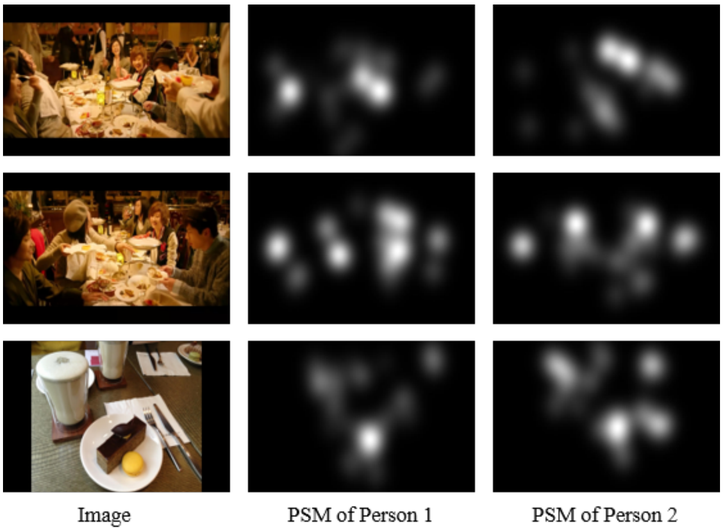

Figure 4. Examples for explaining the diversity of images and the variance of PSMs. Since images in the first and second rows are visually similar, AIS selects either one. On the other hand, for the image in the third row, since PSMs of person 1 and person 2 are similar, AIS does not select this image.

2.3. FPSP Based on Person Similarity

In this subsection, we explain FPSP using person similarity and the predicted PSM by the multi-task CNN. We predict the PSM of the new person p^{new} based on the similar persons' PSMs predicted by the multi-task CNN. First, we predict $\boldsymbol{S}^{\mathrm{out}}(p, \boldsymbol{X}^{\mathrm{sel}}_c)$ by inputting the target image into the multi-task CNN and using Equation (2), where $\boldsymbol{X}^{\mathrm{sel}}_c$ $(c = 1, 2, \ldots, C)$ are the C images selected in Section 2.2. Next, from the predicted PSMs $\boldsymbol{S}^{\mathrm{out}}(p, \boldsymbol{X}^{\mathrm{sel}}_c)$, we calculate cross correlation as a similarity score β^p between the target person p^{new} and person p included in the PSM dataset as follows:

$$\beta^p = \frac{1}{C}\sum_{c=1}^{C} \mathrm{corr}\left(\boldsymbol{S}^{\mathrm{PSM}}(p^{\mathrm{new}}, \boldsymbol{X}_c^{\mathrm{sel}}), \boldsymbol{S}^{\mathrm{out}}(p, \boldsymbol{X}_c^{\mathrm{sel}})\right), \tag{6}$$

where $\mathrm{corr}(\cdot,\cdot)$ calculates the cross correlation. Note that $\boldsymbol{S}^{\mathrm{PSM}}(p^{\mathrm{new}}, \boldsymbol{X}_c^{\mathrm{sel}})$ is obtained by using the gaze data for the target person p^{new}. This means that the new person p^{new} needs to view only the selected C images to obtain the gaze data for calculating the PSM $\boldsymbol{S}^{\mathrm{PSM}}(p^{\mathrm{new}}, \boldsymbol{X}_c^{\mathrm{sel}})$. Then, for eliminating the influence from dissimilar persons, we only select similar persons based on the selection coefficient a^p as follows:

$$a^p = \begin{cases} 1 & (\beta^p > \tau) \\ 0 & (\mathrm{otherwise}), \end{cases} \tag{7}$$

where τ is a pre-determined threshold value. Finally, by using the similarity score and the selection coefficient, we calculate the person similarity between the new person p^{new} and person p as follows:

$$w^p = \frac{a^p \beta^p}{\sum_{p'} a^{p'} \beta^{p'}}. \tag{8}$$

By using the person similarities w^p and similar persons' PSMs predicted by the multi-task CNN, we can simply predict the PSM $\boldsymbol{S}^{\mathrm{FPSP}}(p^{\mathrm{new}}, \boldsymbol{X}^{\mathrm{tgt}})$ of the new person p^{new} for the target image $\boldsymbol{X}^{\mathrm{tgt}}$ as follows:

$$\boldsymbol{S}^{\mathrm{FPSP}}(p^{\mathrm{new}}, \boldsymbol{X}^{\mathrm{tgt}}) = \sum_{p=1}^{P} w^p \boldsymbol{S}^{\mathrm{out}}(p, \boldsymbol{X}^{\mathrm{tgt}}). \tag{9}$$

Therefore, by using the person similarity w^p, the proposed method enables the prediction of the PSM of the new person from a small amount of training gaze data.

3. Experiment

In this section, the effectiveness of the FPSP based on AIS is shown from results of experiments. Section 3.1 shows the experimental settings, and Section 3.2 shows the performance evaluation and discussion.

3.1. Experimental Settings

In this subsection, we explain our experimental settings. We used the PSM dataset [20] that consisted of 1600 images and their corresponding gaze data for 30 persons who have normal or corrected-to-normal vision. The gaze data were obtained when each person gazed at each image for three seconds under free-viewing conditions. Moreover, we calculated PSMs based on gaze data by following [24]. We randomly chose 500 images as test images and used the remaining 1100 images as training images. Moreover, we selected C images from the training images and changed the number of the selected images, C, in $\{10, 20, \ldots, 100\}$. In this experiment, we randomly chose 10 persons as the new target persons in Section 2 and used the remaining 20 persons as those used for the training. We used the PSM calculated on the basis of gaze data as Ground Truth (GT). Moreover, the multi-task CNN was optimized on the basis of stochastic gradient descent [25], and then we set mini-batch size, learning rate, momentum and the number of iterations as 9, 0.00003, 0.9 and 1000, respectively. We experimentally set the threshold value τ to 0.7, where its determination will be investigated in future work. In the proposed method, $\boldsymbol{S}^{\mathrm{USM}}(\boldsymbol{X}_n)$ can be calculated as an average of the visual attention of training 20 persons.

For confirming the effectiveness of FPSP including the image selection scheme, we performed qualitative evaluation and quantitative evaluation. In the quantitative evaluation, we used the difference between the predicted PSM and its GT based on Pearson's correlation coefficient (CC), Kullback–Leibler divergence (KLdiv), and histogram intersection (Sim) [26] by following [27]. We also performed two kinds of comparative experiments. In the first comparative experiment, for revealing the effectiveness of our PSM prediction method based on a small amount of gaze data, we compared our method with the following four comparative methods that predict the USM chosen from the MIT saliency benchmark [28]:

- A USM prediction method based on low level visual features (Itti) [1]
- A USM prediction method based on a graph approach (GBVS) [2]
- A USM prediction method based on the separation of foreground and background in images (signature) [3]
- One of the state-of-the-art USM prediction methods based on deep learning (SalGAN) [4] which was trained from the SALICON dataset [29].

Moreover, we compared our method with the following two PSM prediction methods using a small amount of gaze data:

- A PSM prediction method based on visual similarities (Baseline1) [30]
- A PSM prediction method based on visual similarities and spatial information (Baseline2) [22].

It should be noted that the above comparative methods were trained by using the selected images since we assume that the target person views only selected images. In the second comparative experiment, for revealing the effectiveness of our image selection method, we compared our method with the following image selection methods:

- Image selection based on visual features (ISVF)
  Images having with a low similarity to those of visual features to other images were selected. We adopted the outputs of the final convolution layer of pre-trained DenseNet201 [31] as visual features.
- Image selection focusing on variance of PSMs (ISPSM)
  Images having a high variance of PSMs included in the PSM dataset were selected.

3.2. Performance Evaluation and Discussion

In this subsection, we confirm and discuss the experimental results. Figures 5–12 and Table 2 shows experimental results. First, Figure 5 shows the predicted results of one person and reveals that the FPSP method enables predicting the PSM that is the most similar to GT among all of the PSMs predicted by comparative methods. In Table 2, we show the average results, and it can be confirmed that FPSP based on AIS is the most effective for the PSM prediction in any evaluation indices. Therefore, by comparing the averages, we confirm the effectiveness of FPSP.

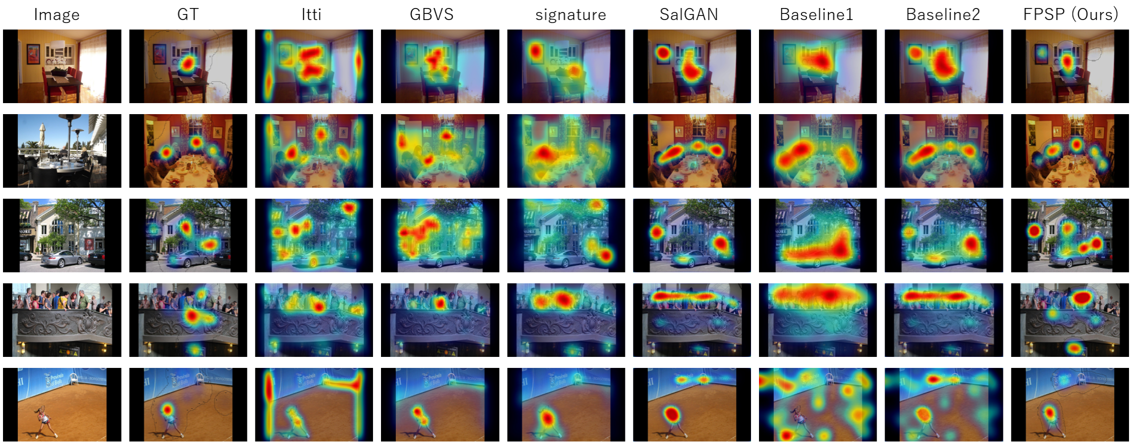

Figure 5. Qualitative results for one person predicted by the FPSP and the comparative methods. In this figure, training images of baselines 1 and 2 and FPSP were selected by AIS.

Table 2. Comparison of performance in multiple evaluation indices. The mark (↑) means that the higher the index becomes, the higher the performance increases. Similarly, the mark (↓) means that the lower the index becomes, the higher the performance increases. Note that 100 (=C) selected images were used for training in baselines 1 and 2 and FPSP. It should be noted that the bold font represents the highest value in its evaluation index.

Methods	**CC↑**	**Sim↑**	**KLdiv↓**
Itti	0.3218	0.3911	9.0397
Gignature	0.4126	0.4122	8.0410
SalGAN	0.6345	0.5689	3.5597
Baseline1 based on ISVF	0.0953	0.3140	11.029
Baseline1 based on ISPSM	0.0762	0.3100	11.161
Baseline1 based on AIS	0.4013	0.4165	7.641
Baseline2 based on ISVF	0.4842	0.4274	4.014
Baseline2 based on ISPSM	0.4761	0.4170	3.057
Baseline2 based on AIS	0.5972	0.5032	4.133
FPSP based on AIS (Ours)	**0.7845**	**0.6557**	**1.083**

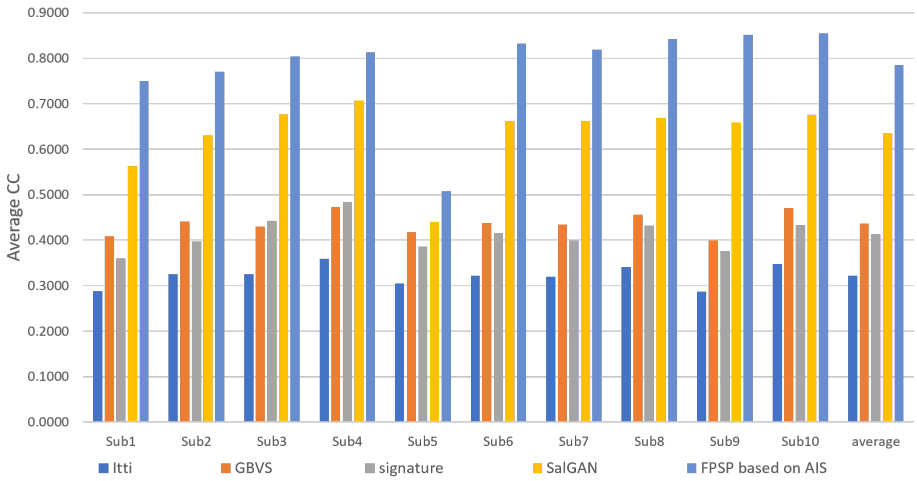

Figure 6. Average CC of each target person (↑) with comparison to USM prediction methods.

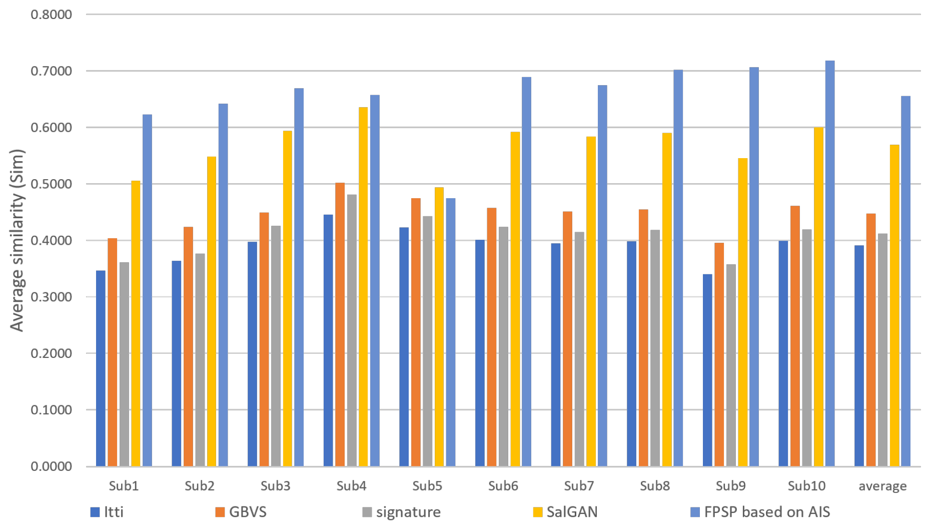

Figure 7. Average similarity (Sim) of each target person (↑) with comparison to USM prediction methods.

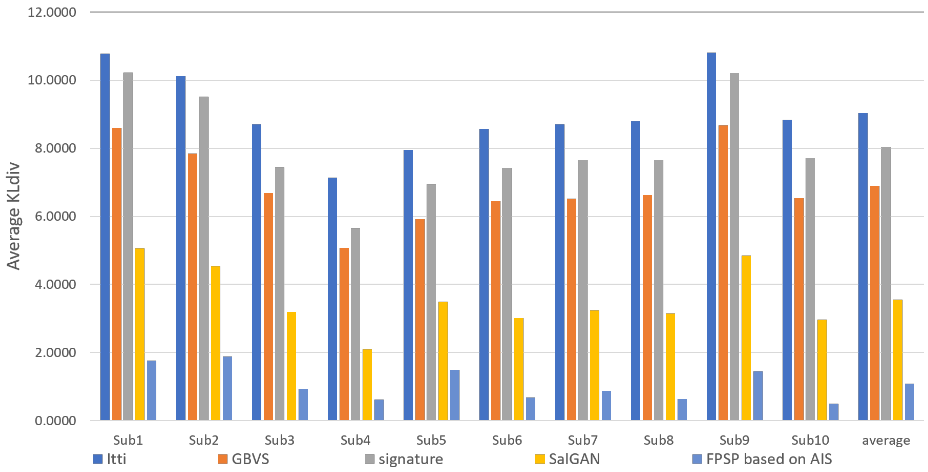

Figure 8. Average KLdiv of each target person (↓) with comparison to USM prediction methods.

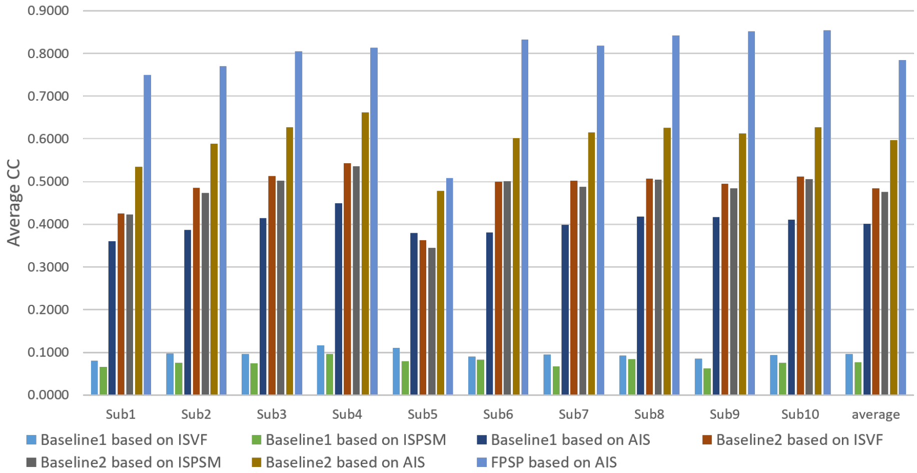

Figure 9. Average CC of each target person (↑) with comparison to PSM prediction methods.

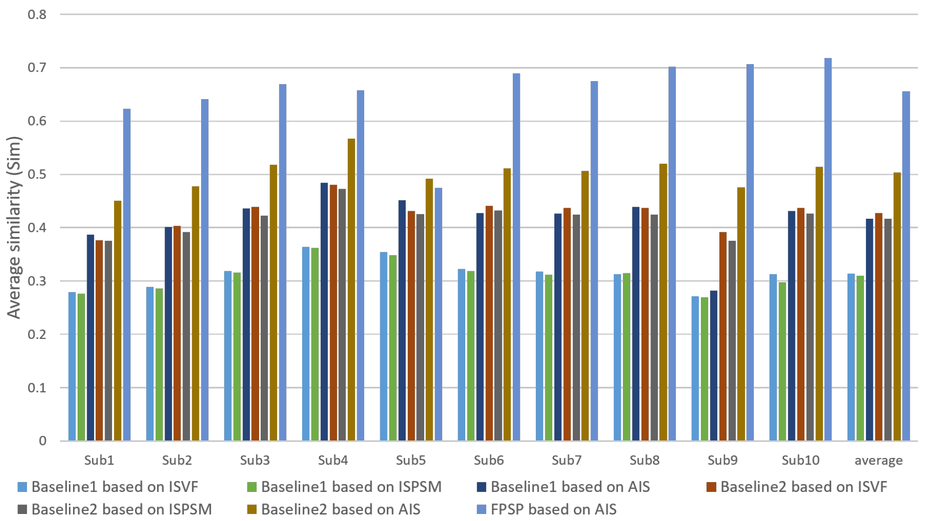

Figure 10. Average similarity (Sim) of each target person (↑) with comparison to PSM prediction methods.

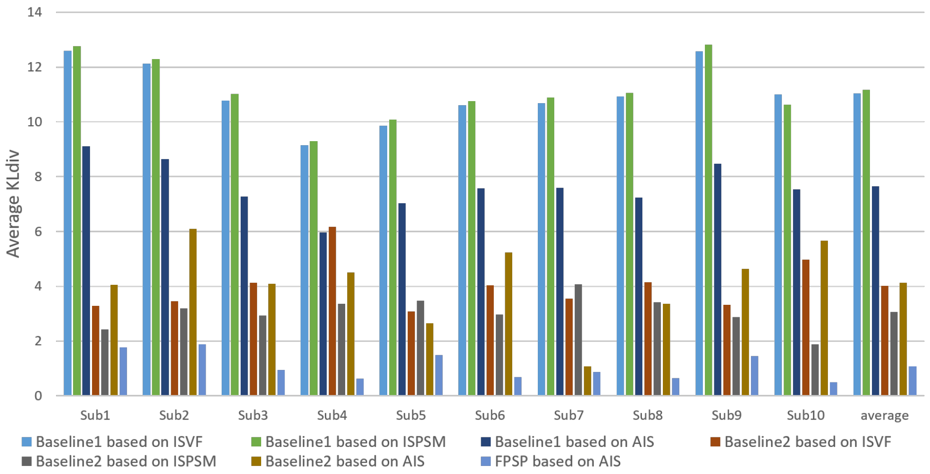

Figure 11. Average KLdiv of each target person (↓) with comparison to PSM prediction methods.

We show the results predicted by FPSP based on AIS and the USM prediction methods for each subject in Figures 6–8. Note that we denote 10 target persons as Subs 1–10 in these figures. These figures show that FPSP enables the person-specific prediction for most persons more successfully compared to the USM prediction methods. Specifically, FPSP outperforms SalGAN, which is one of the state-of-the-art USM prediction methods. Thus, we confirm the effectiveness of the construction of the prediction model for each person. Furthermore, we show the results predicted by FPSP based on AIS and the PSM prediction methods for each subject in Figures 9–11. These figures show that the results predicted by FPSP are higher than those of other PSM prediction methods. Thus, FPSP enables more accurate prediction than baseline PSM prediction methods. Therefore, the effectiveness of FPSP is verified in the first experiment.

Next, we discuss the difference between AIS, ISVF, and ISPSM in the second experiment. Focusing on the baselines in Table 2, we can confirm that the use of AIS is the most effective image selection method.

Furthermore, Figure 12 shows the performance of FSPS with changes in the number of training images when the training images are selected by AIS, ISVF and ISPSM for the calculation of the person similarity. In CC and KLdiv, the results of FPSP based on AIS are robust to changes in the number of training images and constantly higher than that of AIS and ISPSM. In other words, FPSP based on AIS enables accurately predicting the PSM of the target person just by gazing at 10 images included in the PSM dataset. Thus, it is convinced that our image selection method, AIS, is also effective for FPSP. Therefore, the effectiveness of FPSP based on AIS is verified by the experimental results.

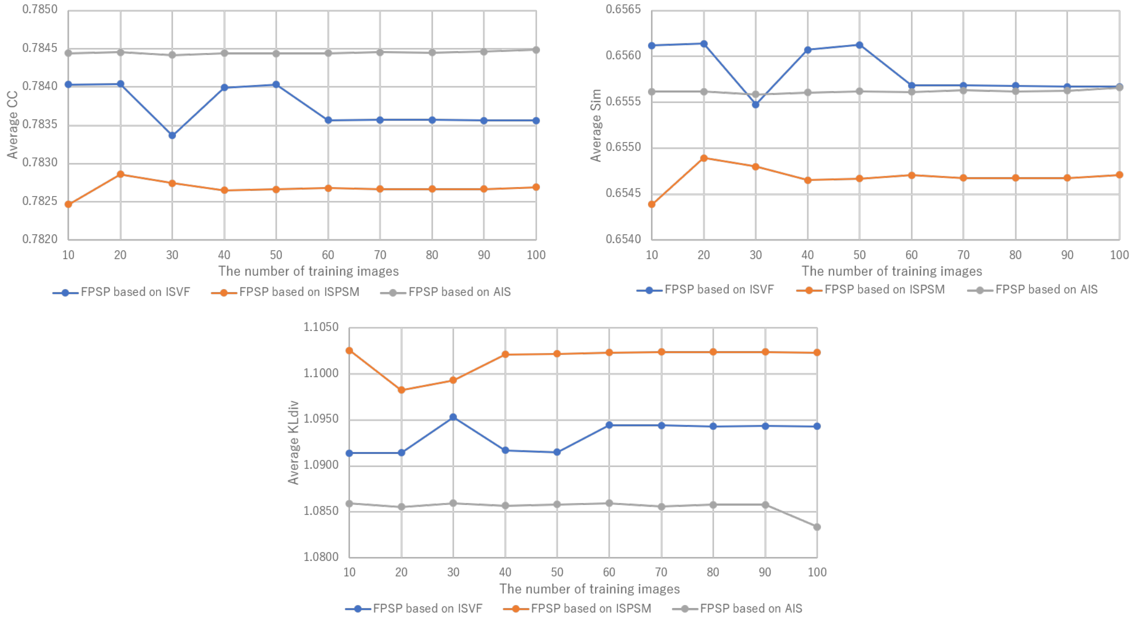

Figure 12. The prediction performance with changes in the number of training images. The robustness of FPSP based on AIS is verified.

We summarize the discussions. We confirm the effectiveness of the proposed PSM prediction method, FPSP, in Figure 5 and Table 2 from the perspective of the qualitative and quantitative evaluations by focusing the average results. Moreover, by comparing FPSP with USM prediction methods and baseline PSM prediction methods for each person in Figures 6–11, it is verified that FPSP enables the accurate prediction for each person. Finally, Figure 12 confirms the robustness and effectiveness of AIS for FPSP. Therefore, we reveal that FPSP based on AIS enables the accurate prediction with a small number of training images and reduces the burden of persons to obtain their gaze data for the PSM prediction.

4. Conclusions

In this paper, we have proposed few-shot personalized saliency prediction based on adaptive image selection considering object and visual attention. FPSP enables the accurate PSM prediction with the small number of training images. Moreover, AIS realizes that the number of images that the new person views becomes smaller. Finally, FPSP based on AIS enables the accurate prediction with the small number of training images and reduces the burden of persons to obtain their gaze data for the PSM prediction. Experimental results showed the effectiveness of our proposed method.

Author Contributions: Conceptualization, Y.M., K.M., T.O., and M.H.; methodology, Y.M., K.M., T.O., and M.H.; software, Y.M.; validation, Y.M., K.M., T.O., and M.H.; data curation, Y.M.; writing—original draft preparation, Y.M.; writing—review and editing, K.M., T.O., and M.H.; visualization, Y.M.; funding acquisition, T.O. and M.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported by the MIC/SCOPE #18160100.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Itti, L.; Koch, C.; Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 1254–1259. [CrossRef]
2. Harel, J.; Koch, C.; Perona, P. Graph-based visual saliency. In Proceedings of the the Advances in Neural Information Processing Systems 20: 21st Annual Conference on Neural Information Processing Systems 2007, Vancouver, BC, Canada, 3–6 December 2007.
3. Hou, X.; Harel, J.; Koch, C. Image signature: Highlighting sparse salient regions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 194–201.
4. Pan, J.; Ferrer, C.; McGuinness, K.; O'Connor, N.; Torres, J.; Sayrol, E.; Giro, X. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv* **2017**, arXiv:1701.01081.
5. Setlur, V.; Takagi, S.; Raskar, R.; Gleicher, M.; Gooch, B. Automatic image retargeting. In Proceedings of the 4th International Conference on Mobile and Ubiquitous Multimedia, Christchurch, New Zealand, 8–10 December 2005.
6. Fang, Y.; Zhang, C.; Li, J.; Lei, J.; Da Silva, M.P.; Le Callet, P. Visual attention modeling for stereoscopic video: A benchmark and computational model. *IEEE Trans. Image Process.* **2017**, *26*, 4684–4696. [CrossRef] [PubMed]
7. Itti, L. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Trans. Image Process.* **2004**, *13*, 1304–1318. [CrossRef] [PubMed]
8. Li, S.; Xu, M.; Ren, Y.; Wang, Z. Closed-form optimization on saliency-guided image compression for HEVC-MSP. *Trans. Multimedia* **2017**, *20*, 155–170. [CrossRef]
9. Gasparini, F.; Corchs, S.; Schettini, R. Low-quality image enhancement using visual attention. *Opt. Eng.* **2007**, *46*. [CrossRef]
10. Fan, F.; Ma, Y.; Huang, J.; Liu, Z. Infrared image enhancement based on saliency weight with adaptive threshold. In Proceedings of the 3rd International Conference on Signal and Image Processing (ICSIP 2018), Shenzhen, China, 13–15 July 2018.
11. Alwall, N.; Johansson, D.; Hansen, S. The gender difference in gaze-cueing: Associations with empathizing and systemizing. *Pers. Indiv. Differ.* **2010**, *49*, 729–732. [CrossRef]
12. Fan, S.; Shen, Z.; Jiang, M.; Koenig, B.; Xu, J.; Kankanhalli, M.; Zhao, Q. Emotional attention: A study of image sentiment and visual attention. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018.
13. Imafuku, M.; Kawai, M.; Niwa, F.; Shinya, Y.; Inagawa, M.; Myowa-Yamakoshi, M. Preference for dynamic human images and gaze-following abilities in preterm infants at 6 and 12 months of age: An Eye-Tracking Study. *Infancy* **2017**, *22*, 223–239. [CrossRef]
14. Xu, Y.; Gao, S.; Wu, J.; Li, N.; Yu, J. Personalized saliency and its prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, 2975–2989. [CrossRef] [PubMed]
15. Gygli, M.; Grabner, H.; Riemenschneider, H.; Nater, F.; Van Gool, L. The interestingness of images. In Proceedings of IEEE International Conference on Computer Vision (ICCV 2013), Sydney, Australia, 1–8 December 2013; pp. 1633–1640.
16. Li, Y.; Xu, P.; Lagun, D.; Navalpakkam, V. Towards measuring and inferring user interest from gaze. In Proceedings of International Conference on World Wide Web Companion, Perth, Australia, 3–7 April 2017.
17. Bazrafkan, S.; Kar, A.; Costache, C. Eye gaze for consumer electronics: Controlling and commanding intelligent systems. *IEEE Consum. Electron. Mag.* **2015**, *4*, 65–71. [CrossRef]
18. Zhao, Q.; Chang, S.; Harper, M.; Konstan, J. Gaze prediction for recommender systems. In Proceedings of the ACM Conference on Recommender Systems, Boston, MA, USA, 15–19 September 2016.
19. Xu, Y.; Li, N.; Wu, J.; Yu, J.; Gao, S. Beyond universal saliency: Personalized saliency prediction with multi-task CNN. In Proceedings of the International Joint Conferences on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017.
20. Yin, X.; Liu, X. Multi-task convolutional neural network for pose-invariant face recognition. *Trans. Image Process.* **2017**, *27*, 964–975. [CrossRef] [PubMed]
21. Moroto, Y.; Maeda, K.; Ogawa, T.; Haseyama, M. Estimation of user-specific visual attention based on gaze information of similar users. In Proceedings of the IEEE 8th Global Conference on Consumer Electronics (GCCE 2019), Las Vegas, NV, USA, 15–18 October 2019.

22. Moroto, Y.; Maeda, K.; Ogawa, T.; Haseyama, M. User-specific visual attention estimation based on visual similarity and spatial information in images. In Proceedings of the IEEE International Conference on Consumer Electronics—Taiwan (IEEE 2019 ICCE-TW), Ilan, Taiwan, 20–22 May 2019.
23. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
24. Judd, T.; Ehinger, K.; Durand, F.; Torralba, A. Learning to predict where humans look. In Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV 2009), Kyoto, Japan, 29 September–2 October 2009.
25. Bottou, L. Large-scale machine learning with stochastic gradient descent. In Proceedings of the 19th international symposium on computational statistics, Paris, France, 22–27 August 2010.
26. Judd, T.; Durand, F.; Torralba, A. *A Benchmark of Computational Models of Saliency to Predict Human Fixations*; Technical Report; MITCSAIL-TR-2012-001; Massachusetts Institute of Technology Press: Cambridge, MA, USA, 2012.
27. Bylinskii, Z.; Judd, T.; Oliva, A.; Torralba, A.; Durand, F. What do different evaluation metrics tell us about saliency models? *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 740–757. [CrossRef] [PubMed]
28. Bylinskii, Z.; Judd, T.; Borji, A.; Itti, L.; Durand, F.; Oliva, A.; Torralba, A. Mit Saliency Benchmark. 2015. Available online: http://saliency.mit.edu/ (accessed on 11 April 2020).
29. Jiang, M.; Huang, S.; Duan, J.; Zhao, Q. Salicon: Saliency in context. In Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015), Boston, MA, USA, 7–12 Jun 2015.
30. Moroto, Y.; Maeda, K.; Ogawa, T.; Haseyama, M. User-centric visual attention estimation based on relationship between image and eye gaze data. In Proceedings of the 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE 2018), Nara, Japan, 9–12 October 2018.
31. Huang, G.; Liu, Z.; Maaten, L.; Weinberger, K. Densely connected convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017.