2021 Thirteenth International Conference on Quality of Multimedia Experience (QoMEX)

Focus Guided Light Field Saliency Estimation

Ailbhe Gill†, Emin Zerman†, Martin Alain†, Mikael Le Pendu‡, and Aljosa Smolic†
† V-SENSE, School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
Emails: {gilla3, zermane, alainm, smolica}@tcd.ie
‡ Inria Centre de Recherche Rennes Bretagne Atlantique, 35042 Rennes, France
Email: mikael.le-pendu@inria.fr

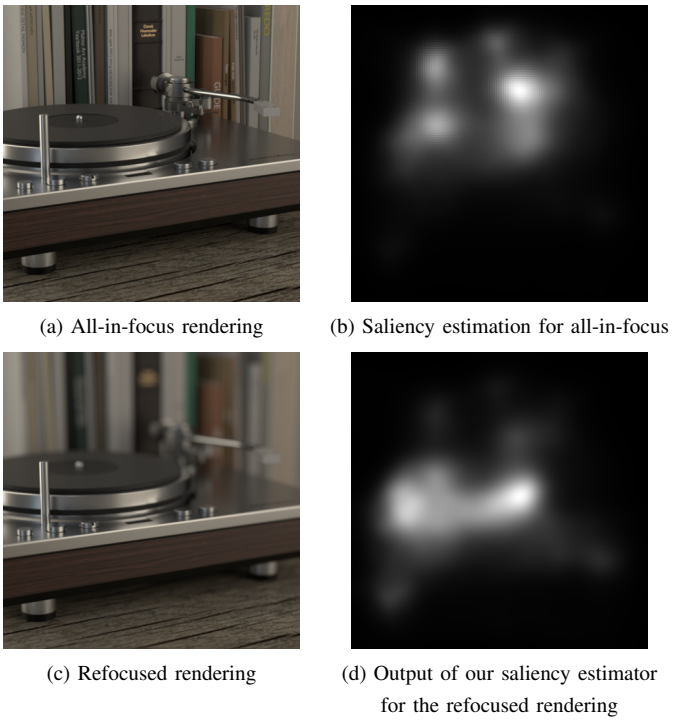
***Abstract*—Light field imaging enables us to capture all light rays in a visual scene. As light fields are four-dimensional, their captures come with an increased amount of information to take advantage of. This has stimulated ongoing light field specific research into virtual viewpoints and shallow depth of field rendering, commonly called refocusing. However, the computation time and memory required to perform these operations can make tasks such as real-time rendering impractical. One solution is to exploit the salient information of light fields to focus resources on regions that attract visual attention when using these algorithms. Although saliency estimation methods for light fields have been previously explored, these focus mainly on salient object segmentation with the goal of generating one saliency map per light field.**

Aiming to create a basis for a 4D saliency prediction model analogous to light fields, this paper proposes a saliency estimation method specific to light fields that considers the refocusing operation. The proposed method modifies an existing view rendering algorithm with focus guidance, obtained from the light field disparity. This facilitates the construction of saliency maps without the need to render the corresponding view itself, which will help to speed up processing operations that are compatible. The results show that the proposed saliency estimation approach yields very good predictions of visual attention across multiple planes of the light field. We anticipate that this approach can be extended for a range of rendering applications.

***Index Terms*—light field, saliency, refocusing, rendering, visual attention**

I. Introduction

Light field (LF) imaging technology aims to capture and recreate all the rays passing through an area in 3D space [1]. A common method for capturing LFs is to sample the light rays using two parallel planes. This enables the capturing of angular information in addition to spatial information. With the two-plane parametrization, an LF can be represented as a 4D function that is defined both on angular (s, t) and spatial (u, v) axes: $(s, t, u, v) \rightarrow L(s, t, u, v)$. The captured LFs are commonly represented as a matrix of $S \times T$ views of $U \times V$ spatial resolution. These $U \times V$ sized images are named sub-aperture images (SAI), which we denote by $I_{s,t}(u, v) = L(s, t, u, v)$ for convenience. On the one hand, this increased dimensionality brings utility: LFs can be used in many different applications including estimating the geometry or the depth of the scene [2], rendering new views from different viewpoints [1], and changing the focus (or refocusing) of the scene [3], [4], see Fig. 1.(c). On the other hand, it brings challenges for the visual perception aspects of this new form of media.



(a) All-in-focus rendering (b) Saliency estimation for all-in-focus (c) Refocused rendering (d) Output of our saliency estimator for the refocused rendering

Fig. 1: Visualization of (a) an all-in-focus rendering of a LF, (b) its saliency estimation result using a state-of-the-art VA model Deepgaze II, (c) a refocused rendering of a LF, and (d) the result of the proposed saliency estimation method.

Understanding viewers' visual attention (VA) is important for various applications such as compression, transmission, rendering, and visualization for both traditional 2D-images and LFs. It is also crucial to be able to estimate the saliency map before the corresponding view is actually rendered, especially in applications where the saliency map is used in the rendering process itself, e.g. compression [5] and foveated rendering [6]. However, collecting VA in user studies is not always feasible, and so to predict the VA distribution, automatic saliency estimation algorithms are used.

Throughout this paper, we define *visual attention* as where people look when viewing a visual scene and *saliency maps*

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/RP/27760.

978-1-6654-3589-5/21/$31.00 ©2021 IEEE

as the mapping of the estimated VA of a visual stimulus. For VA estimation, eye-tracking data collected from participants is used as ground truth. So, in this context, *saliency* represents the probabilistic distribution of eye fixation.

Saliency estimation for traditional 2D images is well-established for use in a variety of applications and for different definitions of saliency. However, presently saliency estimation methods for LFs [7]–[10] focus on detecting and segmenting the salient objects in a scene (cf. Section II). Saliency estimation for different novel renderings of the LF using other definitions of saliency is still an open problem.

Traditionally saliency research has focused on the task of eye fixation prediction, where a saliency map assigns a probability of visual importance to every pixel of an image. A light field is defined as a collection of rays, which can be represented as L shown above. In this paper, we define the concept of LF saliency as the probability of visual importance of every ray of a LF, and introduce the corresponding representation as a saliency field Ψ below (cf. Section III-B). We aim to use this representation to estimate the VA of refocused views of the LF, see Fig. 1. To achieve this, we propose a focus guided LF VA prediction method. For this method, we modify a classical refocus rendering algorithm by integrating the disparity information relevant for the refocusing operation. The proposed algorithm is validated on an LF VA database visually and quantitatively. Our approach can be used to estimate saliency for the refocusing operation without having to render the views. Our results show that the integration of the focus guidance improves the saliency estimation and helps yield an accurate VA prediction.

II. Background & Related Work

A. Light field refocusing

A common LF operation is to render a 2D image simulating a traditional photographic camera with a narrow depth of field, with the ability to choose the focal plane, also called refocusing. A refocus image I_r can be produced through use of the well-known shift-and-sum algorithm [3], in which it is obtained as a linear combination of shifted LF SAIs:

$$I_r(u, v, \delta_F) = \sum_{s,t} A(s,\, t)\, I_{s,t}(u_F,\, v_F), \tag{1}$$
$$u_F = u + (s - s_r)\delta_F,$$
$$v_F = v + (t - t_r)\delta_F$$

where (s_r, t_r) corresponds to the position of the refocus image on the camera plane, δ_F is the disparity value corresponding to the focus distance with (u_F, v_F) the corresponding pixel shift, and A is a filter that defines the synthetic aperture. Intuitively, the shift-and-sum algorithm aligns the regions of the SAIs corresponding to the target disparity δ_F. High frequency textures and edges are thus preserved for these regions, but blurred otherwise. Increasing the size of the aperture filter A will combine more SAIs and result in a shallower depth of field, as for traditional cameras.

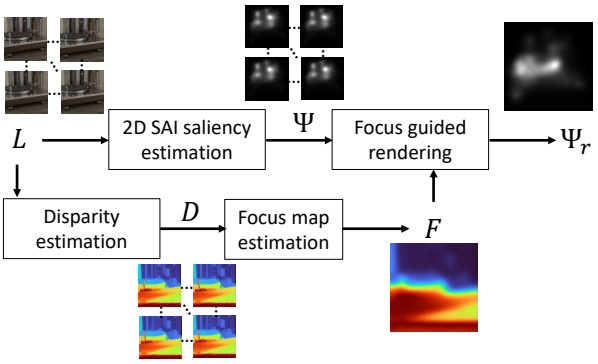


Fig. 2: Pipeline of the proposed approach.

B. Saliency estimation

There are many different types of saliency prediction models. One of the ways to categorize these methods is by their definition of saliency. Some models treat saliency as segments of objects which stand out in a visual scene. These segments of semantic objects can then be used for multimedia analysis. Others define visual saliency as the probabilistic distribution of eye fixation. These models are trained and evaluated on their ability to predict saliency by using eye-tracking data as ground truth. For our research in VA, we are concerned with all regions that draw a viewer's gaze. Therefore, this latter definition of visual saliency is more suitable for our work.

The top performing visual saliency models for 2D-images are tested on the MIT/Tuebingen Saliency Benchmark [11]. DeepGaze II [12] is one such model that we use as part of our pipeline. For a given input image, it uses the pre-trained VGG-19 network to obtain feature maps. These are then passed through a small readout network trained and evaluated for visual saliency prediction using information gain.

LF saliency is a growing field of research. Current LF saliency prediction models [7]–[9] focus on object-based methods of prediction. The related ground truth datasets consist of binary maps of segmented objects and are used for training and evaluation. In its current form, saliency estimation of LFs is mostly based on an all-in-focus rendered RGB image of the LF, and the SAIs and depth map are used to incorporate additional information to improve the prediction performance. There has been only one model that uses multiple refocused renderings as a 4D input but their goal was still to output a single 2D map with the salient object segmented [10].

However, work by Gill et al. [13] shows that a 2D saliency map can not fully explain the saliency of LFs with regards to VA. They demonstrate, from eye-tracked fixation data of different LF renderings that there is variation in VA across the different renderings. They also observed that participants gaze was drawn to in-focus regions. To the best of our knowledge, this is the first study that tackles LF VA for rendered images.

III. Proposed Method

In this section, we describe the proposed focus guided saliency estimation ($FGSE$) algorithm for LFs and its integration into a rendering framework.

A. Focus guided saliency

As explained in the introduction, we define in this paper the concept of LF saliency as the probability of visual importance for every ray of a LF. Formally, we denote the saliency field as a 4D function $\Psi(s,t,u,v)$. For convenience, we define a "saliency SAI" as $\Psi_{s,t}(u,v) = \Psi(s,t,u,v)$. While LF SAIs are natural images, which contain low to high frequencies, saliency SAIs only contain low frequencies and do not have high frequency textures or edges, see Fig. 1.(b). Therefore, the existing shift-and-sum refocusing algorithm described in (1) can not be directly applied to the saliency field.

Based on the assumption that gaze is attracted by in-focus regions as found in [13], we propose to process the saliency SAIs using a modified shift-and-sum algorithm guided by a focus map. We obtain the focus map from the disparity maps estimated from the LF, and we denoted the 4D disparity field as $D(s,t,u,v)$. The saliency field Ψ is obtained by independently estimating the saliency of the LF SAIs with the 2D saliency estimator DeepGaze II [12]. The overall pipeline of the proposed approach is shown in Fig. 2.

As the ground truth VA maps were obtained by applying Gaussian filtering (corresponding to 1° visual angle) at fixation points to take visual acuity into consideration, we apply the same blur to the disparity field input of our method in order to closer match the properties of these VA maps. In [13], VA maps are obtained with a Gaussian kernel:

$$B(u,v) = \frac{1}{2\pi\sigma_u\sigma_v}\exp\left(-\frac{u^2}{2\sigma_u^2} - \frac{v^2}{2\sigma_v^2}\right) \tag{2}$$

where $\sigma_u = \frac{47.66}{1080}U$ and $\sigma_v = \frac{42.67}{1080}V$. We use the same Gaussian kernel to blur the disparity maps:

$$D_b(s,t,u,v) = B(u,v) \circledast D(s,t,u,v), \forall(s,t) \tag{3}$$

where $\circledast$ is the convolution operator. In addition to approximating the VA maps' properties, blurring the disparity maps is also advantageous as we can use fast disparity estimation, for which errors are removed by the blurring process. In our experiment we used the method proposed by Chen at al. to estimate the disparity field [2].

The 5D focus map $F(s,t,u,v,\delta)$ can then be obtained from the blurred disparity field for a given target disparity δ as:

$$F(s,t,u,v,\delta) = \exp\left(\frac{-||D_b(s,t,u,v) - \delta||^2}{\sigma_F^2}\right) \tag{4}$$

We choose to express the focus map as soft probability using a Gaussian distribution rather than a binary mask to compensate for remaining errors in the blurred disparity field. The parameter σ_F controls the "depth of field" of the focus map. We observed in the ground truth VA maps from [13] that VA depends on the strength of the defocus blur. As the maximum strength of the defocus blur depends on the maximum disparity, we introduce an intermediate parameter σ_D, such that $\sigma_F = \sigma_D * (max(D) - min(D))$, where $max(D)$ and $min(D)$ are the higher and lower bound of the disparity range respectively. The parameter σ_D allows for easy controlling of the focus map depth of field for all LFs in the dataset, rather than experimentally defining σ_F for each LF.

B. Integration into rendering

The main idea of the proposed focus-guided rendering method is to modify the shift-and-sum algorithm to weight the saliency SAIs with the focus map:

$$\Psi_r(u,v,\delta_F) = \sum_{s,t} A(s,\,t)\, F(s,t,u_F,\,v_F,\,\delta_F)\Psi_{s,t}(u_F,\,v_F) \tag{5}$$

By the properties of the shift and sum, all the shifted focus maps are aligned and almost equal. In addition, given that the saliency SAIs are composed of low frequency values, we can use the following approximation:

$$F_r(u,v,\delta_F) \simeq F(s,t,u_F,\,v_F,\delta_F), \forall(s,t) \tag{6}$$

The algorithm can thus be simplified as:

$$\Psi_r(u,v,\delta_F) = F_r(u,v,\delta_F) \sum_{s,t} A(s,\,t)\,\Psi_{s,t}(u_F,\,v_F) \tag{7}$$

With this simplification we observed experimentally that the processing time is 25% faster compared to the direct approach of (5), while maintaining similar saliency estimation performance (see Table I)[1].

IV. Experimental Results

To validate the proposed focus guided saliency estimation approach for LFs, we make use of an LF VA database and compare our results with a state-of-the-art saliency estimation method. Here, we briefly describe the database, selected saliency estimation method, and selected metrics. Then, we present and discuss results.

A. Database

In our study, we use the ground truth visual attention data collected by Gill *et al.* [13] for our LF rendering approach. This dataset was chosen as it is the only one that has collected eye-fixation data and for LF renderings on multiple planes of focus. This is necessary for our research in building and substantiating a four-dimensional light field saliency field.

The visual stimuli presented to participants was generated using 5 different renderings approaches on 20 LFs. We only consider the 34 stimuli from this database that correspond to refocus images of 2D full-parallax LFs - two different focal renderings of 17 LFs. These renderings were named as "Region-1" and "Region-2" in the database paper, and we keep the same notation in our paper for consistency. The LFs we selected were acquired by various means from three LF datasets specifically the EPFL Light Field Image Dataset [14] using a camera with a microlens array, the Stanford (New) Light Field Archive [15] using a camera array, and the HCI Heidelberg 4D Light Field Dataset [16] using computer generated imagery.

[1]For more details and the code, see https://v-sense.scss.tcd.ie/research/light-fields/light-field-saliency-estimation/

TABLE I: Analysis of the proposed method's parameters[†].

Saliency method	AUC↑	NSS↑	CC↑	KLD↓	SIM↑
$FGSE$ Eq. 5 - w/o blur	0.844	1.614	0.672	0.659	0.636
$FGSE$ Eq. 7 - w/o blur	0.844	1.608	0.671	0.680	0.635
$FGSE$ Eq. 5 - w/ blur	**0.845**	1.615	0.678	**0.616**	0.639
$FGSE$ Eq. 7 - w/ blur	**0.845**	**1.618**	**0.680**	0.619	**0.640**

[†]All $FGSE$ methods use $\sigma_D = 0.4$. **Boldface** indicates the best result in each column.

B. Saliency estimation method

Both for the estimation of the saliency SAIs $\Psi_{s,t}$, and as an anchor metric for validation, DeepGaze II [17] was selected as one of the highest performing saliency estimation algorithms according to MIT/Tŭbingen Saliency Benchmark [11]. It takes as input a regular 2D-image and outputs a saliency map which represents the likelihood of eye fixation.

C. Selected evaluation metrics

To evaluate the proposed method quantitatively, we selected five evaluation metrics for saliency: area under curve (AUC), normalized scanpath saliency (NSS) , Pearson's correlation coefficient (CC), Kullback-Leibler divergence (KLD), and Similarity (SIM). These were selected as they are the most commonly reported metrics in saliency evaluation [18]. To compute these metrics we used open source code[2] [19].

These metrics measure our saliency estimator's performance varying in approach and criteria. AUC and NSS are location-based similarity metrics, CC and SIM are distribution-based similarity metrics, and KLD is a distribution-based dissimilarity metric [18]. AUC treats evaluation as a classification problem. Evaluating the estimated saliency against ground truth fixations, the true positive rate and false positive rate are found. AUC measures the area under the receiver operating characteristic curve plotted using these values. It scores closer to 1 the better the estimation. NSS computes the average normalized saliency between the saliency map and the ground-truth fixations. The higher the score the better the estimator predicts VA. CC calculates the linear correlation between two heatmaps: the estimated saliency and the ground truth VA distribution. KLD measures how far apart the saliency estimation is from the underlying VA distribution, where a higher score indicates higher dissimilarity. Lastly, SIM outputs the similarity between two saliency maps viewed as histograms. A SIM of 1 indicates the distributions are the same and 0 suggests there is no overlap.

D. Quantitative analysis

In Table I, evaluation metrics are reported for the two variations of the proposed focus guided rendering method described in (5) and (7), with and without the disparity blurring described in (2) and (3). The algorithm with blur applied performs better than that without. The simplification of the algorithm using (7) results in a worse score for KLD but improvements in the NSS, CC and SIM metrics.

[2]https://github.com/dariozanca/FixaTons

TABLE II: Metric results[‡] for the proposed $FGSE$ method compared with the baseline shift & sum saliency estimation ($SSSE$) without focus guidance.

Saliency method	AUC↑	NSS↑	CC↑	KLD↓	SIM↑
$SSSE$	0.817	1.348	0.568	0.695	0.583
$FGSE_{\sigma_D=0.7}$	0.831	1.463	0.618	0.627	0.610
$FGSE_{\sigma_D=0.6}$	0.834	1.497	0.632	0.614	0.618
$FGSE_{\sigma_D=0.5}$	0.839	1.546	0.652	*0.602*	0.628
$FGSE_{\sigma_D=0.4}$	0.845	1.618	0.680	0.619	0.640
$FGSE_{\sigma_D=0.3}$	*0.847*	1.713	0.713	0.790	*0.649*
$FGSE_{\sigma_D=0.2}$	0.835	*1.744*	***0.717***	1.445	0.629
$FGSE_{\sigma_D=0.1}$	0.781	1.572	0.637	3.882	0.512
DeepGaze II	**0.851**	**1.745**	0.703	**0.585**	**0.653**

[‡]DeepGaze II results are reported for readers' information. **Boldface** indicates the best score for each column, and *Italic* indicates the best results for the $FGSE$ method.

In Table II, we compare the performance of our estimator for different values of σ_D to that of Deepgaze II run on the refocused rendering and the no focus guidance baseline.

Our estimated saliency method achieves the best score for the CC metric and results comparable to DeepGaze II for the other metrics. The worse AUC score but very close NSS score compared to Deepgaze II suggests that our model has less low valued false positives but also less intense saliency at fixation locations. This shows the effectiveness of our proposed algorithm considering the DeepGaze II method needs the LF rendering operation to be completed before saliency estimation whereas our method computes the entire saliency field from only the SAI input without rendering. Overall, the differences between the metric values of DeepGaze II and the highest values of the $FGSE$ method (for different σ_D values) are very small. Additionally, $FGSE$ beats the baseline $SSSE$ with respect to AUC, NSS, CC, and SIM scores except for two instances. The KLD scores are mixed as the dissimilarity depends on the spread of the estimated saliency map. However, for higher σ_D, $FGSE$ attains lower (i.e., better) KLD values compared to $SSSE$, and for some σ_D, it scores close to the KLD value of DeepGaze II. Considering all the metrics, we chose $\sigma_D = 0.4$ where our model performs well overall, with high AUC, NSS, CC, and SIM scores and low KLD.

Upon observation of model performance with different σ_D, we see that there is a bias-variance tradeoff in choosing σ_D. The performance decreases at the extremities 0.7 and 0.1 and is best between 0.3 and 0.5. Lowering σ_D increases the influence of focus guidance as the "depth of field" of the focus map increases. Very low σ_D causes high bias in our estimator as the model is too simplistic and only considers the focus map. Thus, less emphasis is put on saliency prediction of the overall image. Conversely, very high σ_D leads to high variance of the model and it places little weight on the in-focus region. High sigma means that the predicted saliency maps for all renderings mostly resemble the SAI saliency estimation. We found that $\sigma_D = 0.4$ balances this tradeoff the best for the stimuli we tested across all metrics.

E. Rendering results

The qualitative performance of our model's output is demonstrated in Fig. 3, which displays the saliency maps

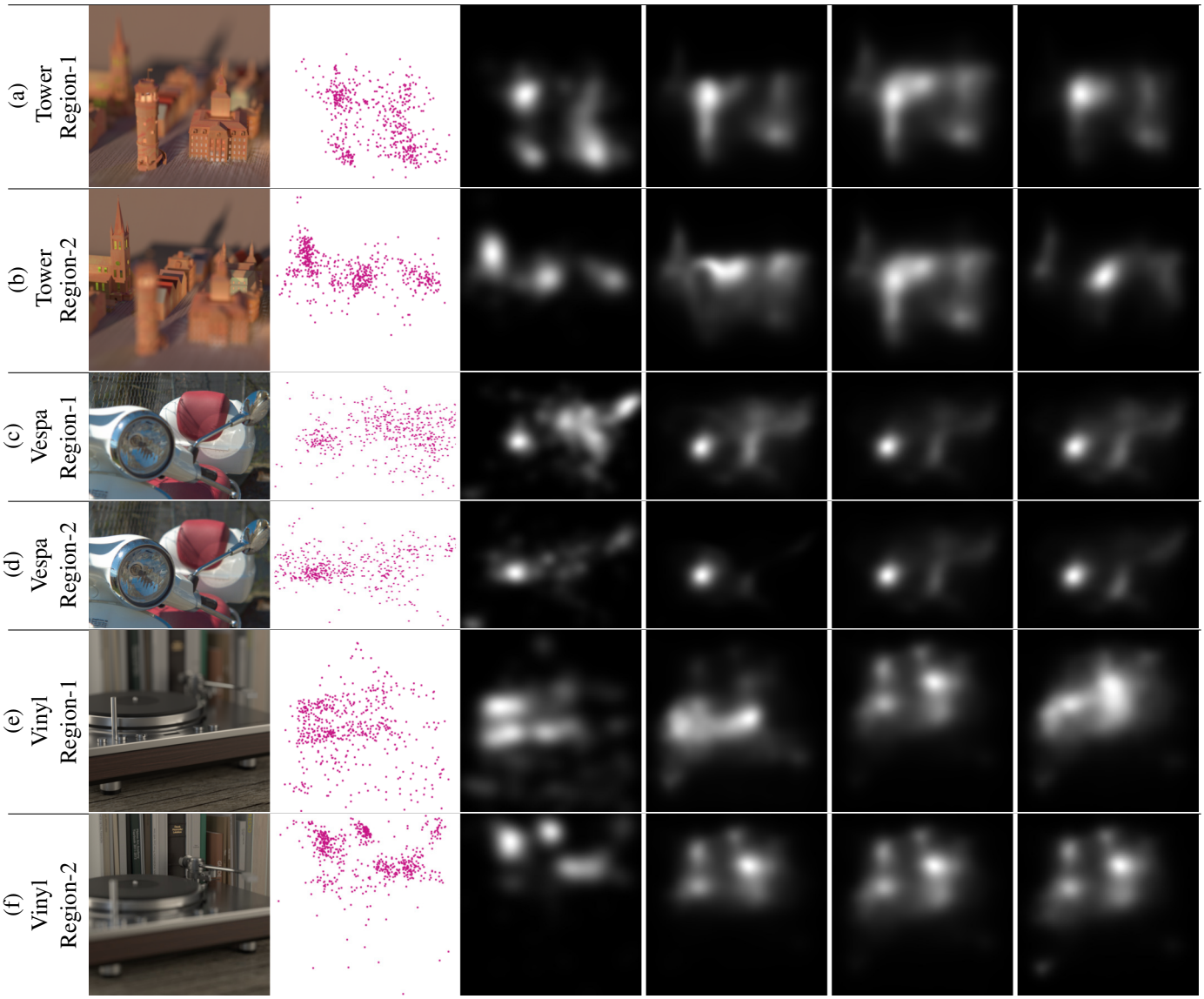


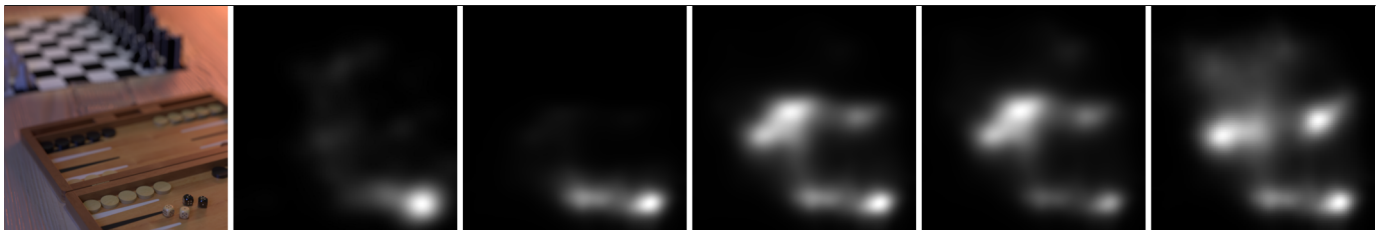
Fig. 3: The results of our focus guided model $FGSE$ shown alongside the stimulus, the ground truth fixations and VA map as well as the no focus guidance $SSSE$ baseline and the results of DeepGaze II run directly on the stimulus. Here the stimuli are the *Region-1* and *Region-2* renderings of a selection of the light fields tested.

outputted by our $FGSE$ method using the simplification of (7) with blurred disparity map (D_b), and $\sigma_D = 0.4$. For comparison, we also provide images alongside our predictions of the following: RGB stimulus, ground truth (GT) fixations, corresponding GT VA maps, Deepgaze II run directly on each stimulus, and the baseline shift & sum saliency estimation ($SSSE$) without focus guidance. For each LF, we used two stimuli to test our model: renderings *Region-1* and *Region-2*. These were chosen because they have sufficiently distant planes rendered to be in focus, emphasising the difference when refocusing [13]. Our model's output can be broken down into two main components.

Firstly, our model closely estimates the variations in concentration of the saliency observed in the ground truth at certain regions. This concentration depends on whether or not the regions appear on the focal plane and therefore guides visual attention. These differences can be observed between the renderings of the Vespa LF Fig. 3.(c) and (d).

Secondly, our model accurately matches the underlying VA across different renderings of the same LF, particularly when a shift in the salient region is observed in the ground truth due to varying the focal plane. This is seen when eye gaze follows a line of focus, for example in the Vinyl LF Fig. 3.(e) and (f). Our model better predicts VA shifts between objects as they come in and out of focus compared to the baseline and is at least on par with the Deepgaze II model run directly on the stimuli as seen in the Tower LF Fig. 3.(a) and (b).

As discussed in the previous section, the σ_D parameter in



Board games Region-2 | GT Visual Attention | $\sigma_D = 0.2$ | $\sigma_D = 0.4$ | $\sigma_D = 0.6$ | DeepGaze II

Fig. 4: The Boardgames *Region-2* rendering, the VA map, the $FGSE$ output using three different σ_D values and DeepGaze II run directly on the stimulus.

our algorithm controls the "depth of field" of the focus map used to generate the estimated saliency map. Consequently, increasing σ_D decreases the influence of the focus guidance on our models predictions. For example, in the Board games Region-2 rendering Fig. 4, the lower $\sigma_D = 0.2$ produces a more visually similar map to the ground truth. For $\sigma_D = 0.4$, the estimated saliency over extends outside the region of focus.

V. Conclusion

In this study, we considered the light field as a scene representation, from which novel views could be rendered. We developed a single cohesive four-dimensional saliency field model for estimating the VA of a LF. Rather than treating the saliency of refocused renderings as separate entities, we employed this model and built a saliency field from which the saliency map corresponding to any refocus rendering could be generated. This model lays the groundwork for predicting the saliency field of various types of LF renderings and for generating most salient renderings. Furthermore, this model has the potential to develop LF applications that can guide a viewers gaze to desired regions.

We tested the efficacy of our model for VA prediction and found that it performs as good as a state-of-the-art visual attention model without the need to render the refocused image. Our model shows that it is possible to generate the saliency of any refocus rendering from only the SAI captures and the disparity map without the need to refocus the entire LF. Our algorithm could be further optimized as a branch of future research to reduce computational complexity, e.g. by estimating the saliency field at lower resolution.

As the σ_D parameter controls the extent by which the focal plane affects the saliency estimation, we observed that there is a tradeoff when choosing σ_D. This parameter influences the performance of each individual LF. While in this paper we took into account the disparity range of each LF, future work could explore additional parameters that influence the "depth of field" of the focus map, such as the aperture size and shape, to automatically compute σ_F for each individual LF.

In future, our model can be extended to other refocusing algorithms [4] and combined with view synthesis to generate saliency maps of other novel views.

References

[1] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. SIGGRAPH*, 1996, pp. 31–42.

[2] Y. Chen, M. Alain, and A. Smolic, "Fast and accurate optical flow based depth map estimation from light fields," in *Proceedings of the Irish Machine Vision and Image Processing Conference*, 2017.

[3] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," *Stanford Technical Report CSTR*, vol. 2, pp. 1–11, 2005.

[4] M. Le Pendu, C. Guillemot, and A. Smolic, "A Fourier disparity layer representation for light fields," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5740–5753, 2019.

[5] K. Wu, Z. Liao, Q. Liu, Y. Yin, and Y. Yang, "A global co-saliency guided bit allocation for light field image compression," in *Data Compression Conference (DCC)*. IEEE, 2019.

[6] X. Meng, R. Du, and A. Varshney, "Eye-dominance-guided foveated rendering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 5, pp. 1972–1980, 2020.

[7] Y. Piao, Z. Rong, M. Zhang, and H. Lu, "Exploit and replace: An asymmetrical two-stream architecture for versatile light field saliency detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 865–11 873.

[8] J. Zhang, Y. Liu, S. Zhang, R. Poppe, and M. Wang, "Light field saliency detection with deep convolutional networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 4421–4434, 2020.

[9] Y. Jiang, T. Zhou, G.-P. Ji, K. Fu, Q. Zhao, and D.-P. Fan, "Light field salient object detection: A review and benchmark," 2021.

[10] Y. Liu, H. Xiao, H. Tan, and P. Li, "Are RGB-based salient object detection methods unsuitable for light field data?" *EURASIP Journal on Image and Video Processing*, vol. 2020, no. 1, pp. 1–17, 2020.

[11] M. Kümmerer, Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "MIT/Tübingen saliency benchmark," https://saliency.tuebingen.ai/results.html, Accessed: 2021-02-06.

[12] M. Kümmerer, T. S. Wallis, and M. Bethge, "DeepGaze II: Reading fixations from deep features trained on object recognition," *arXiv preprint arXiv:1610.01563*, 2016.

[13] A. Gill, E. Zerman, C. Ozcinar, and A. Smolic, "A study on visual perception of light field content," in *Irish Machine Vision and Image Processing Conference (IMVIP)*, Sept 2020.

[14] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *8th International Conference on Quality of Multimedia Experience (QoMEX)*, 2016.

[15] V. Vaish and A. Adams, "The (new) stanford light field archive," http://lightfield.stanford.edu, 2008, [online].

[16] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 19–34.

[17] M. Kummerer, T. S. Wallis, L. A. Gatys, and M. Bethge, "Understanding low-and high-level contributions to fixation prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4789–4798.

[18] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740–757, 2018.

[19] D. Zanca, V. Serchi, P. Piu, F. Rosini, and A. Rufa, "FixaTons: A collection of human fixations datasets and metrics for scanpath similarity," *CoRR*, vol. abs/1802.02534, 2018.