Translating Adult's Focus of Attention to Elderly's

Onkar Krishna[1] Go Irie[1] Takahito Kawanishi[1] Kunio Kashino[1] Kiyoharu Aizawa[2]

[1] NTT Corporation [2] The University of Tokyo

***Abstract*—Predicting which part of a scene elderly people would pay attention to could be useful in assisting their daily activities, such as driving, walking, and searching. Many computational models for predicting focus of attention (FoA) have been developed. However, most of them focus on mimicking adult FoA and do not work well for predicting elderly's, due to age-related changes in human vision. Is it possible to leverage the prediction results made by an FoA model of general adults to accurately predict elderly's FoA, rather than training a new network from scratch? In this paper, we consider a novel problem of translating adult's FoA to elderly's and propose an approach based on deep image-to-image translation. Our model is trained by minimizing both Kullback-Leibler divergence and adversarial loss to approximate the joint probability distribution of adult and elderly FoA. Experiments on two datasets demonstrate that our model gives remarkable prediction accuracy.**

I. Introduction

The world's population is ageing. For instance, in Japan, which is considered to have the highest percentage of elderly, seniors accounted for 25% of the total population in 2015, and this trend is likely to continue; It is expected to reach 30% of the population by 2025 and nearly 40% by 2055. Given the scale and the trend of the situation, there will be a great demand for AI systems that can monitor and support the daily activities of the elderly. Our work in this paper aims to develop a computer vision system that predicts the focus of attention (FoA) of the elderly to support various daily activities such as driving, walking, and searching.

Computational models of saliency [1] are frequently used to predict FoAs when people are watching the scene. Significant efforts have been made on developing effective saliency models. The early attempts including the seminal work by Itti and Koch [1] developed bottom-up models of saliency based on hand-crafted image features. Considering the difficulty of designing effective hand-crafted features that can capture a variety of human FoAs, modern approaches rely on deep learning to obtain an end-to-end mapping from the input image to the FoA map. For example, [2] used a Convolutional Neural Network (CNN) to learn effective features from data and to make FoA predictions with the help of prior information. [3] proposed a model based on a 3D convolutional neural network called C3D to model FoAs in car driving scenarios. [4] proposed to use convolutional Long-Short Term Memory (LSTM) network to iteratively refine the predicted FoA maps. [5] also proposed a sequential CNN-LSTM model for video saliency estimation. However, one drawback of existing models are focused on and evaluated for adult's FoA, hence cannot work well for the prediction of elderly's. Indeed, it has been claimed that aging process adversely affects the saccadic eye-movement functioning, which results in crucial changes in scene viewing behaviors [6]. Furthermore, some studies [7], [8] suggest that human FoA significantly changes with aging.

Motivated by these observations, in this paper, we propose a FoA prediction model targeted to elderly. A straightforward approach would be to train a prediction model, e.g., CNN, on the eye-gaze data of elderly participants from scratch. However, collecting a sufficient amount of training data of elderly's FoAs to train a brand new deep CNN is challenging due to their physical or health conditions.

We consider a new framework to provide a data-efficient approach for training a model of elderly's FoAs. Although the FoAs by adults and elderly generally look quite different as we will show later in Fig. 2, their tendencies can still be well characterized by the scene they watch. Aiming at leveraging the correlations between the adults' and elderly's FoAs, we propose a deep image-to-image translation approach. Given a scene, our method translates the FoA map of adults predicted by a state-of-the-art method for the adults, e.g., [3], [2], to that of the elderly's. Our translation model is obtained by an encoder-decoder-type deep CNN. The training of our model is performed to minimize the Kullback-Leibler (KL) divergence between the ground truth and predicted FoA maps of elderly. Moreover, we aim at simultaneously minimizing the adversarial loss so that the model can capture the underlying correlations between the FoAs of adults and elderly for the scene in the form of joint probability distribution. Although a few recent attempts consider age-dependent saliency models [9], [10], this is the first work that introduces the image-to-image translation framework to the age-dependent FoA prediction task, to the best of our knowledge. We construct two datasets covering both of adults' and elderly's FoA maps in two different scenarios, i.e., task-based viewing while driving on a car driving simulator and free-viewing while walking on a crowded street. Evaluation experiments on both datasets show that our model gives remarkable prediction accuracy for elderly's FoA.

II. Methods

First of all, we briefly explain our problem considered in this paper. Given a sequence of video frames, our task is to predict which part of each frame an elderly person would pay attention to while driving a car or walking on a street. More concretely, given a sequence of k successive frames denoted by $\mathcal{F}_n = \{f_{n-k+1}, f_2 ..., f_n\}$, our task is to predict the FoA on f_n. The FoA is predicted in the form of the probability of how likely each pixel of f_n is attended by an observer. We denote the predicted FoA of elderly and adults for f_n by e_n and

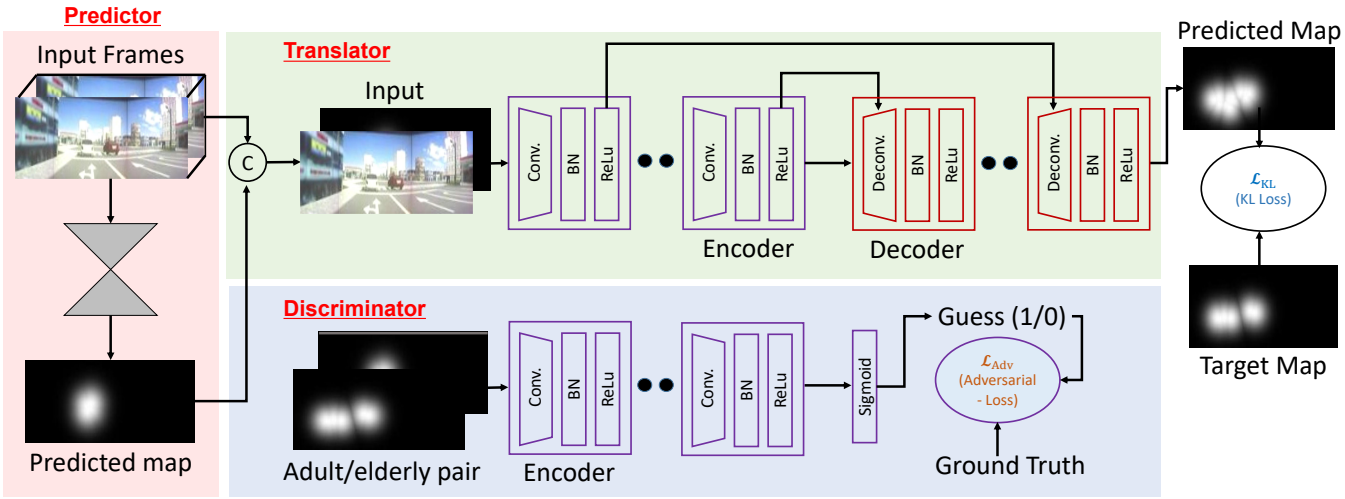

Fig. 1. **Overview of our proposed network**. The translation framework mainly consist of two major networks; translator and discriminator. Translator receives adult's FoA concatenated with the corresponding RGB frame as a input, where the adult's FoA is predicted by the predictor network which is a state-of-the-art method of adult's FoA prediction.

a_n, respectively. The ground truth FoA of elderly observers is denoted by e_n^*.

We approach to the task in two steps; we first predict the adult's FoA a_n, and then transform it to the elderly's FoA e_n. The schematic overview of our model is illustrated in Fig. 1. In the first step, given an input sequence of a video, the adult's FoA a_n is predicted by using a *predictor network* which is trained on a large fixation data of adult observers. While there are a lot of existing methods that can predict the adult's FoAs, we in this work specifically use the state-of-the-art models, namely [3] or [2], depending on evaluation scenarios we will describe later in Sec. IV. In the second step, the *translator network* is used to get the elderly's FoA map e_n from both of the current video frame f_n and the predicted adult's FoA a_n by the predictor network. We use another network called *discriminator network* only for training to facilitate the training of the translator network in an adversarial learning framework. The task of the discriminator is to judge if the input, i.e., e_n or e_n^*, is the one produced by the translator network ("fake") or not ("real").

Hereafter, we first give the details of our translator and discriminator networks and then describe our training procedure.

A. Model Configurations

Our translator and discriminator networks are both based on CNNs. The specific configurations are given as follows.

Translator Network. The translator network has a simple encoder-decoder architecture. After receiving a concatenated input of the current RGB video frame f_n and the corresponding adult's FoA a_n, the encoder applies a sequence of convolution layers to reduce the spatial size of the input, and the decoder applies a sequence of deconvolution layers to recover the original resolution. The encoder/decoder is designed to have eight 4×4 convolution/deconvolution layers, each followed by the rectified linear activation (ReLU) function. The first three layers of the encoder are consisted of 64, 128, and 256 channels, respectively, whereas the fourth to the last layers consist of 512 channels for each. The decoder is designed to have an almost symmetrical structure with the encoder, i.e., each of the first five layers of it has 512 channels and the following three have 256, 128, and 1 channels, respectively. In order to improve the quality of the generated images, the idea of skip connections [11] which directly connects each of the intermediate encoder layers to the decoder layer having the same spatial resolution is adapted. The skip connections provide an option of by-passing the encoder/decoder part if it doesn't have a use for it.

Discriminator Network. The discriminator is used to improve the translator network in the training process (we will give the detail in Sec. II-B). This is to judge if the provided input is "fake" or "real", i.e., if it is generated by the translator or not. It receives a combination of the two inputs; an adult's FoA a_n and predicted e_n or ground truth elderly's FoA e_n^*. It applies five convolution layers followed by ReLU, except for the last layer where a sigmoid function is used instead of ReLU. There are 64, 128, 256, and 512 channels used for the first to the fourth layers, respectively. Following the idea of PatchGAN proposed in [12], the discriminator outputs the patch-level confidence of the classification result.

B. Model Training

Let us denote translator and discriminator networks as T and D, respectively. Overall, we train these two networks in a unified training framework that solves the following optimization problem with respect to their parameters.

$$\min_T \max_D \mathcal{L}_{\text{Adv}}(T, D) + \lambda \mathcal{L}_{\text{KL}}(T). \quad (1)$$

This objective is consisted of two major loss terms; content loss $\mathcal{L}_{KL}$ and adversarial loss $\mathcal{L}_{\text{Adv}}$. The translator T tries to minimize the objective through the gradient decent, while

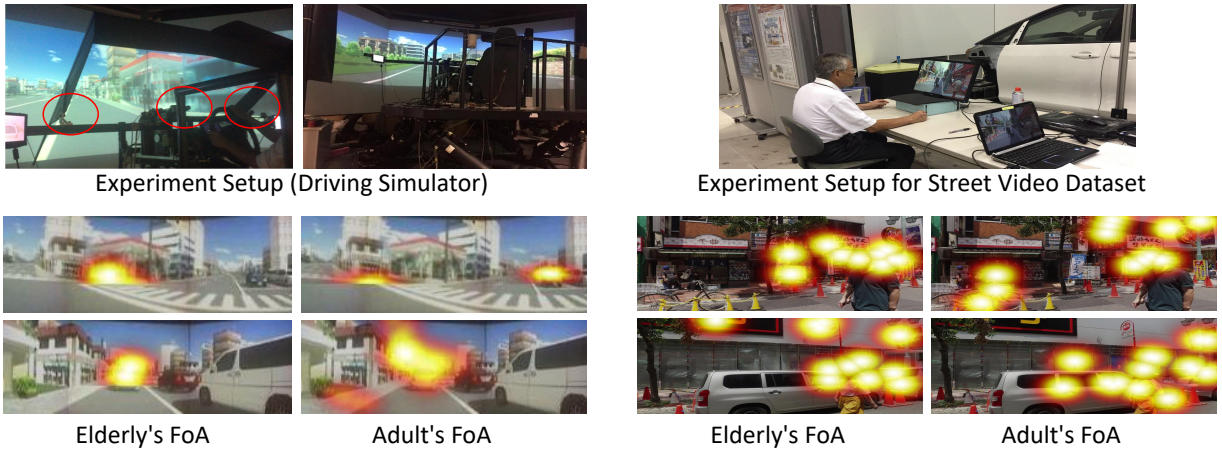

Fig. 2. **Data collection setup.** Left: Setup for Driving Dataset collection. The eye-gaze of participants were recorded by three eye-trackers (in red circle) while driving on a simulator shown in the figure. Right: Setup for Street Video Dataset collection. Examples of the scenes and resulting FoA maps of adult's and elderly's are shown in the bottom part.

the discriminator D tries to maximize it with gradient ascent. Below we give the details of each term one-by-one.

Content Loss. The content loss requires the translator network T to output the ground truth FoA e_n^* in per pixel bases by comparing the predicted FoA of elderly $e_n = T(a_n)$ with e_n^*. Denote by $e_{n,i}$ and $e_{n,i}^*$ the i-th pixel of e_n and e_n^*, respectively. Considering that $e_{n,i}$ and $e_{n,i}^*$ are both probability values, the most appropriate form of the loss function would be the KL-divergence which allows to directly compare the two probability distributions.

$$\mathcal{L}_{\text{KL}}(T) = \sum_n \sum_i e_{n,i}^* (\log(e_{n,i}^*) - \log(e_{n,i})) \tag{2}$$

Adversarial Loss. As explained in the introduction, the assumption behind our approach is that the tendencies of FoAs by adults and elderly are characterized by the scene. The content loss defined above only aims at approximating the conditional distribution of e_n^* with given $\mathcal{F}_n$, i.e., $p(e_n^*; \mathcal{F}_n)$, which may not fully capture the underlying correlation between e_n and a_n. Hence, we introduce an additional loss to explicitly model the joint distribution $p(a_n, e_n^*; \mathcal{F}_n)$. This can be achieved by the generative adversarial learning [11], which is specifically formulated as follows.

$$\begin{aligned}\mathcal{L}_{\text{Adv}}(T, D) = {} & E_{(a_n, e_n^*) \sim p(a_n, e_n^*; \mathcal{F}_n)}[\log D(a_n, e_n^*)] \\ & + E_{a_n \sim p(a_n; \mathcal{F}_n), e_n \sim p(e_n | a_n; \mathcal{F}_n)}[1 - \log D(a_n, e_n)]\end{aligned} \tag{3}$$

By minimizing this with respect to the translator network T under the assumption that the ideal discriminator network D^* is given, T is trained so that the distribution $p(a_n, e_n; \mathcal{F}_n) = p(a_n; \mathcal{F}_n) p(e_n | a_n; \mathcal{F}_n)$ matches the desirable joint distribution $p(a_n, e_n^*; \mathcal{F}_n)$ in terms of Jensen-Shannon divergence. This is approximately achieved by an alternating update of T and D.

III. Dataset Creation

We describe our datasets used in our experiments described in the next section.

Many benchmark datasets for image/video-based FoA estimation have been developed consisting of ground truth FoA maps build by aggregating fixations of several observers for an image[1]. However, observers of these datasets are restricted to a certain age group, typically around 18-35 years old, which cannot be used for the purpose of this research that focuses on elderly's FoAs.

We therefore construct two new datasets that cover both of the adults and elderly observers to evaluate our model. We consider two different scenarios, namely task-based viewing and free-viewing. The first dataset is called **Driving Dataset**, which was collected under one of the most important applications of FoA estimation, a car driving scenario. The second one is called **Street Video Dataset** which is created to simulate free-viewing on a street.

The details of the dataset creation process are described below.

A. Participants

For both datasets, 18 observers belonging to two different age groups, adults and elderly, were recruited (9 observers in each group). The adult and elderly observers had mean age of 26 and 75 years, respectively. The number of observers in our experiment is almost consistent with several previous studies [14], [15], [16], [13], [17], where the total number of observers ranges from 8 to 15 for each image. All the observers were driving licence holders for more than five years and had normal or corrected to the normal vision.

B. Driving Dataset

Each observer was asked to use a driving simulator and to safely drive a car to reach a certain destination. Our simulator is the same as the one used in [18] to analyze the drivers' behaviors[2]. The simulator shows each participant a video sequence consisting of $10,000$ frames of a road

[1]For example, a comprehensive list of available datasets can be found at [13].

[2]To accurately simulate real driving scenarios, the simulator consists of brake and accelerator pedals, an electric steering system with a torque generator, and a stereo sound system to provide sounds. See [18] for more details.

environment. The video was designed in a way that it includes different contexts in terms of landscape and traffic conditions. Specifically, each video frame shows different types of objects such as roads, pedestrians, traffic signs, and cars. A gaze tracking system called Smart-Eye is used to record the gaze movement of each participant while driving in real time. The setup is shown in Fig. 2.

While the same road environment is used for all the participants and they drove in the same road, the two drivers cannot experience the same scene at the same time due to the differences in the driving speed. More specifically, two drivers can not be at the same time to the same place on the road. This imposes a temporal alignment problem of the frames for the videos of all the 18 observers. To resolve this issue, we use the dynamic time warping (DTW) technique to find the alignments of each of 17 videos to a remaining one which we call reference video. As a result, all the 17 videos got temporally aligned with respect to the reference video. After the temporal alignment, the recorded fixations of participants in each age group are overlaid frame-by-frame, which provide us two FoA maps for each video frame showing places attended by two age groups of adults and elderly, respectively. Besides, FoA maps of each observer's are obtained by convolving a Gaussian filter to each fixation point as in the standard protocol for saliency estimation experiments [13].

Consequently, we obtained $9,713$ continuous FoA maps correspond to the $9,713$ frames of the video stimuli for both age groups. We used $7,716$ frames for training and the rest for testing. Some examples of the adult's and elder's FoA maps corresponding to some video frames are shown in the bottom of Fig. 2.

C. Street Video Dataset

Unlike Driving Dataset, this dataset was created to reproduce the elderly's FoA trend in a free-viewing scenario, where the observer has no task in hand. The same participant was again involved and asked to freely look at the video displayed on the monitor. The video was shot in first-person view at a resolution of 1080×1980 pixels on a busy street market with pedestrians and several shops on both sides. We used a 24 inch display to play the video sequence and an eye-tracking system called Tobii to record the eye-gaze movement of each participant while viewing this video from a distance of approximately 60 cm. The experiment setup is shown in Fig. 2. The same processes as the case of Driving Dataset are used to yield FoA maps of each age group. Finally we obtained $4,425$ maps, and $3,532$ are used for training and the rest for testing. Some examples of the resulting maps are shown in the bottom of Fig. 2.

IV. Experiments

We empirically demonstrate the effectiveness of our framework and compare with existing FoA estimation methods on both Driving Dataset and Street Video Dataset.

TABLE I
COMPARISON WITH BASELINES ON DRIVING DATASET.

Algorithm	CC↑	SIM↑	KL↓	Time (sec.)↓
[19]	0.13	0.22	5.60	6.31
[20]	0.09	0.26	4.90	6.43
[2]	0.26	0.42	9.97	**2.71**
[3]	0.64	0.53	4.06	7.48
[3] (fine-tuned)	0.66	0.55	3.89	7.48
Ours	**0.91**	**0.79**	**0.80**	7.56

TABLE II
COMPARISON WITH BASELINES ON STREET VIDEO DATASET.

Algorithm	CC↑	SIM↑	KL↓	Time (sec.)↓
[21]	0.24	0.49	0.82	4.10
[22]	0.22	0.47	1.00	6.33
[5]	0.27	0.46	2.21	9.23
[2]	0.27	0.47	1.36	**2.74**
[2] (fine-tuned)	0.58	0.57	**0.58**	**2.74**
Ours	**0.72**	**0.71**	0.94	2.93

A. Implementation Details

For the predictor network of our model, we used two existing networks depending on the dataset, specifically, [3] for Driving Dataset and [2] for Street Video Dataset. The reason for choosing [3] for Driving Dataset is that it is especially designed and trained for saliency estimation in a driving scenario. In contrast, [2] is designed to predict saliency in free-viewing scenarios, which is suited our Street Video Dataset. According to their network configurations, the numbers of input video frames are set to $k = 16$ and $k = 1$ for Driving Dataset and Street Video Dataset, respectively, which are those used in [3] and [2].

For our translator and discriminator networks, we trained them from scratch for both datasets during 50 epochs by using Adam with the learning rate of 2×10^{-4} and momentum of 0.5. The weight for the KL-divergence loss term λ is fixed to 100. Note that only the parameters of the the translator and discriminator networks are updated during the training, while those of the predictor network are all fixed with the values pre-trained on the datasets of adults' FoAs used in their original papers [3] and [2].

B. Experiment Setup

To measure the quality of the predicted FoA maps, we used three popular metrics including Pearson's Correlation Coefficient (CC), Similarity (SIM), and KL-Divergence [23]. Higher values of CC and SIM mean better performance, while lower KL-Divergence is better.

We compare our method with several baselines [5], [19], [20], [21], [22] including both deep and non-deep methods. In addition, to prove the effectiveness of our translator network, we also compared our method with [3] and [2] (which are used as our predictor network) on Driving Dataset and Street Video Dataset, respectively. For these two models, we prepared both pre-trained and fine-tuned versions (with Driving Dataset and Street Video Dataset, respectively) for fair comparisons. For running the baslines, we used the codes and pre-trained weights provided by each author group and freely available on their project page. We make sure that same hardware

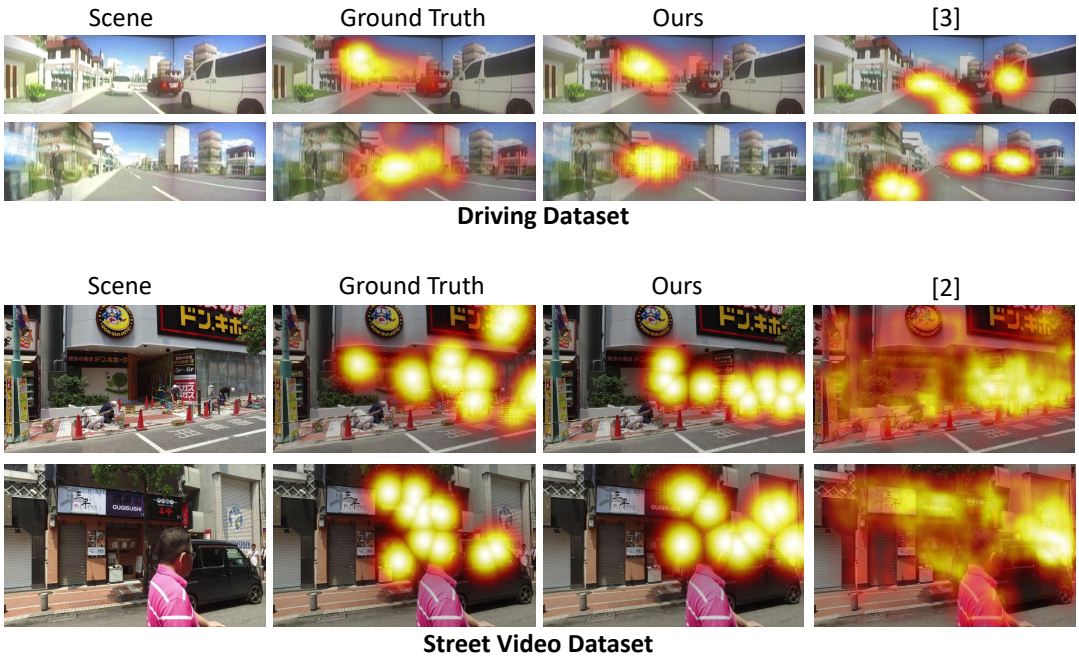

Fig. 3. **Qualitative results**. From left to right, input frame, ground truth eldery's map, predicted FoA maps by ours and the baseline method ([3] or [2]).

TABLE III
ABLATION STUDY.

Algorithm	CC↑	SIM↑	KL↓	Time (sec.)↓
Driving Dataset				
w/o RGB	0.77	0.68	2.18	**7.53**
w/ RGB	**0.91**	**0.79**	**0.80**	7.56
Street Video Dataset				
w/o RGB	0.64	0.66	1.13	**2.89**
w/ RGB	**0.72**	**0.71**	**0.94**	2.93

environment is used for each method and hyper-parameters of these methods are carefully tuned.

C. Quantitative Results

We first report the quantitative performance of our method evaluated in terms of the quality of the predicted map with respect to the ground truth map, and, we also record average test time in predicting eldery's FoA for a video frame.

Results on Driving Dataset. Results are shown in Table I. We can see that our model outperforms all the baselines with significant margins. [20] and [19] poorly perform in this dataset. The reason is that these methods fully rely on bottom-up hand-crafted features, which suffers a few bottlenecks such as feature selection and integration. Furthermore, the bottom-up guidance of eye movements is less prominent in such a task-based viewing scenario.

Our method performs better than other deep learning-based methods [2], [3]. This suggests that even the models trained with large-scale training datasets of adults' FoA maps are not sufficient to accurately predict the elderly's. We also fine-tuned the pre-trained model of [3] with our eye-gaze data of elderly's in order to do a fair comparison. However, our model still performs better than the case. This shows that our model is better trained with a small number of training data, demonstrating the effectiveness of our approach based on image-to-image transformation.

We also measure the prediction time of the methods to evaluate how much complexity our method adds to the predictor network which has exactly the same architecture as [3]. The rightmost column of Table I shows the results. Although our method is not very fast when compared with all the baselines, the difference from the [3] is reasonably small. This is because our translator network is a 2D fully-convolutional network consisting of only 16 layers with small kernels, which is fairly compact compared to the predictor network [3] that uses multiple C3D networks. Ours achieves huge performance gain with this slight expense of run time. Although [2], basing on a 2D fully-convolutional network, is the fastest despite taking a higher resolution image than ours, its accuracy is not satisfactory on this dataset.

Results on Street Video Dataset. Results on Street Video Dataset are shown in Table II. We can see that our model on this dataset also outperforms all the baselines with huge margins, which demonstrates the generalizability of our method. Unlike the cases of Driving Dataset, the methods based on handcrafted features [22], [21] perform relatively well on this dataset. This is because such bottom-up methods tend to work better in free-viewing scenarios. However, the gain of ours is still huge. The methods with deep learning [5], [2] still perform poorly when compared to ours, as the training data used are collected from adult participants only. We also compared ours with the fine-tuned version of [2] with our elderly eye-gaze data, ours still performs better than the case. The rightmost column of Table II shows the computation time taken in predicting FoA for a single frame. Again, the time added by our translator network is reasonably small compared to the

predictor network [2].

Ablation Study. Besides our full proposed model, we also tested a variant of our translator network that takes only the predicted adult's FoA (without the original RGB video frame) as its input. This is to understand the significance of context (RGB frame) while predicting elder's FoA.

Table III shows that our full model (w/ RGB) is better than that does not look at the RGB video frame (w/o RGB). This may be because ours w/ RGB can capture the content-dependent interactions to the eldery's FoA by using $p(e_n|a_n;\mathcal{F}_n)$ rather than $p(e_n|a_n)$. These results overall show that the superiority of our approach based on deep image-to-image translation to the existing methods.

D. Qualitative results

We show qualitative results in Fig. 3. The results show the remarkable ability of our model in predicting elderly's FoA. The examples for Driving Dataset show that our model accurately mimics the gaze tendency of elderly people, who tend to focus only on cars in the same lane, whereas the predicted FoA maps by the most competitive baseline [3] pays attention to the cars in different lanes and even to the objects in the environment. These results reflect typical behaviors of elderly and adults, respectively. A similar trend is seen in the examples of Street Video Dataset. Ours tends to look only at a few RoIs of the scenes, which is close to the behaviors observed in the ground truth images, while [2] covers a larger area. These results indicate that our method can appropriately mimic the FoAs of elderly's.

V. Conclusion

In this paper, we attempted to accurately predict the elderly's FoA by introducing a deep image-to-image translation framework. The proposed model is an encoder-decoder type network that takes adult's FoA generated by the state-of-the art FoA prediction method as an input and translates it to the elderly's FoA. The evaluation experiments are performed on two different datasets to cover both free-viewing and task-based viewing scenarios. Both qualitative and quantitative results show that our model can mimic the elderly's FoAs more accurately compared with the state-of-the-art baseline methods originally designed and evaluated for adult's FoA prediction.

We believe that this paper will open up a new direction of utilizing image-to-image transformation for FoA estimation. Our study showed that the gaze tendency of observers of different age groups can be effectively transformed by a fairly simple image-to-image transformation network. Exploring applicability of this framework to other types of human attributes may also be an interesting future direction of this research.

Acknowledgement. We would like to thank Prof. Kimihiko Nakano of The University of Tokyo for availing the driving simulator and assisting with the data collection.

References

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 20, no. 11, pp. 1254–1259, 1998.

[2] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "A deep multi-level network for saliency prediction," in *Proc. International Conference on Pattern Recognition (ICPR)*, 2016.

[3] A. Palazzi, D. Abati, S. Calderara, F. Solera, and R. Cucchiara, "Predicting the driver's focus of attention: the DR(eye)VE project," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 47, no. 7, pp. 1720–1733.

[4] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an LSTM-based saliency attentive model," *IEEE Trans. Image Processing (TIP)*, vol. 27, no. 10, pp. 5142–5154, 2018.

[5] L. Jiang, M. Xu, T. Liu, M. Qiao, and Z. Wang, "DeepVS: A deep learning based video saliency prediction approach," in *Proc. European Conference on Computer Vision (ECCV)*, 2018.

[6] A. Peltschae, A. Hemraja, A. Garciaab, and D. Munoz, "Age-related trends in saccade characteristics among the elderly," *Neurobiology of Aging*, vol. 32, no. 4, pp. 669–679, 2011.

[7] D. J. Madden, "Aging and visual attention," *Current Directions in Psychological Science*, vol. 16, no. 2, pp. 70–74, 2007.

[8] O. Krishna, T. Yamasaki, A. Helo, R. Pia, and K. Aizawa, "Developmental changes in ambient and focal visual processing strategies," *Electronic Imaging*, vol. 2017, no. 14, pp. 224–229, 2017.

[9] O. Krishna, A. Helo, P. Rämä, and K. Aizawa, "Gaze distribution analysis and saliency prediction across age groups," *PLoS One*, vol. 13, no. 2, 2018.

[10] A. T. Zhang and B. O. Le Meur, "How old do you look? Inferring your age from your gaze," in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2018.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2014.

[12] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *Proc. European Conference on Computer Vision (ECCV)*, 2016.

[13] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "Mit saliency benchmark dataset," http://saliency.mit.edu/datasets.html.

[14] S. Fan, Z. Shen, M. Jiang, B. L. Koenig, J. Xu, M. S. Kankanhalli, and Q. Zhao, "Emotional attention: A study of image sentiment and visual attention," in *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[15] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *Journal of Vision*, vol. 14, no. 1:28, pp. 1–20, 2014.

[16] T. Judd, F. Durand, and A. Torralba, "Fixations on low-resolution images," *Journal of Vision*, vol. 11, no. 4, pp. 14–14, 2011.

[17] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2008.

[18] Z. Wang, R. Zheng, T. Kaizuka, and K. Nakano, "Relationship between gaze behavior and steering performance for driver-automation shared control: A driving simulator study," *IEEE Trans. Intelligent Vehicles (TIV)*, vol. 4, no. 1, pp. 154–166, 2019.

[19] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[20] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Trans. Image Processing (TIP)*, vol. 24, no. 11, pp. 4185–4196, 2015.

[21] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2007.

[22] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, pp. 1489–1506, 2000.

[23] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 41, no. 3, pp. 740–757, 2016.