 Copyright © 2025 by ITE Transactions on Media Technology and Applications (MTA)

Paper

Few-shot Personalized Saliency Prediction Based on Interpersonal Gaze Patterns

Yuya Moroto †, Keisuke Maeda ††, Takahiro Ogawa (member)††, Miki Haseyama (member)††

Abstract This study proposes a few-shot personalized saliency prediction method that leverages interpersonal gaze patterns. Unlike general saliency maps, personalized saliency maps (PSMs) capture individual visual attention and provide insights into individual visual preferences. However, predicting PSMs is challenging because of the complexity of gaze patterns and the difficulty of collecting extensive eye-tracking data from individuals. An effective strategy for predicting PSMs from limited data is the use of eye-tracking data from other persons. To efficiently handle the PSMs of other persons, this study focuses on the selection of images to acquire eye-tracking data and the preservation of the structural information of PSMs. In the proposed method, these images are selected such that they bring more diverse gaze patterns to persons, and structural information is preserved using tensor-based regression. The experimental results demonstrate that these two factors are beneficial for few-shot PSM prediction.

Key words: Salinecy prediction, personalized saliency map, tensor-based regression, person similarity, adaptive image selection.

1. Introduction

Humans can selectively obtain vital information from the abundant visual information in the complex real world through their visual system. Many researchers have attempted to introduce such human mechanisms into image processing models[1)2)3)]. Specifically, a saliency map, which represents the salient parts that are more noticeable than the neighboring parts, is predicted to reproduce the human instinctive visual perception[1)4)5)6)7)]. A saliency map is predicted for each image without personalization. However, different individuals focus on different areas even when viewing the same scene, that is, individual differences exist[8)9)10)]. To model individual visual attention, saliency maps have been personalized over the past few years[11)12)13)14)15)]. A traditional saliency map and its personalization are distinguished by referring to them as a universal saliency map (USM) and a personalized saliency map (PSM), respectively. A USM omits individual differences, whereas a PSM is predicted for each person. Personalized visual preferences can be reflected by differences between PSMs[16)17)18)]; thus, individuality can be useful in many situations. Such individuality has significant potential in applications where personal visual preferences play a crucial role. For instance, PSM prediction can be applied to targeted advertising, where understanding individual visual preference helps determine advertisement placements or make UI designs[19)20)]. Another example is a recommender system, which highlights the regions or contents related to users' interests in images or products[21)22)]. Moreover, PSMs can contribute to transferring skills related to tacit knowledge, which is knowledge that is difficult to verbalize but manifests in habitual behaviors or internalized expertise[23)24)]. These applications demonstrate the practical relevance of predicting PSMs beyond USMs. Here, to obtain a PSM for unseen images in advance, the PSM should be predicted from the individual gaze pattern tendency.

To model individual gaze patterns, the relationship between visual stimuli, e.g., images, and the individual PSM should be analyzed based on eye-tracking data obtained from each person in the past. Then, the gaze patterns emerging in the images are complex and different, and these characteristics lead to the difficulty of PSM prediction. To extract individual gaze pattern tendencies, several researchers have collected eye-tracking data for thousands of images[11)12)14)16)]. The prediction mod-

Received February 3, 2025; Revised May 24, 2025; Accepted July 18, 2025

† Graduate School of Information Science and Technology, Hokkaido University
(Sapporo, Japan)
†† Faculty of Information Science and Technology, Hokkaido University
(Sapporo, Japan)

els adopted in these studies are based on deep learning, which requires a massive amount of training data for each person. The large-scale PSM dataset is openly available; however, the acquisition of a massive amount of individual eye-tracking data can be a significant burden and time-consuming task for new persons in the application. Consequently, a PSM prediction method with a limited amount of training eye-tracking data is required.

To predict a PSM from a limited amount of data, an effective strategy is to use PSMs obtained from persons with similar gaze patterns to the target person. To determine whether a person has gaze patterns similar to those of the target person, several pairs of eye-tracking data for the same images are required. However, such pairs cannot be acquired in large quantities, and the selection of images to acquire eye-tracking data is an important process. In a previous study[25], images that induce the scattering of gazes were selected using adaptive image selection (AIS) to efficiently and steadily obtain the similarity of gaze patterns between the target and other persons (called training persons in this study). Additionally, in a previous study[13], the collaborative multi-output Gaussian process regression (CoMOGP)[26] was used with the PSMs obtained from training persons for PSM prediction. However, such regression-based methods require vector-format input, and the structural information of PSMs cannot be effectively used.

Structural information is an important clue to detecting salient areas in the human visual system[1]. The method proposed in the previous work[1] extracts handcrafted image features and takes the center-surround differences of them. In this method, as a result of feature extraction, several feature maps are calculated with considering the pixel positions and their relationships, that is, the two-dimensional spatial configuration is preserved. Inspired by this process, we hypothesized that such two-dimensional spatial configuration is useful for PSM prediction and captures the relative positioning and distribution of salient regions within an image. In this way, this paper focuses on the two-dimensional spatial configuration as the structural information for performance improvement of PSM prediction. Then, it is necessary to construct a PSM prediction method that considers structural information that is compatible with the effective use of PSMs predicted for several training persons. Therefore, to improve few-shot PSM prediction, it is desirable to collaboratively incorporate the adaptive selection of images to acquire eye-tracking data and preservation of the structural information of PSMs predicted for training persons.

We propose a few-shot personalized saliency prediction method based on interpersonal gaze patterns. In the proposed method, we collaboratively use AIS[25] and the tensor-based regression model[27]. The AIS scheme focuses on the variety of selected images and the variation in PSMs obtained from the training persons for selecting images. Through the AIS scheme, we can efficiently and steadily obtain the similarity of gaze patterns between the target and training persons. In addition, the input and output of the tensor-based regression model[27] are in a multi-array tensor format; thus, it predicts the PSMs of the target person from the PSMs of training persons while preserving the structural information. Therefore, we realize the effective selection of images to acquire eye-tracking data and preservation of the structural information of PSMs predicted for training persons.

2. Related Works

2.1 USM Prediction

In the field of image processing, USM prediction is a traditional research subject. Specifically, early USM models were constructed based on hand-crafted image features until the development of deep learning methods[1)4)5]. In contrast, deep learning methods, e.g., convolutional neural networks (CNNs), generative adversarial networks (GANs), and vision transformers, have outperformed these models, which do not require training phase[6)7]. Although many USM prediction methods have been proposed, they have limitations in terms of the performance improvement of PSM prediction because they do not account for individual differences.

2.2 PSM Prediction

The advancement of measurement instruments has sparked interest in PSM prediction over the last decade. The open large-scale dataset significantly contributes to the construction of PSM prediction models[11]. Specifically, a multi-task CNN-based model achieved highly accurate PSM prediction by focusing on the difference between USMs and PSMs[11]. In addition, a CNN model with person-specific information encoded filters (CNN-PIEF) was proposed as an extended version of the previous study[16]. In CNN-PIEF, the embeddings of person-specific information enable the personalization of the prediction model. In addition, in a previous work[29], a model based on conditional GANs with person clusters

Table 1 Comparison of representative methods for USM and PSM prediction.

Method	Personalization	Order of Training data	Additional Data
USM Prediction			
Computational models[1)4)5)]	-	-	-
Deep learning-based USM prediction[6)7)]	-	10^4	-
PSM Prediction			
Multi-task CNN[11)]	✓	10^3 / Person	-
CNN-PIEF[16)]	✓	10^3 / Person	Person-specific information
Sherkati et al.[29)]	✓	10^3 / Person	Person-specific information
Strohm et al.[30)]	✓	10^3 / Person	Pretrained person embedding network
Person similarity-based approach[13)25)31)] (Our setting)	✓	$\mathbf{10^2}$ / Person	Other person's eye tracking data

constructed from person-specific information was proposed, and in another study[30)], a siamese CNN-based model for learning user embedding was proposed. These models successfully predicted PSMs using deep learning. However, deep learning requires a sufficient amount of training eye-tracking data to train CNNs, and significant amounts of training data are required to make predictions for a new person.

In this regard, several studies have attempted to reduce the amount of training data by using eye-tracking data obtained from persons with similar gaze patterns to the target person[13)25)31)]. However, in these methods, the structural information of PSMs cannot be effectively used. Structural information is an important clue to detecting salient areas in the human visual system[1)]. Thus, it is necessary to construct a PSM prediction method that considers structural information that is compatible with the effective use of PSMs predicted for several training persons.

We show the comparison of the representative methods for USM and PSM prediction in Table 1. As shown in the table, while the USM prediction methods can be trained on a large amount of training data by collecting gaze data from various people, the PSM prediction methods should be trained on a limited amount of training data for each person due to personalized prediction. Furthermore, deep learning-based PSM prediction methods[11)16)29)30)] still require a large amount of training data, about 1,000 eye-tracking data per person, which is a heavy burden on the collection of training data from new persons. With regard to CNN-PIEF[16)] and Sherkati et al.[29)], PSMs can be predicted by using the person-specific information, e.g., age, and gender, but such information is not always available in the real-world applications. In the study by Strohm et al.[30)], the person embedding network is pretrained on a large amount of training data. On the other hand, our setting requires a moderate amount of training data, 100 training data per person, but offers an adequate balance between personalization and scalability. Therefore, the performance comparison is difficult due to the difference of the training data and the research objective.

3. Proposed Few-shot PSM Prediction

The proposed few-shot PSM prediction comprises three phases, and the entire flow is depicted in Fig. 1. Here, we assume that there are P training persons with a massive amount of eye-tracking data and a target person with a limited amount of eye-tracking data. First, the multi-task CNN[28)] is trained to predict the PSMs of the training persons by referring to the previous studies[11)16)]. Next, we select common images that the target person gazes at based on the AIS scheme[25)]. The common images are selected such that they bring more diverse gaze patterns to persons. Finally, the proposed method predicts the PSM using tensor-to-matrix regression[27)] with the PSMs of the training persons. Therefore, by efficiently using the interpersonal gaze patterns, we can effectively select images to acquire eye-tracking data and preserve the structural information of PSMs predicted for training persons.

3.1 Multi-Task CNN for Training Persons

To train the multi-task CNN model[28)], we prepare the training 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) and their USMs $\boldsymbol{U}(\boldsymbol{X}_n) \in \mathbb{R}^{d_1 \times d_2}$, where $d_1 \times d_2$ and d_3 denote the size of the image and the color channel, respectively. To effectively obtain the predicted PSMs of training persons, previous studies[11)16)] have adopted the specific approach of predicting the difference map

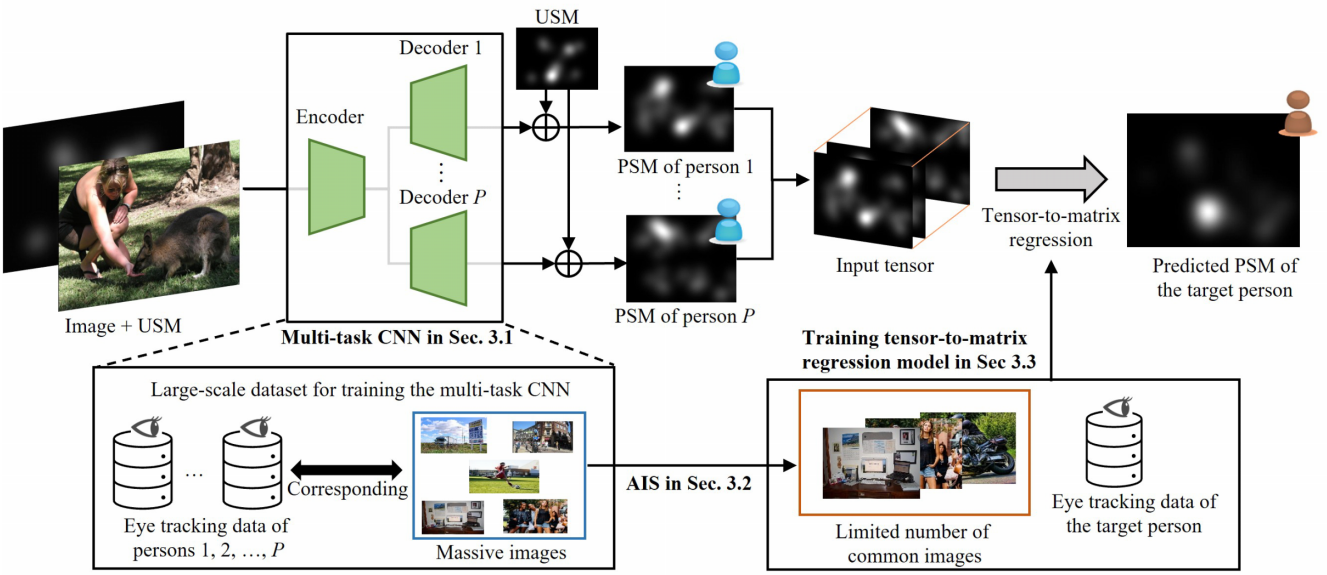

Fig. 1 Entire flow of the proposed PSM prediction method consisting of three phases. In the first phase, the multi-task CNN[28] predicts the PSMs of P training persons. Next, using the AIS scheme[25], I images are selected as common images that the target person gazes at. Finally, the PSM is predicted using tensor-to-matrix regression [27] with the PSMs of training persons.

$\boldsymbol{M}(\boldsymbol{X})_p \in \mathbb{R}^{d_1 \times d_2}$ $(p = 1, 2, \ldots, P)$ between USMs and PSMs as follows:

$$\boldsymbol{M}(\boldsymbol{X})_p = \boldsymbol{S}(\boldsymbol{X})_p - \boldsymbol{U}(\boldsymbol{X}), \tag{1}$$

where $\boldsymbol{S}(\boldsymbol{X})_p$ denotes the PSM of the pth training person based on the eye-tracking data for the image $\boldsymbol{X}$. Next, to simultaneously predict the PSMs of training persons, we construct a multi-task CNN consisting of one image encoder and P PSM decoders and optimize their trainable parameters by minimizing the following objective function:

$$\sum_{p=1}^{P}\sum_{n=1}^{N}\sum_{l=1}^{L} ||\hat{\boldsymbol{M}}_l(\boldsymbol{X}_n)_p - \boldsymbol{M}(\boldsymbol{X}_n)_p||_F^2, \tag{2}$$

where $\hat{\boldsymbol{M}}_l(\boldsymbol{X}_n)_p$ $(l = 1, 2, \ldots, L$; L being the number of convolution layers in one decoder) denotes a predicted difference map calculated from the lth layer, and $||\cdot||_F^2$ denotes the Frobenius norm.

Given the test image $\boldsymbol{X}_{\text{tst}}$, the predicted PSM of the pth person is calculated as follows:

$$\hat{\boldsymbol{S}}(\boldsymbol{X}_{\text{tst}})_p = \hat{\boldsymbol{M}}_L(\boldsymbol{X}_{\text{tst}})_p + \boldsymbol{U}(\boldsymbol{X}_{\text{tst}}). \tag{3}$$

Therefore, the multi-task CNN can simultaneously predict the PSMs of the training persons and consider the relationship between these PSMs.

3.2 Adaptive Image Selection for PSM Prediction

We select a few images from the N training images to obtain the tendency of the target and training persons to be similar. To effectively analyze such similarity, the I common images that produce more diverse gaze patterns to persons are selected using the AIS scheme[25]. Specifically, the AIS scheme focuses on various common images and variations in PSMs obtained from the training persons. To simultaneously consider these factors, the AIS scheme uses the variation in PSMs for objects in each image.

First, we calculate the PSMs and their variance for each object $\boldsymbol{B}_{n,j}$ $(j = 1, 2, \ldots, J$; J being the number of object categories in the training images) in the training images $\boldsymbol{X}_n$. Then, object detection [32] is applied to the training images to obtain a rectangle with dimensions of $d_{n,j}^h \times d_{n,j}^w$ for the jth object in the ith image. The PSM variance $q_{n,j}$ for object $\boldsymbol{B}_{n,j}$ is calculated as follows:

$$q_{n,j} = \frac{1}{d_{n,j}^h d_{n,j}^w P}\sum_{p=1}^{P} ||\bar{\boldsymbol{S}}(\boldsymbol{B}_{n,j})_p \odot \bar{\boldsymbol{S}}(\boldsymbol{B}_{n,j})_p||_F^1, \tag{4}$$

$$\bar{\boldsymbol{S}}(\boldsymbol{B}_{n,j})_p = \boldsymbol{S}(\boldsymbol{B}_{n,j})_p - \frac{1}{P}\sum_{p=1}^{P}\boldsymbol{S}(\boldsymbol{B}_{n,j})_p, \tag{5}$$

where $\boldsymbol{S}(\boldsymbol{B}_{n,j})_p$ denotes the PSM for object $\boldsymbol{B}_{n,j}$ of person p, and $\odot$ denotes the operator of the Hadamard

product. Then, we set $q_{n,j} = 0$ when $\boldsymbol{X}_n$ does not include the jth object and set the largest $q_{n,j}$ when the image $\boldsymbol{X}_n$ includes several mth objects. Then, we obtain the sum of $q_{n,j}$ for nth image as follows:

$$\bar{q}_n = \sum_{j=1}^{J} q_{n,j}. \tag{6}$$

Finally, using $\bar{q}_n$, we select the top I images as common images under the constraint to maximize the number of object categories in these images. Consequently, the selected common images have multiple object categories, and the objects in these images exhibit a high PSM variance. Specifically, the AIS scheme focuses on the PSM variance to analyze the differences in gaze patterns and persons' visual preferences, and a higher number of object categories leads to greater scene diversity. For images, the AIS scheme has demonstrated its strength when image selection for training images exhibits high diversity. In addition, the AIS scheme is based on object-level gaze analysis, and it contributes to maintaining the diversity of visual information. In contrast, for PSMs, the target person's gaze pattern can be represented by combining the gaze patterns of multiple persons even if no person has a similar gaze pattern, and this possibility increases with the number of training persons. In addition, if the eye-tracking data of a single person contain noise or outliers, the effect can be reduced in the AIS scheme by focusing on the PSM variance of several persons.

3.3 PSM Prediction via Tensor-to-Matrix Regression

This subsection presents the tensor-to-matrix regression model for the few-shot PSM prediction. The comparison of the tensor-input regression with the vector-input regression is shown in Fig. 2. The tensor-input regression is superior to the vector-input regression in the points that it can preserve the structural information, which means the two-dimensional spatial configuration in this paper, within and between PSMs.

In the proposed method, we do not explicitly define or compute gaze pattern similarity between individuals. Instead, the similarity is implicitly captured through the training process of the tensor-to-matrix regression model. Specifically, the model is trained to predict the target person's PSMs using the PSMs of training persons as input. During training, the weights are adjusted such that individuals with more similar gaze tendencies contribute more to the prediction. Therefore, similarity is learned as part of the regression process, without measuring similarities or prior clustering of individuals.

The PSMs predicted in Sec. 3.1 are used to predict the PSM of the target person. Several PSMs are treated as input. The input tensor $\mathcal{S}(\boldsymbol{X}_i) \in \mathbb{R}^{P \times d_1 \times d_2}$ $(i = 1, 2, \ldots, I)$ corresponding to the image $\boldsymbol{X}_i$ chosen in Sec. 3.2 is constructed as follows:

$$\mathcal{S}(\boldsymbol{X}_i) = [\hat{\boldsymbol{S}}(\boldsymbol{X}_i)_1, \hat{\boldsymbol{S}}(\boldsymbol{X}_i)_2, \ldots, \hat{\boldsymbol{S}}(\boldsymbol{X}_i)_P]. \tag{7}$$

In addition, we prepare the supervised PSM $\boldsymbol{S}(\boldsymbol{X}_i)_{p^{\mathrm{tst}}}$ of the target person p^{tst} for the input tensor $\mathcal{S}(\boldsymbol{X}_i)$. Here, we assume that the target person gazes only at the common images selected in Sec. 3.2, and we can obtain the supervised PSM $\boldsymbol{S}(\boldsymbol{X}_i)_{p^{\mathrm{tst}}}$. In a tensor-to-matrix regression scenario, the weight tensor $\mathcal{W} \in \mathbb{R}^{P \times d_1 \times d_2 \times d_1 \times d_2}$ is used to predict the PSM of a newly given image as follows:

$$\boldsymbol{S}_{\mathrm{TReg}}(\boldsymbol{X}_{\mathrm{tst}})_{p^{\mathrm{tst}}} = \langle \mathcal{S}(\boldsymbol{X}_{\mathrm{tst}}), \mathcal{W} \rangle_3, \tag{8}$$

where $\langle \cdot, \cdot \rangle_Q$ denotes the tensor product and Q denotes the number of input arrays.

To optimize the weight tensor $\mathcal{W}$, we minimize the sum of the squared errors using L_2 regularization as follows:

$$\min_{\mathrm{rank}(\mathcal{W}) \leqq R} \sum_{i=1}^{I} ||\boldsymbol{S}(\boldsymbol{X}_i)_{p^{\mathrm{tst}}} - \langle \mathcal{S}(\boldsymbol{X}_i), \mathcal{W} \rangle_3||_F^2 + \lambda ||\mathcal{W}||_F^2. \tag{9}$$

Note that it is difficult to solve this minimization problem because the inputs and outputs are in a multi-array format. Thus, by referring to a previous study[27], we assume that $\mathcal{W}$ has the reduced PARAFAC/CANDECOMP (CP)-rank such that $\mathrm{rank}(\mathcal{W}) \leqq R$ and solve Eq. (9) under this constraint. Although tensor-based regression tends to suffer from high-dimensional problems and overfitting, the low-rank approximation mitigates such problems. In addition, L_2 regularization suppresses overfitting. The reduced CP-rank constraint serves as a dimensionality reduction strategy, which utilizes a multi-way structure to improve generalization and prediction efficiency. In this formulation, the ridge (L2) regularization is further interpreted as a Bayesian prior, which provides statistical motivation for the prediction procedure[27]. Therefore, using the tensor-to-matrix regression model, the proposed method preserves structural information without vectorizing the input tensor and output matrix.

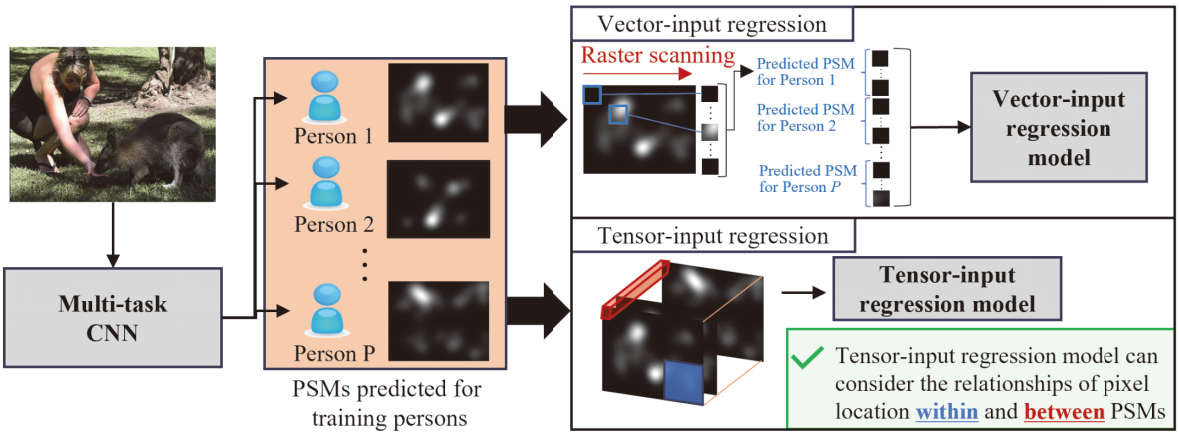

Fig. 2 Comparison of the tensor-input regression with the vector-input regression in the proposed approach. The tensor-input regression is superior to the vector-input regression in the points that it can preserve the structural information, which means the two-dimensional spatial configuration in this paper, within and between PSMs.

4. Experiments

4.1 Dataset

In this experiment, an open-source large-scale PSM dataset[16] was used. The PSM dataset was constructed such that the included images have high diversity and is suitable for performance and robustness evaluation. The PSM dataset comprises 1,600 images with corresponding eye-tracking data obtained from 30 participants. The participants had normal or corrected visual acuity and gazed at one image for three seconds under free viewing conditions. To evaluate the predicted PSMs, we constructed the PSMs of each participant for all images from the eye-tracking data as the ground truth (GT) map based on a previous work[33]. As the USM used in the proposed method, we adopted the mean PSMs of the training persons to reduce the influence of USM prediction errors.

In the proposed method, we required training images with eye-tracking data to train the multi-task CNN model and common images selected from the training images to train the tensor-to-matrix regression model. Thus, 1,100 images were randomly selected for training and the remaining 500 images were used as test images in the experiment. In addition, I common images were selected from the training images based on the AIS scheme. Note that the common images selected by AIS are used exclusively for training the prediction model. The evaluation is performed on a separate test set that is not influenced by the AIS for ensuring that the prediction results are not biased by the image selection mechanism. In addition, we randomly selected 20 participants as training persons, and the remaining 10 persons were treated as target persons. The number of training subjects and common images were determined empirically to achieve a practical balance between training and test data. The 20 training persons were randomly selected to ensure sufficient variability in gaze patterns. The I common images were selected from training images via the AIS scheme. If the number of common images is too small, the prediction may become unstable due to depending on a limited set of gaze patterns. While, if the number of common images is too large, the target person should gaze at the massive number of images, which is unrealistic, although the prediction may become stable. In this way, we experimentally determined the number of common images, and its value is described in Sec. 4.2. Although the eye-tracking data of the target persons were available, we only used the eye-tracking data of the target persons as the common images for PSM prediction because we assumed that the target persons gazed at the common images.

4.2 Experimental Settings

We optimized the multi-task CNN model in Sec. 3.1 and the tensor-to-matrix regression model in Sec. 3.3. The multi-task CNN model was optimized via the stochastic gradient descent[34] by referring to a previous study[16], and the number of layers (L), momentum, batch size, epoch, and learning rate were set to 3, 0.9, 9, 1,000, and 3.0×10^{-5}, respectively. In addition, the tensor-to-matrix regression model was optimized by simply differentiating the weight parameters with tensor unfolding. Furthermore, we set $I = 100$ and conducted additional experiments focusing on hy-

perparameter analysis by varying the hyperparameters of the tensor-to-matrix regression model over $R \in \{5, 10, \ldots, 50\}$ and $\lambda \in \{0.01, 0.1, \ldots, 10000\}$ to examine their effect on performance.

To objectively evaluate the proposed method, we adopted several USM and PSM prediction methods as comparison methods. We adopted the following USM prediction methods: Signature[5], GBVS[4], Itti[1], SalGAN[6], and Contextual[7]. Signature, GBVS, and Itti are computational models that predict USMs only from input images. SalGAN and Contextual are deep learning-based models trained on the SALICON dataset[35], which is a large-scale eye-tracking dataset, without considering personalization. The following two few-shot PSM prediction (FPSP) methods using only common images and their eye-tracking data were adopted.

Baseline1: PSM prediction using visual similarity between target and common images[36].

Baseline2: PSM prediction based on Baseline1 and USM prediction[37].

In addition, we compared three PSM prediction methods with settings similar to those of the proposed method: similarity-based FPSP[25], CoMOGP-based FPSP[13], and object-based gaze similarity (OGS)-based FPSP[31]. Note that although other PSM prediction methods exist[11)16)29)30], they cannot learn from a small amount of training data as discussed in Sec. 2. 2. Therefore, we adopted only the aforementioned comparison methods in our experiment.

As the evaluation metrics, we adopted the Kullback–Leibler divergence (KLdiv) and cross-correlation (CC) between the predicted PSM and the GT map based on previous research[38]. Specifically, KLdiv was used to evaluate the similarity of the distribution, that is, structural similarity, and CC was used to evaluate pixel-based similarity. By using these two metrics, both the global and local similarities between the predicted PSMs and their GT maps can be evaluated.

4. 3 Results and Discussion

Figure 3 presents the predicted results. Table 2 presents the quantitative evaluation results. As shown in Fig. 3, the PSMs predicted by the proposed method exhibit a distribution close to the GTs, demonstrating the effectiveness of preserving structural information. In addition, as shown in Table 2 that compares the proposed and comparison methods, the proposed method outperforms all comparison methods in terms of KLdiv. This result confirms that tensor-to-matrix regression is effective for PSM prediction considering structural information. The effectiveness of personalization for saliency prediction was confirmed because the proposed method outperformed the USM prediction methods. In addition, by comparing the proposed method with other PSM prediction methods, the effectiveness of focusing on structural information is confirmed. Moreover, Baselines1 and 2 do not use the PSMs of other persons, and the comparison between these methods and the proposed method demonstrates the effectiveness of using eye-tracking data obtained from other persons.

Table 2 Quantitative distribution-based evaluation based on KLdiv and pixel-based evaluation based on CC. A lower KLdiv value indicates higher performance, whereas a higher CC value indicates higher performance.

Methods	**KLdiv↓**	**CC↑**
Signature[5]	8.04	0.413
GBVS[4]	6.89	0.437
Itti[1]	9.04	0.322
SalGAN[6]	3.56	0.635
Contextual[7]	3.57	0.674
Baseline1[36]	7.64	0.401
Baseline2[37]	4.13	0.597
Similarity-based FPSP[25]	1.82	0.735
CoMOGP-based FPSP[13]	1.38	0.765
OGS-based FPSP[31]	1.09	**0.781**
Proposed Method ($R = 50$, $\lambda = 1000$)	**1.00**	0.775

The "KLdiv" of OGS-based FPSP[31] is similar to that of the proposed method. Here, OGS-based FPSP predicts PSMs by searching for objects in input images from the training images and using the PSMs corresponding to the regions of searched objects. In cases where an input image contains objects that were not present in the training dataset, OGS-based FPSP may be difficult to predict PSMs because it relies on PSMs of known objects. Consequently, its applicability may be limited when input images contain objects that were not seen during training. In contrast, the proposed method focuses on the relationships between persons' gaze patterns without semantic information, e.g., objects. Thus, our approach remains applicable even when the input images include objects that were not present in the training dataset. This characteristic allows the proposed method to be used in a broader range of scenarios.

The "CC" of the proposed method is comparable to that of other PSM methods such as the OGS-based

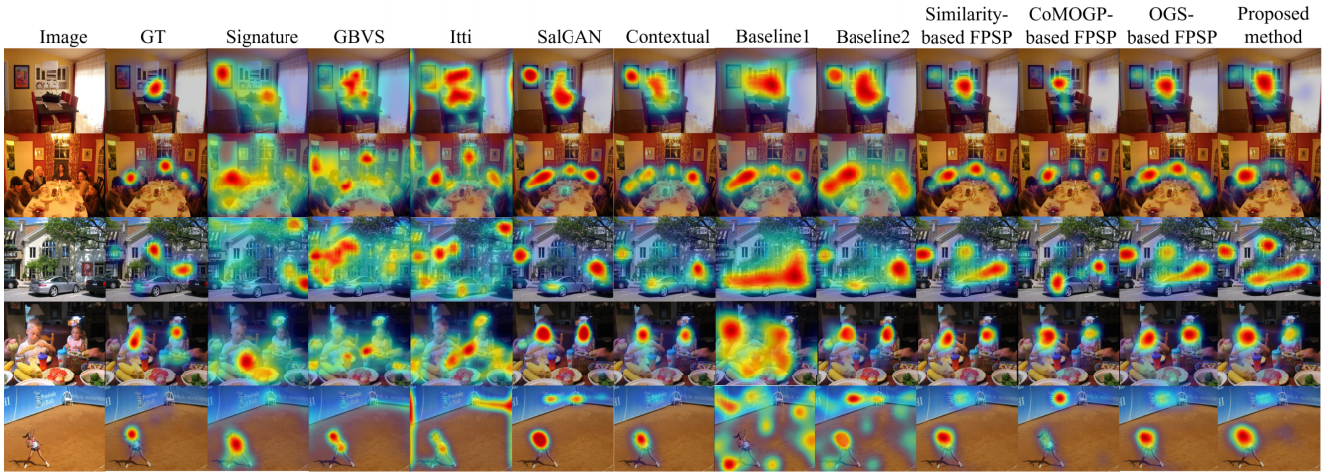

Fig. 3 Examples of predicted PSMs.

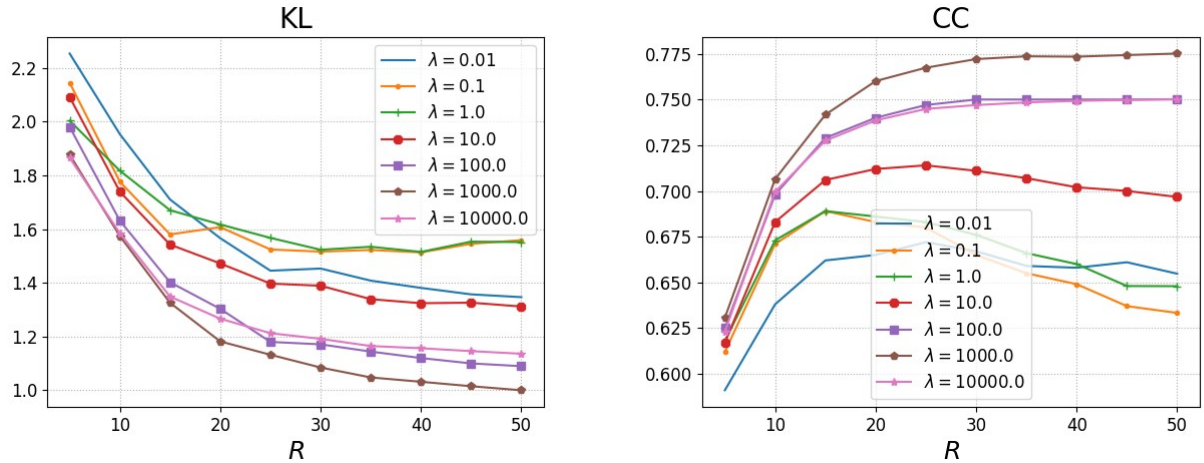

Fig. 4 Changes in the values of the evaluation metrics in response to changes in hyperparameters of tensor-to-matrix regression.

FPSP; however, it is not the best. One possible reason is that the expressive capacity of the model is limited by the rank R in the low-rank tensor approximation of the weight tensor in Sec. 3.3. In general, increasing R enhances the model's ability to capture complex patterns, including fine-grained local information; however, it also substantially increases computational cost. The discussion of hyperparameters in the tensor-to-matrix regression model is described below. Another reason why the "CC" of the proposed method is lower than that of OGS-based FPSP is that the proposed method does not incorporate explicit semantic information such as object categories. In contrast, the OGS-based FPSP utilizes object-level matching to transfer saliency information, which results in higher CC scores. Additionally, the images used in our experiments contain a wide variety of general-purpose images, and many objects in the test images also appear in the training images. This enables the OGS-based FPSP to utilize its object-based mechanism effectively. However, as discussed above, the OGS-based FPSP has limited applicability when the test images include unseen objects. In such scenarios, its object-dependent mechanism cannot be applied. In contrast, the proposed method relies on interpersonal gaze similarity rather than object-specific information, which supports the applicability of the proposed method, particularly in real-world settings where unseen objects appear.

Here, "CC" is a pixel-based evaluation, whereas "KLdiv" is a distribution-based evaluation. Thus, the proposed method can preserve structural information because of its high "KLdiv." Here, as mentioned in Sec. 1, structural information is an important clue to detecting salient areas in the human visual system[1]. From this viewpoint, we confirmed that the proposed method is valid because it is not significantly inferior to OGS-based FPSP in terms of "CC" and superior in terms of "KLdiv." Therefore, we emphasize the effectiveness of

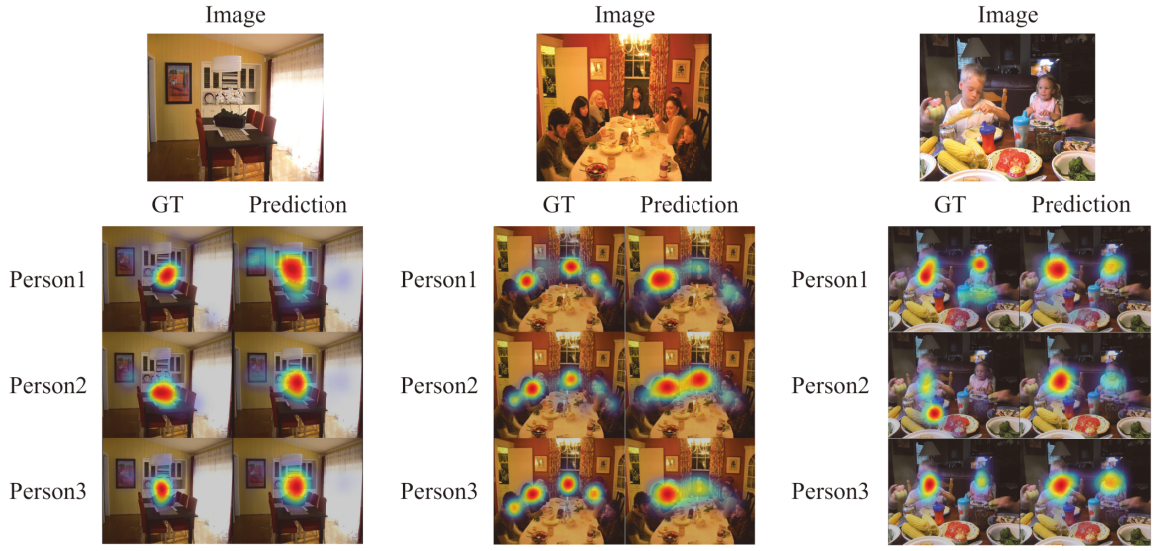

Fig. 5 Examples of PSMs predicted for several persons using the proposed method.

the proposed method for PSM prediction by preserving the structural information.

We confirmed the evaluation scores in response to changes in the hyperparameters of the tensor-to-matrix regression model through additional experiments focusing on hyperparameter analysis. Figure 4 shows the evaluation scores in response to R and λ. As shown in the figure, as R increases, the performance improves, whereas $\lambda = 1000$ achieves the best performance regardless of R. Although these results were obtained experimentally, the trend was stable. Beyond this value, further increases in λ led to performance degradation. Therefore, we adopted $\lambda = 1000$ as a setting corresponding to a local maximum in performance. The results also indicate that the performance improves as R increases, but the improvement saturates around $R = 50$. Beyond this point, further increases in R lead primarily to longer computation times without meaningful gains in prediction performance. Therefore, we adopted $R = 50$ as a practical compromise between model performance and efficiency. In addition, an extremely high λ value is not required because λ is a regularization hyperparameter. Therefore, we confirm the desirable hyperparameters of tensor-to-matrix regression in the proposed method. While this study limits the exploration to R, further investigation with more efficient model designs to reduce computational cost is a promising direction for future work. In addition, we adopted the CP-rank decomposition for the weight tensor in the tensor-to-matrix regression model by referring to a previous study[27]. The validation of other rank decomposition methods is also a future work.

To show the difference in PSMs across individuals, we present the examples of PSMs predicted for several persons using the proposed method in Fig. 5. Most of these examples were correctly predicted by the proposed method. While some predicted PSMs in Fig. 5 did not perfectly align with the actual gaze locations. For example, the main focus might be slightly shifted for certain individuals, such as Person2 in the right column. However, the proposed method still captured the overall attended regions with reasonable performance. It is important to note that human gaze behavior can vary not only due to the visual stimulus but also based on transient factors such as short-term intentions or long-term mental states. Therefore, perfectly predicting individual attention is inherently difficult. Despite this variability, the proposed method was able to approximate the attention tendencies of each person, which demonstrates its effectiveness in modeling personalized saliency.

5. Conclusions

This study has proposed a few-shot PSM prediction method based on interpersonal gaze patterns. The proposed method incorporates the adaptive image selection scheme and tensor-to-matrix regression for effective image selection of images and the preservation of structural information, respectively. By treating the input and output PSMs without vectorization, the proposed method preserves structural information. The experiments on the open dataset demonstrate the effectiveness of incorporating these factors.

Acknowledgement

This work was partly supported by the JSPS KAKENHI Grant Numbers JP24K02942, JP23K21676, and JP23K11211.

References

1) Laurent Itti, Christof Koch, and Ernst Niebur : "A Model of Saliency-based Visual Attention for Rapid Scene Analysis," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254–1259 (1998)
2) Inam Ullah, Muwei Jian, Sumaira Hussain, Jie Guo, Hui Yu, Xing Wang, and Yilong Yin, "A Brief Survey of Visual Saliency Detection," Multimedia Tools and Applications, vol. 79, pp. 34605–34645 (2020)
3) Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li, "Salient Object Detection: A Survey," Computational Visual Media, vol. 5, pp. 117–150 (2019)
4) Jonathan Harel, Christof Koch, and Pietro Perona, "Graph-based Visual Saliency," Advances in Neural Information Processing Systems, pp. 545–552 (2007)
5) Xiaodi Hou, Jonathan Harel, and Christof Koch, "Image Signature: Highlighting Sparse Salient Regions," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 34, no. 1, pp. 194–201 (2012)
6) Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i-Nieto, "Salgan: Visual Saliency Prediction with Generative Adversarial Networks," arXiv preprint arXiv:1701.01081 (2017)
7) Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel, "Contextual Encoder–decoder Network for Visual Saliency Prediction," Neural Networks, vol. 129, pp. 261–270 (2020)
8) Aoqi Li and Zhenzhong Chen, "Personalized Visual Saliency: Individuality Affects Image Perception," IEEE Access, vol. 6, pp. 16099–16109 (2018)
9) Evan F Risko, Nicola C Anderson, Sophie Lanthier, and Alan Kingstone, "Curious Eyes: Individual Differences in Personality Predict Eye Movement Behavior in Scene-viewing," Cognition, vol. 122, no. 1, pp. 86–90 (2012)
10) Olivier Le Meur, Antoine Coutrot, Zhi Liu, Rämä, Adrien Le Roch, and Andrea Helo, "Visual Attention Saccadic Models Learn to Emulate Gaze Patterns from Childhood to Adulthood," IEEE Trans. Image Processing, vol. 26, no. 10, pp. 4777–4789 (2017)
11) Yanyu Xu, Nianyi Li, Junru Wu, Jingyi Yu, and Shenghua Gao, "Beyond Universal Saliency: Personalized Saliency Prediction with Multi-task CNN," in Proc. Int'l Joint Conf. Artificial Intelligence, pp. 3887–3893 (2017)
12) Sikun Lin and Pan Hui, "Where's Your Focus: Personalized Attention," arXiv preprint arXiv:1802.07931 (2018)
13) Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Few-shot Personalized Saliency Prediction Using Person Similarity Based on Collaborative Multi-output Gaussian Process Regression," in Proc. IEEE Int'l Conf. Image Processing, pp. 1469–1473 (2021)
14) Shi Chen, Nachiappan Valliappan, Shaolei Shen, Xinyu Ye, Kai Kohlhoff, and Junfeng He, "Learning from Unique Perspectives: User-aware Saliency Modeling," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, pp. 2701–2710 (2023)
15) Qiong Wang, Meriem Outtas, Julie Fournier, Elise Etchamendy, Myriam Chérel, and Lu Zhang, "Predicting Personalized Saliency Map for People with Autism Spectrum Disorder," in Proc. Int'l Conf. Content-based Multimedia Indexing, pp. 34–40 (2023)
16) Yanyu Xu, Shenghua Gao, Junru Wu, Nianyi Li, and Jingyi Yu, "Personalized Saliency and its Prediction," IEEE Trans. Pattern Analysis and Machine Intelligence, pp. 2975–2989 (2018)
17) Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool, "The Interestingness of Images," in Proc. IEEE Int'l Conf. Computer Vision, pp. 1633–1640 (2013)
18) Yixuan Li, Pingmei Xu, Dmitry Lagun, and Vidhya Navalpakkam, "Towards Measuring and Inferring User Interest from Gaze," in Proc. Int'l Conf. World Wide Web Companion, pp. 525–533 (2017)
19) Camilo Fosco, Vincent Casser, Amish Kumar Bedi, Peter O'Donovan, Aaron Hertzmann, and Zoya Bylinskii, "Predicting Visual Importance Across Graphic Design Types," in Proc. Annual ACM Symposium on User Interface Software and Technology (UIST), pp. 249–260 (2020)
20) Quanlong Zheng, Jianbo Jiao, Ying Cao, and Rynson W.H. Lau, "Task-driven Webpage Saliency," in Proc. European Conference on Computer Vision (ECCV), pp. 287-302 (2018)
21) Yoon Ho Cho, Jae Kyeong Kim, and Soung Hie Kim, "A Personalized Recommender System Based on Web Usage Mining and Decision Tree Induction," Expert Systems with Applications, vol. 23, no. 3, pp. 329-342 (2002)
22) Mohamed Nader Jelassi, Sadok Ben Yahia, and Engelbert Mephu Nguifo, "A Personalized Recommender System Based on Users' Information in Folksonomies," in Proc. Int'l Conf. World Wide Web (WWW), pp. 1215-1224 (2013)
23) Jun Nakamura, and SanetakeNagayoshi, "The Pottery Skills and Tacit Knowledge of a Maser: An Analysis Using Eye-tracking Data," Procedia Computer Science, vol. 159, pp. 1680-1687 (2019)
24) Weiwei Yu, Dian Jin, Wenfeng Cai, Feng Zhao, and Xiaokun Zhang, "Towards Tacit Knowledge Mining within Context: Visual Cognitive Graph Model and Eye Movement Image Interpretation," Computer Methods and Programs in Biomedicine, vol. 226, pp. 107107 (2022)
25) 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, vol. 20, no. 8: 2170, pp. 1-15(2020)
26) Trung V Nguyen, Edwin V Bonilla, "Collaborative Multi-output Gaussian Processes," in Proc. Association for Uncertainty in Artificial Intelligence, pp. 643–652 (2014)
27) Eric F Lock, "Tensor-on-tensor Regression," Journal of Computational and Graphical Statistics, vol. 27, no. 3, pp. 638–647 (2018)
28) Xi Yin and Xiaoming Liu, "Multi-task Convolutional Neural Network for Pose-invariant Face Recognition," IEEE Trans. Image Processing, vol. 27, no. 2, pp. 964–975 (2017)
29) Rezvan Sherkati and James J Clark, "Clustered Saliency Prediction," arXiv preprint arXiv:2207.02205 (2022)
30) Florian Strohm, Mihai Bâce, and Andreas Bulling, "Learning User Embeddings from Human Gaze for Personalized Saliency Prediction," in Proc. ACM on Human-Computer Interaction, vol. 8, pp. 1–16 (2024)
31) Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Few-shot Personalized Saliency Prediction with Similarity of Gaze Tendency Using Object-based Structural Information," in Proc. IEEE Int'l Conf. Image Processing, pp. 3823–3827 (2022)
32) Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun, "Yolox: Exceeding Yolo Series in 2021," arXiv preprint arXiv:2107.08430 (2021)
33) Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba, "Learning to Predict Where Humans Look," in Proc. IEEE Int'l Conf. Computer Vision, pp. 2106–2113 (2009)
34) Léon Bottou, "Large-scale Machine Learning with Stochastic Gradient Descent," in Proc. Int'l Conf. Computational Statistics, pp. 177–186 (2010)
35) Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao, "Salicon: Saliency in Context," in Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, pp. 1072–1080 (2015)
36) Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "User-centric Visual Attention Estimation Based on Relationship between Image and Eye Gaze Data," in Proc. Global Conf. Consumer Electronics, pp. 44–45 (2018)
37) Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "User-specific Visual Attention Estimation Based on Visual Similarity and Spatial Information in Images," in Proc. IEEE Int'l Conf. Consumer Electronics-Taiwan, pp. 479–480 (2019)
38) Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand, "What Do Different Evaluation Metrics Tell Us about Saliency Models?," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 41, no. 3, pp. 740–757 (2018)

Yuya Moroto received his B.S. degree in Electronics and Information Engineering from Hokkaido University, Japan in 2019, and his M.S. and Ph.D. degrees in Information Science and Technology from Hokkaido University, Japan in 2021 and 2024. His research interests include computer vision, biological information analysis, multimodal signal processing and its applications. He is a member of IEEE.

Keisuke Maeda received his B.S., M.S., and Ph.D. degrees in Electronics and Information Engineering from Hokkaido University, Japan in 2015, 2017, and 2019. At present, he is currently a Specially Appointed Associate Professor in the Data-Driven Interdisciplinary Research Emergence Department, Hokkaido University. His research interests include multimodal signal processing and machine learning and its applications. He is an IEEE and IEICE member.

Takahiro Ogawa received his B.S., M.S. and Ph.D. degrees in Electronics and Information Engineering from Hokkaido University, Japan in 2003, 2005 and 2007, respectively. He joined Graduate School of Information Science and Technology, Hokkaido University in 2008. He is currently a professor in the Faculty of Information Science and Technology, Hokkaido University. His research interests are AI, IoT and big data analysis for multimedia signal processing and its applications. He was a special session chair of IEEE ISCE2009, a Doctoral Symposium Chair of ACM ICMR2018, an organized session chair of IEEE GCCE2017-2019, a TPC Vice Chair of IEEE GCCE2018, a Conference Chair of IEEE GCCE2019, etc. He has been also an Associate Editor of ITE Transactions on Media Technology and Applications. He is a senior member of IEEE and a member of ACM, IEICE and ITE.

Miki Haseyama received her B.S., M.S. and Ph.D. degrees in Electronics from Hokkaido University, Japan in 1986, 1988 and 1993, respectively. She joined the Graduate School of Information Science and Technology, Hokkaido University as an associate professor in 1994. She was a visiting associate professor of Washington University, USA from 1995 to 1996. She is currently a professor in the Faculty of Information Science and Technology Division of Media and Network Technologies, Hokkaido University. Her research interests include image and video processing and its development into semantic analysis. She has been a Vice-President of the Institute of Image Information and Television Engineers, Japan (ITE), an Editor-in-Chief of ITE Transactions on Media Technology and Applications, a Director, International Coordination and Publicity of The Institute of Electronics, Information and Communication Engineers (IEICE). She is a member of the IEEE, IEICE, Institute of Image Information and Television Engineers (ITE) and Acoustical Society of Japan (ASJ).