Predicting Visual Importance of Mobile UI Using Semantic Segmentation

Ami Yamamoto, Yuichi Sei[a], Yasuyuki Tahara[b] and Akihiko Ohsuga[c]
Graduate School of Informatics and Engineering, The University of Electro-Communinacions, Tokyo, Japan

Keywords: Deep Learning, Visual Importance, Mobile Interface, User Interface for Design.

Abstract: When designing a UI, it is necessary to understand what elements are perceived to be important to users. The UI design process involves iteratively improving the UI based on feedback and eye-tracking results on the UI created by the designer, but this iterative process is time-consuming and costly. To solve this problem, several studies have been conducted to predict the visual importance of various designs. However, no studies specifically focus on predicting the visual importance of mobile UI. Therefore, we propose a method to predict visual importance maps from mobile UI screenshot images and semantic segmentation images of UI elements using deep learning. The predicted visual importance maps were objectively evaluated and found to be higher than the baseline. By combining the features of the semantic segmentation images appropriately, the predicted map became smoother and more similar to the ground truth.

1 INTRODUCTION

In recent years, mobile terminals, as typified by smartphones, have spread rapidly, and the rate of Internet usage via mobile terminals has also increased. In tandem with this growth, the types of applications available on mobile devices and the number of downloads continue to increase, and consumer activities centered on lifestyle and entertainment, such as shopping, payment, and video streaming, are also expanding. As a result, mobile application developers and designers need to understand the elements that will engage users to develop more usable applications. Designers make improvements based on feedback and eye tracking results on the UI they create. However, these methods require research for each design, which is time-consuming and costly.

In this study, we focused on visual importance, rather than visual saliency, which has been widely studied, as a metric for quantitatively evaluating design. Visual saliency is estimated from actual eye gaze information obtained by eye tracking, whereas visual importance data is created by mapping the areas that users perceive as important when they look at a design, regardless of their gaze. Therefore, visual importance is strongly related to semantic categories such as text and images, as well as position and hue (Bylinskii et al., 2017).

Since mobile terminal screens are smaller than PC screens, the number of objects that can be displayed on a single screen is smaller, and visual saliency manifests itself differently on mobile terminals than on PCs (Leiva et al., 2020). Therefore, it is highly likely that the visual importance of mobile devices also tends to be different from that of PCs, which makes it significant to conduct visual importance forecasting specific to mobile UI. In addition, new design patterns and interface elements are frequently introduced in recent mobile application platforms. Furthermore, features such as hover status in PC interfaces are not applicable in mobile UIs (Swearngin and Li, 2019). Since mobile applications are finger-operated, more emphasis is placed on visual importance as a quality characteristic. In addition, since UI is composed of various design elements such as text, images, and buttons, it is reasonable to use visual importance in the optimization of UI design and feedback tools. Accurately predicting the visual importance of a mobile UI can provide designers with real-time feedback and design optimization.

We propose a method for predicting visual importance maps from mobile UI screenshot images and semantic segmentation images of UI elements using deep learning. We investigated three different feature combination methods and evaluated the predicted visual importance maps objectively and subjectively.

[a] https://orcid.org/0000-0002-2552-6717
[b] https://orcid.org/0000-0002-1939-4455
[c] https://orcid.org/0000-0001-6717-7028

Yamamoto, A., Sei, Y., Tahara, Y. and Ohsuga, A.
Predicting Visual Importance of Mobile UI Using Semantic Segmentation.
DOI: 10.5220/0011655800003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 3*, pages 260-266
ISBN: 978-989-758-623-1; ISSN: 2184-433X

The proposed method was rated higher than the baseline.

This paper is organized as follows. Section 2 describes related studies, Section 3 explains the proposed method, and Section 4 presents experimental results and evaluates the proposed method. The results are discussed in Section 5. Finally, Section 6 concludes the paper and presents future prospects.

2 RELATED WORK

Several studies have been conducted on saliency prediction in mobile UI, and Gupta et al. proposed a method to predict saliency for each UI element, focusing on the fact that designers add, remove, and edit elements for each interface component in mobile UI design (Gupta et al., 2018). By using not only UI images but also images at different scales as inputs to the model, local and global features are combined to predict saliency. Leiva et al. created a dataset of saliency in mobile UI and developed a saliency prediction model, SAM (SAM Saliency Attentive Model), and developed a saliency prediction model specific to mobile UI (Leiva et al., 2020). They also conducted a statistical investigation of saliency in mobile UI and showed a strong bias toward the upper left of the screen, text, and images.

Several models have been developed to predict the visual importance of design, and Bylinskii et al. proposed a method to predict the visual importance of graphic design and data visualization using deep learning (Bylinskii et al., 2017). Fosco et al. proposed the Unified Model of Saliency and Importance (UMSI), an integrated model that predicts the visual importance of five design classes (web page, movie poster, mobile UI, infographics, and advertisement) and saliency in natural images (Fosco et al., 2020). UMSI is a deep learning model that automatically classifies classes of input images before predicting their visual saliency and importance.

However, there are no prediction methods specific to the visual importance of mobile UI, and no studies have yet taken into account mobile UI-specific factors such as the placement and categories of UI elements. As mentioned earlier, since visual importance maps have a strong association with UI element categories such as buttons, images, and text, we hypothesized that semantic segmentation images representing UI element placement and categories could improve visual importance prediction performance. Therefore, we propose a method to predict visual importance maps from mobile UI screenshot images and semantic segmentation images of UI elements using deep learning.

3 APPROACH

In this study, a visual importance prediction model was built based on MSI-Net (Kroner et al., 2020), a natural image saliency prediction model, utilizing semantic segmentation images that represent the categories and locations of UI elements. Figure 1 shows the overall diagram of the prediction model.

3.1 Model Architecture

3.1.1 MSI-Net

In the architecture of this model, the UI encoder, ASPP module, and decoder are based on MSI-Net (Kroner et al., 2020), a saliency prediction model for natural images. MSI-Net takes an encoder-decoder structure and incorporates the Atrous Spatial Pyramid Pooling (ASPP) module (Chen et al., 2018) with multiple convolution layers with different expansion rates to extract multi-scale features. ASPP module can estimate the saliency of the entire image, and quantitative and qualitative performance improvements have been reported. The encoder is also based on VGG-16, which removes the stride of the two pooling layers in the second half of the encoder, allowing for a spatial representation of 1/8 of the original input size. This reduces the downscaling effect and allows for higher feature extraction performance. Since the number of trainable parameters for this modified model is the same as for VGG-16, the model can be initialized with weights previously learned in ImageNet (Deng et al., 2009). Since the visual importance dataset is small, it must be pre-trained on the saliency dataset of natural images. We considered this adjustment to be effective for efficient pre-training.

3.1.2 Semantic Segmentation Encoder

In this study, a semantic segmentation encoder that represents the categories and positions of UI elements was incorporated into MSI-Net to build the model. The outputs of the UI encoder and the semantic segmentation encoder were concatenated and used as input to the ASPP module. Semantic segmentation images contain less detail than UI images and can be processed at lower resolutions. Therefore, the semantic segmentation encoder uses half the input image size and fewer convolution layers of the UI encoder.

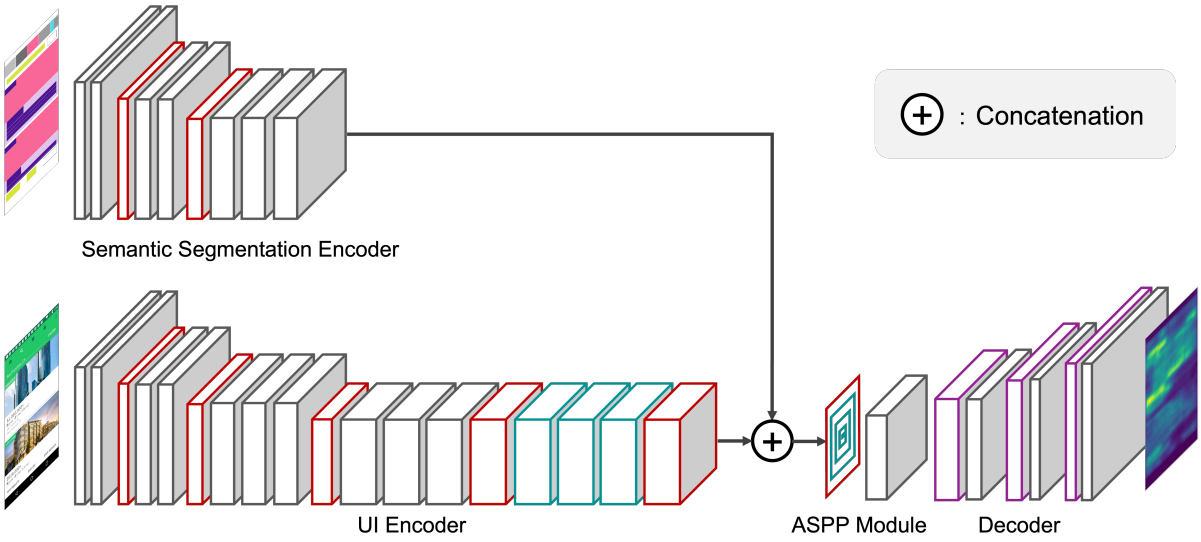

Figure 1: Visual importance prediction model.

3.1.3 Feature Concatenation

In MSI-Net, the outputs of the 10, 14, and 18 layers of encoders are concatenated and used as inputs to the ASPP module to take advantage of the features of the different levels of convolution layers. In our model, the output of the UI encoder, which is the input of the ASPP module, is varied from layers 18 only, 14,18, and 10,14,18 to adjust the feature ratio of the UI and semantic segmentation elements. The dimensions of the features input to the ASPP module are shown in Table 1. In the case of layer 18 only, the features of UI elements are smaller, so the method emphasizes semantic segmentation relatively more. 10, 14, and 18 layers are methods that emphasize the UI itself because the features of UI elements are larger. 14 and 18 layers are in between these two methods, emphasizing the balance between UI elements and semantic segmentation.

Table 1: Dimensions of features to be input to the ASPP module.

Output layer	UI	Segmentation	post-concat
18	512	256	768
14,18	1024	256	1280
10,14,18	1280	256	1536

4 EXPERIMENTS AND RESULTS

In the experiment, MSI-Net without semantic segmentation is used as the baseline model and compared to three proposed methods with different feature dimensions.

4.1 Dataset

For pre-training, we used 10,000 natural images and semantic segmentation training sets and 5,000 test sets from SALICON (Jiang et al., 2015) and MS COCO, using the weights learned in ImageNet as initial values. For the subsequent fine-tuning, we used Imp1k (Fosco et al., 2020) mobile UI data, 160 images from the semantic segmentation training set published on Rico (Deka et al., 2017), and 40 images from the test set. MSI-Net was pre-trained on natural images and fine-tuned on the UI data, without semantic segmentation.

imp1k is a dataset annotated with visual importance in five design classes: web pages, movie posters, mobile UI, infographics, and advertisements. For mobile UI, 200 screenshots were randomly sampled from the Rico dataset and annotated using mouse strokes. The design structure of mobile UI and web pages differs significantly from other design structures and cannot be generalized by models trained to predict the importance of advertisements and posters (Fosco et al., 2020). Therefore, we used 200 pieces of data on mobile UI in the imp1k dataset for this study.

SALICON is a large dataset annotated with the saliency of natural images and is often used to train saliency prediction models. imp1k dataset has a small number of mobile UI data, so we aim to improve model performance by pre-training on the saliency dataset.

Rico contains not only UI screenshot images, but also semantic segmentation related to the meaning

and usage of elements on UI screens (Deka et al., 2017). Therefore, we trained the Rico dataset using only the semantic segmentation images corresponding to the UI contained in imp1k.

4.2 Experimental Settings

In this study, we used KL divergence as the loss function. KL divergence is suitable for models that aim at detecting salient targets because it provides a large penalty for missed predictions. In fine-tuning, the batch size was set to 4, the learning rate to 1e-4, and Adam was used as the optimization function. Regarding the input image size, the natural images are horizontal and the mobile UI images are vertical. Therefore, the natural images were resized to 240×160 for pre-training, and the size was adjusted to 240×160 by adding a margin next to the mobile UI image for subsequent fine-tuning. The same procedure was used for semantic segmentation, and the input image size was set to 120×80.

4.3 Evaluation Metrics and Results

Various metrics have been used to evaluate the performance of predictive models for saliency and visual importance maps. In this study, four indices used in previous studies of visual importance, R^2, CC, $RMSE$, and KL, were used for evaluation (Bylinskii et al., 2017) (Fosco et al., 2020). R^2 is coefficient of determination, CC is correlation coefficient, $RMSE$ is root mean square error, and KL is Kullback-Leibler divergence.

R^2 measures the fit between the estimated map and the ground truth map. It is calculated based on the variability of the data itself and the discrepancy between the predictions. The best fit is 1, and the closer to 1, the better the performance of the model. Given the grand-truth importance map Q and the predicted importance map P, R2 is computed as:

$$R^2(P,Q) = \frac{\sum_{i=1}^{N}(Q_i - P_i)^2}{\sum_{i=1}^{N}(Q_i - \overline{Q})^2} \tag{1}$$

where $\overline{Q} = \frac{1}{N}\sum_{i=1}^{N} Q_i$.

CC means the correlation between the estimated map and the ground truth map. The closer CC is to 1, the stronger the positive correlation, and the closer to 0, the weaker the correlation. CC is computed as:

$$CC(P,Q) = \frac{\sum_{i=1}^{N}(P_i - \overline{P})(Q_i - \overline{Q})}{\sqrt{\sum_{i=1}^{N}(P_i - \overline{P})^2}\sqrt{\sum_{i=1}^{N}(Q_i - \overline{Q})^2}} \tag{2}$$

where $\overline{P} = \frac{1}{N}\sum_{i=1}^{N} P_i$.

$RMSE$ is calculated from the square of the error between the estimated map and the ground truth map. The closer to 0, the higher the prediction accuracy. Because the square is used, the indicator is sensitive to outliers, with a lower rating if the prediction is far off. $RMSE$ is computed as:

$$RMSE(P,Q) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Q_i - P_i)^2} \tag{3}$$

The importance map can be interpreted as representing for each pixel the probability that the pixel is considered visually important. KL is a measure of the distance between the predicted distribution and the ground truth, and represents how closely the probability distribution P approximates the probability distribution Q. A better approximation of the two maps results in a smaller KL, and a KL of 0 indicates that the maps are identical. KL is computed as:

$$KL(P,Q) = \sum_{i=1}^{N}(Q_i \log Q_i - Q_i \log P_i) = L(P,Q) - H(Q) \tag{4}$$

where $H(Q) = -\sum_{i=1}^{N}(Q_i \log Q_i)$ is the entropy of the ground truth importance map and $L(P,Q)$ is the cross entropy of the prediction and ground truth.

The evaluation results are shown in Table 2. The numbers in parentheses indicate the layer of the UI encoder that concatenates the outputs.

Table 2: Performance of visual importance prediction models for mobile UI.

	$R^2 \uparrow$	$CC \uparrow$	$RMSE \downarrow$	$KL \downarrow$
MSI-Net	0.505	0.841	0.102	**0.151**
Ours(10,14,18)	0.548	0.844	0.0959	0.153
Ours(14,18)	**0.639**	**0.845**	**0.0883**	**0.151**
Ours(18)	0.631	0.835	0.0923	0.163

Examples of visual importance maps predicted by the baseline and proposed methods are shown in Figure 2. Figure 3 shows an example where the proposed method did not predict well. The more yellow the pixel is, the higher the visual importance of the area and the bluer the pixel is, the lower the visual importance of the area.

5 DISCUSSION

Table 2 shows that Ours(14,18) was equal to or better than the baseline on all the evaluation indices. Both Ours(10,14,18) and Ours(18) also outperformed the baseline on the R^2 and $RMSE$ metrics, but both were slightly worse than the baseline on the KL metric. Thus, a model that balances image features and

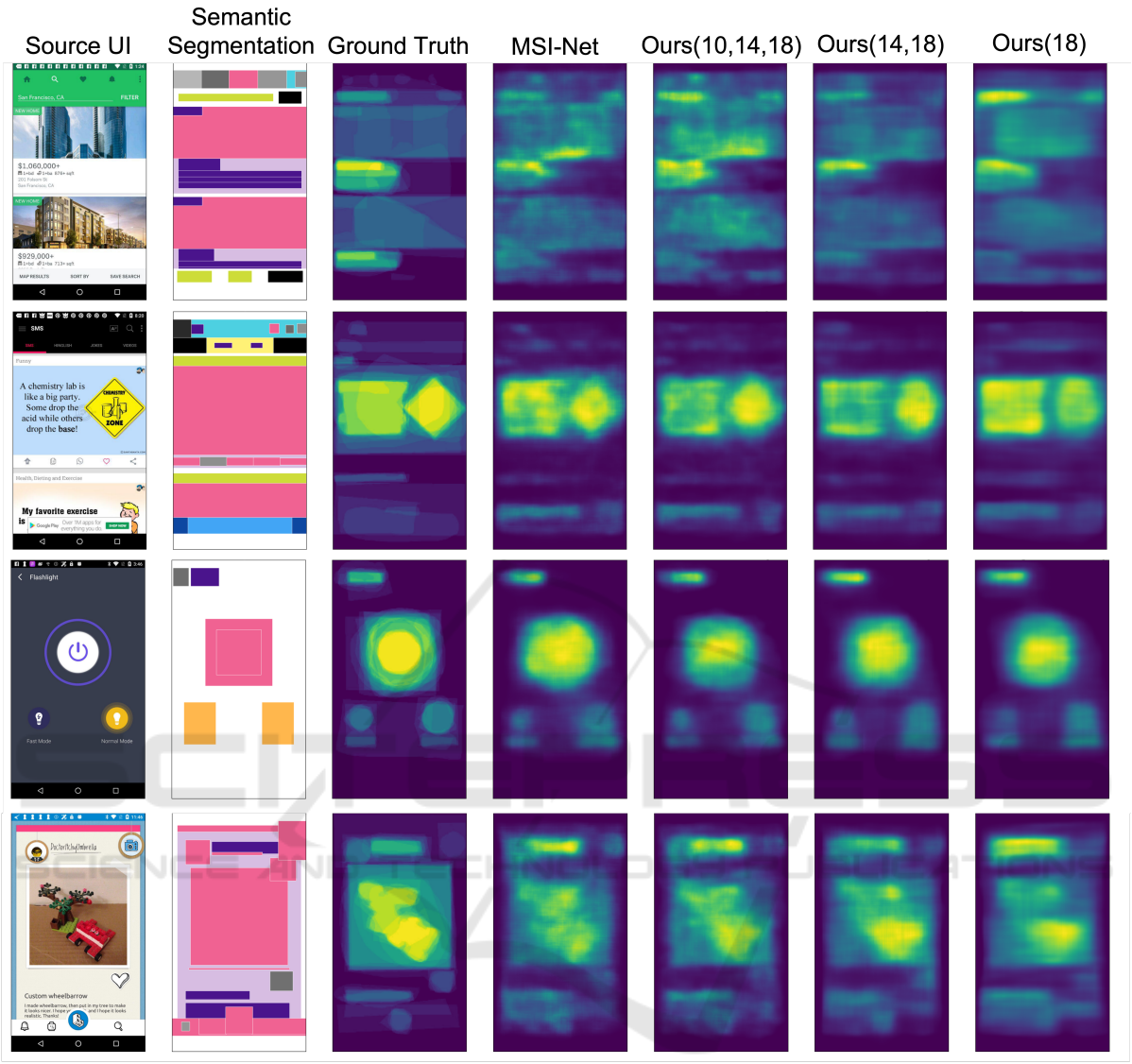

Figure 2: Example of a projected visual importance map.

semantic segmentation features is most suitable for predicting the visual importance of UI images, and the appropriate use of semantic segmentation element features improves the prediction performance of the visual importance map.

To analyze the differences between the mobile UI and the other images, Table 3 shows the results when the prediction model was pre-trained for saliency on natural images. Table 3 shows that Ours(10,14,18) performed best in terms of saliency for natural images, while Ours(14,18), which was superior for mobile UI, performed slightly worse. Even for natural images, the use of semantic segmentation contributes to performance improvement. However, the features of the image itself play a more important role in the saliency of natural images than semantic segmentation. Figure 2 shows that the ground truth of visual importance in mobile UI tends to be distributed across the UI element parts. Comparing Table 2 and Table 3, the difference between MSI-Net and Ours is larger in Table 2, which represents the mobile UI results. Therefore, the benefits of semantic segmentation are particularly large for mobile UI, and our method was effective.

Figure 2 shows that the predicted map is smooth for Ours(18), but it fails to capture image shapes such as rhombuses. In Ours(10,14,18), the image shapes are captured, but the features of the semantic segmentation elements are too small compared to the features of the UI elements, and the results are almost

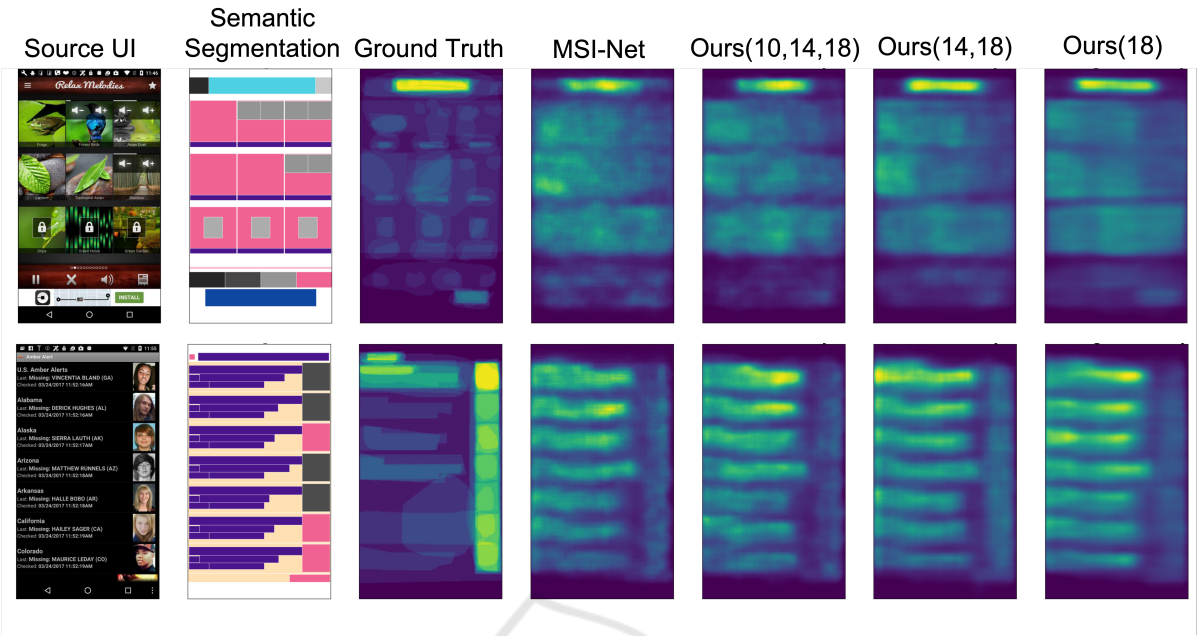

Figure 3: Example of failure to predict well.

Table 3: Performance of saliency prediction models for natural images.

	$R^2 \uparrow$	$CC \uparrow$	$RMSE \downarrow$	$KL \downarrow$
MSI-Net	0.521	0.880	0.113	0.224
Ours(10,14,18)	**0.569**	**0.884**	**0.107**	**0.219**
Ours(14,18)	0.497	0.881	0.116	0.222
Ours(18)	0.482	0.881	0.117	0.226

the same as in MSI-Net. However, Ours(14,18) is able to predict smooth importance maps while preserving image features and using semantic segmentation elements. The results show that using appropriate semantic segmentation features improves the prediction performance of visual importance maps for mobile UIs.

Figure 3 shows an example where the visual importance map could not be predicted well using semantic segmentation. This UI is tiled with images, but the visual importance map in ground truth is based on the features of each image, not the structure of the UI. Since the images in semantic segmentation represent only the structure and categories of the UI, we found that the proposed method does not work well for UIs with strong image features in the visual importance map. As in this example, it is necessary to build a model that is more robust to image features in order to deal with a mobile UI that has a complex UI structure and more prominent image features. However, such a model may have low prediction accuracy for simple mobile UIs.

6 CONCLUSION AND FUTURE WORK

In this study, we proposed a visual importance prediction method that takes UI elements into account with the aim of accurately predicting the visual importance of mobile UI. The evaluation compared the proposed method with a baseline method that does not use semantic segmentation of UI elements. The visual importance map predicted by Ours(14,18) was smoother and closer to ground truth than the baseline. We also adapted the proposed method to natural images to see if semantic segmentation works differently for mobile UI and natural images. The use of semantic segmentation was also effective for natural images, but its effect was weaker than for mobile UI. We found that a balanced use of semantic segmentation features improves the accuracy of predicting the visual importance of the mobile UI.

Since changes are made to UI elements such as buttons and images, rather than to pixels, when developing UI, future research should examine the visual importance of each UI element. In addition, since the experiments in this paper were conducted using only the UI data included in the imp1k dataset, we would like to verify whether visual importance can be predicted in the same way for UIs in different languages. We would also like to apply the visual importance prediction model proposed in this study to optimize mobile UI design and to provide feedback

tools for designers. Specifically, we are considering using our predictive model as an objective function to optimize the color scheme of buttons and text in mobile UI using a genetic algorithm. Optimization allows developers to easily create a UI with the increased importance of the UI components they want to make stand out. For optimization, we would like to perform predictions for novel and unique UIs that are not included in existing datasets and conduct user experiments to see if the predicted visual importance maps are appropriate.

For a better user experience, it is also necessary to analyze where users direct their attention when they see a UI and whether they can understand that UI correctly. There is already related research on how users understand UI, such as icon annotation in mobile UI (Zang et al., 2021) and predicting mobile UI tappability (Swearngin and Li, 2019). We believe that combining these related work with our visual importance predictions will provide more useful feedback to designers.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Numbers JP21H03496, JP22K12157.

REFERENCES

Bylinskii, Z., Kim, N. W., O'Donovan, P., Alsheikh, S., Madan, S., Pfister, H., Durand, F., Russell, B., and Hertzmann, A. (2017). Learning visual importance for graphic designs and data visualizations. *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, pages 57–69.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence 40(4)*, pages 834–848.

Deka, B., Huang, Z., Franzen, C., Hibschman, J., Afergan, D., Li, Y., Nichols, J., and Kumar, R. (2017). Rico: A mobile app dataset for building data-driven design applications. *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, pages 845–854.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *IEEE conference on computer vision and pattern recognition*, pages 248–255.

Fosco, C., Casser, V., Bedi, A. K., O'Donovan, P., Hertzmann, A., and Bylinskii, Z. (2020). Predicting visual importance across graphic design types. *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 249–260.

Gupta, P., Gupta, S., Jayagopal, A., Pal, S., and Sinha, R. (2018). Saliency prediction for mobile user interfaces. *2018 IEEE Winter Conference on Applications of Computer Vision*, pages 1529–1538.

Jiang, M., Huang, S., Duan, J., and Zhao, Q. (2015). Salicon: Saliency in context. *IEEE conference on computer vision and pattern recognition*, pages 1072–1080.

Kroner, A., Senden, M., Driessens, K., and Goebel, R. (2020). Contextual encoder–decoder network for visual saliency prediction. *Neural Networks 129*, pages 261–270.

Leiva, L. A., Xue, Y., Bansal, A., Tavakoli, H. R., Köroğlu, T., Du, J., Dayama, N. R., and Oulasvirta, A. (2020). Understanding visual saliency in mobile user interfaces. *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–12.

Swearngin, A. and Li, Y. (2019). Modeling mobile interface tappability using crowdsourcing and deep learning. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11.

Zang, X., Xu, Y., and Chen, J. (2021). Multimodal icon annotation for mobile applications. *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*, pages 1–11.