August 22-27, 2021 43rd European Conference on Visual Perception

# A DEEP LEARNING SALIENCY MODEL FOR EXPLORING VIEWERS' DWELL-TIME DISTRIBUTIONS OVER AREAS OF INTEREST ON WEBCAM-BASED EYE-TRACKING DATA

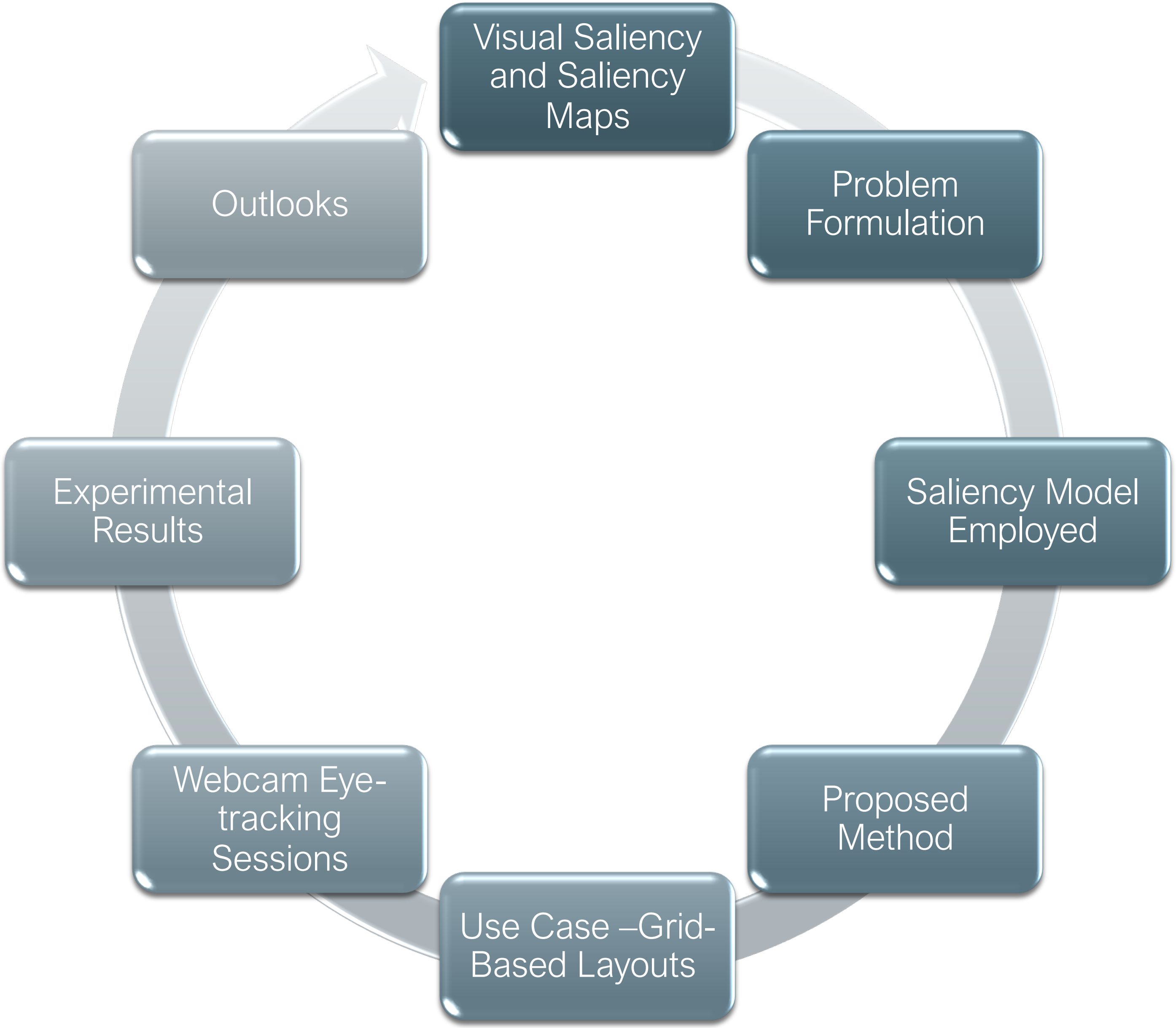
ALESSANDRO BRUNO[1], MAROUANE TLIBA[2], ARZU ÇÖLTEKIN[3]

[1]DEPARTMENT OF COMPUTING AND INFORMATICS AT BOURNEMOUTH UNIVERSITY, POOLE, UNITED KINGDOM

[2]INSTITUT NATIONAL DES TÉLECOMMUNICATIONS ET TIC, ORAN, ALGERIA

[3]UNIVERSITY OF APPLIED SCIENCES AND ARTS NORTHWESTERN SWITZERLAND'S INSTITUTE FOR INTERACTIVE TECHNOLOGIES, BRUGG-WINDISCH, SWITZERLAND

# OUTLINE



Visual Saliency and Saliency Maps

Problem Formulation

Saliency Model Employed

Proposed Method

Use Case –Grid-Based Layouts

Webcam Eye-tracking Sessions

Experimental Results

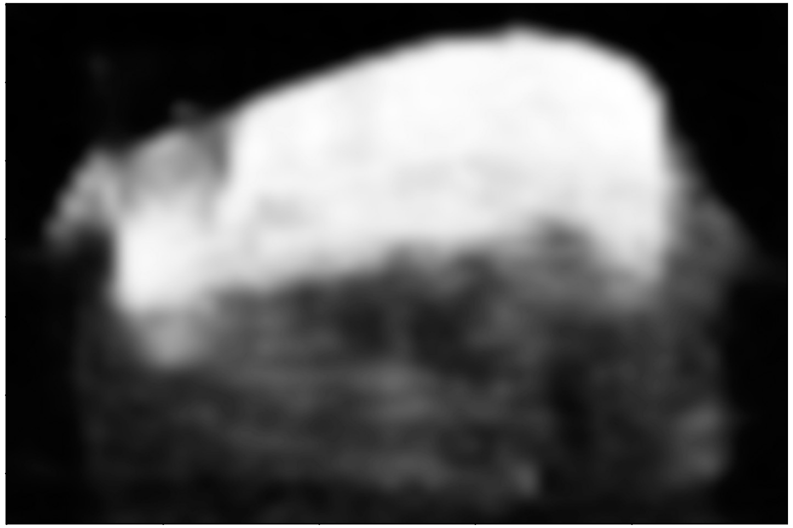
Outlooks

# VISUAL SALIENCY

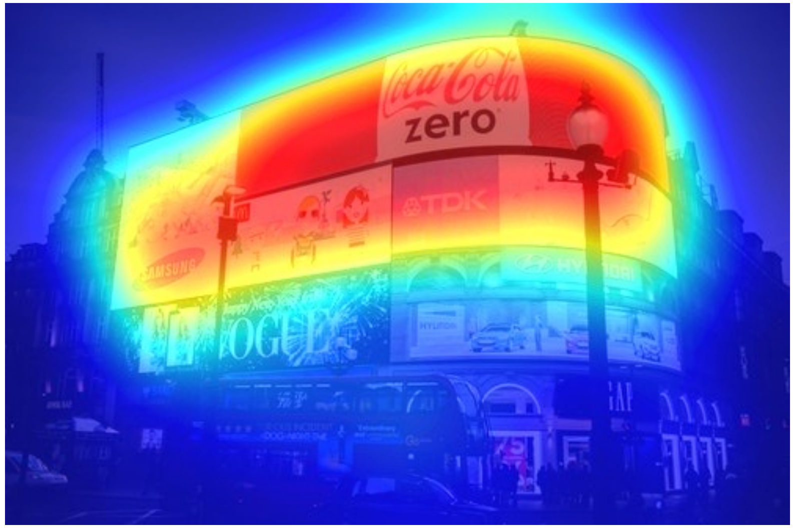
Visual Saliency deals with detecting the most eye-catching regions in images, those regions which naturally stands out of the image. It accounts for bottom-up and top-down visual attention processes over the first few seconds of observation of a given image.

“Visual saliency computation objective can be described as predicting, locating and mining the salient visual information by simulating the corresponding mechanisms in the human vision system.”[2][9]

Input Image



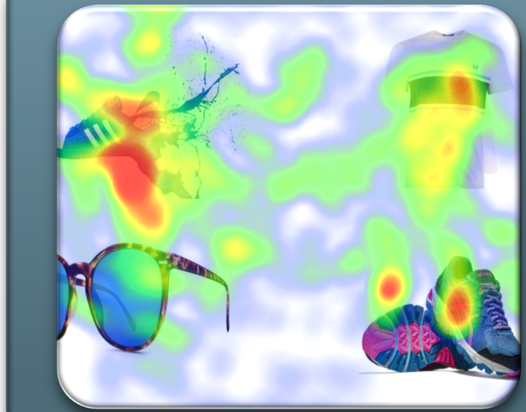
Saliency Map



Heatmap

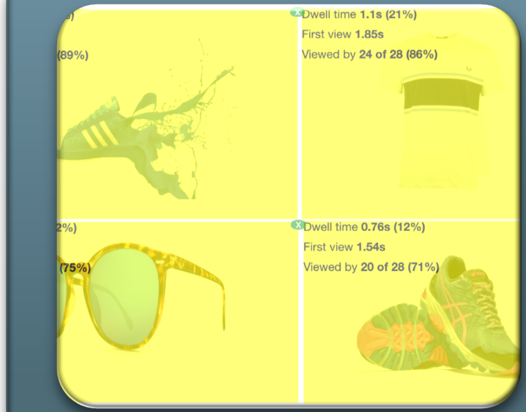
“Given limited computational resources, the human visual system relies on saliency computation to quickly grasp important information from the excessive input from the visual world” [3]

# PROBLEM FORMULATION



## Premise:

- Saliency Maps provide us with predictions of the most eye-catching regions in images.
- Each pixel is, then, encoded with an intensity value in the continuous range [0,1]. A pixel having 0 saliency intensity means that is unlikely to grab viewers' attention; conversely, a pixel showing values around 1 is likely to be 'eye-catching'.
- Saliency is somewhat a biologically inspired research field giving us spatial information on the HVS (Human Vision System) behaviour over the first seconds of observation of a given visual scene.



## Problem Definition:

- Can we use saliency maps to predict viewers' dwell time distribution over different regions of the image?
  - Can we predict what layout configuration better suits some specific requirements?
  - What layout makes a region stand out for longer than other regions?
  - What layout provides an image with the most well-balanced dwell times' distribution over each region?

# GOALS

1
- Running object-oriented saliency over all spatial permutations of regions which an image is made of.

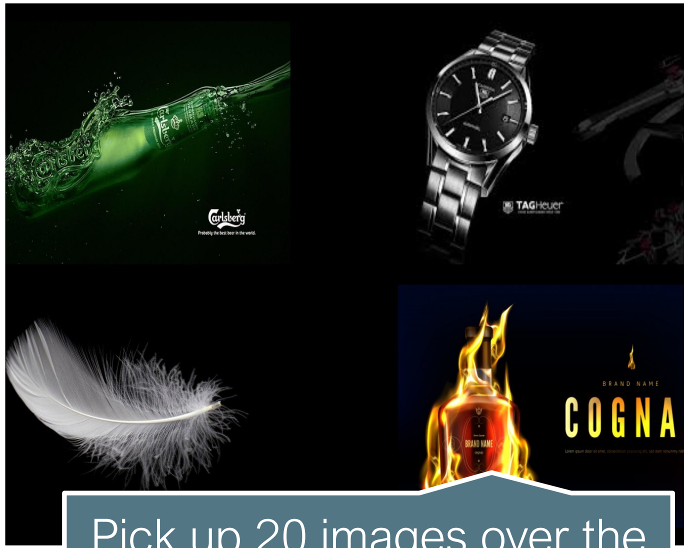
2
- Conduct webcam-based eye-tracking sessions to create a ground-truth for the same images.

3
- Run through all layout configurations and compute local saliency variance throughout all regions for each layout configuration.

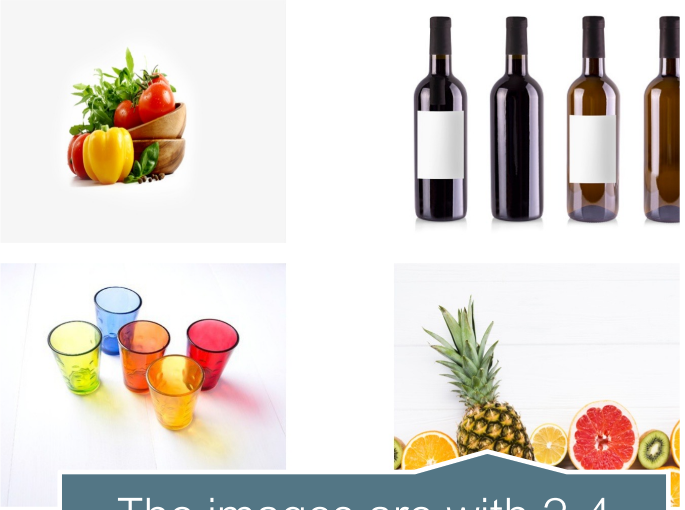
4
- Assess correlations, if any, between the variance of salient local areas and viewers’ dwell times of the same regions in the image.

5
- Develop a fully automatic method providing the best and worst layout content configuration for a given grid-based template and input images, accounting for users’ requirements.

# A PICTURE WORTHS MORE THAN A THOUSAND WORDS…



Pick up 20 images over the Internet and from a publicly available datasets



The images are with 2-4 salient objects against a homogenous background

Objects within the images represent our AOIs (areas of interest)

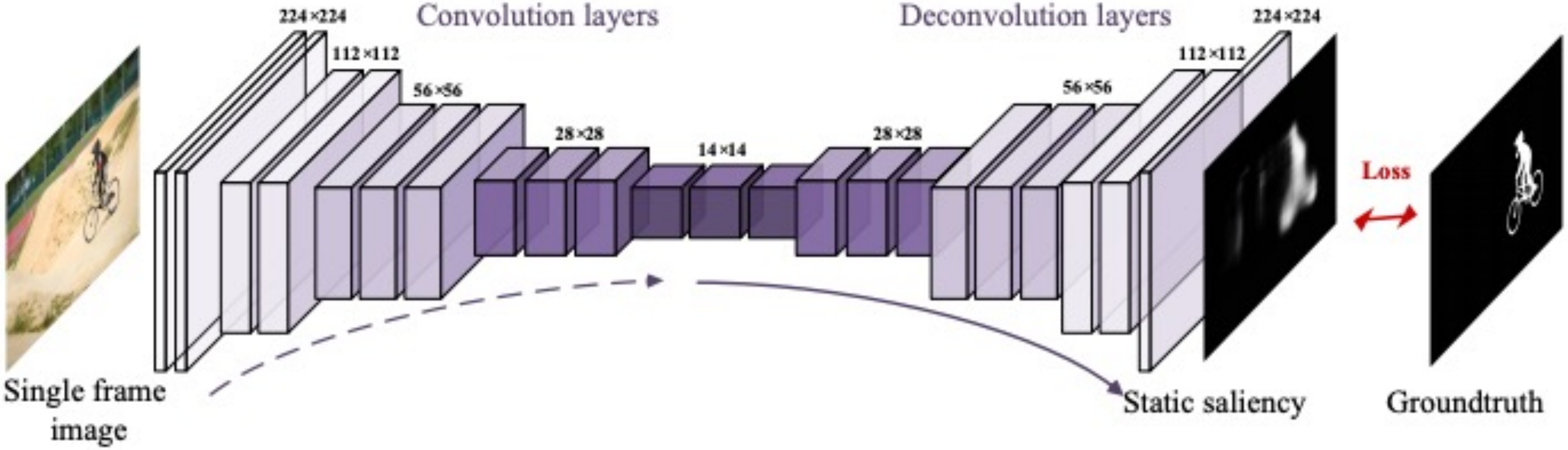
In our work we consider images that consist of different objects that stands out of the background.
What above takes down to a question: “What Saliency Model best suits the extraction of salient regions?”

# OBJECT-ORIENTED VISUAL SALIENCY MODEL

In this work saliency maps are extracted with a Deep Learning model developed by Wang et al. [4].

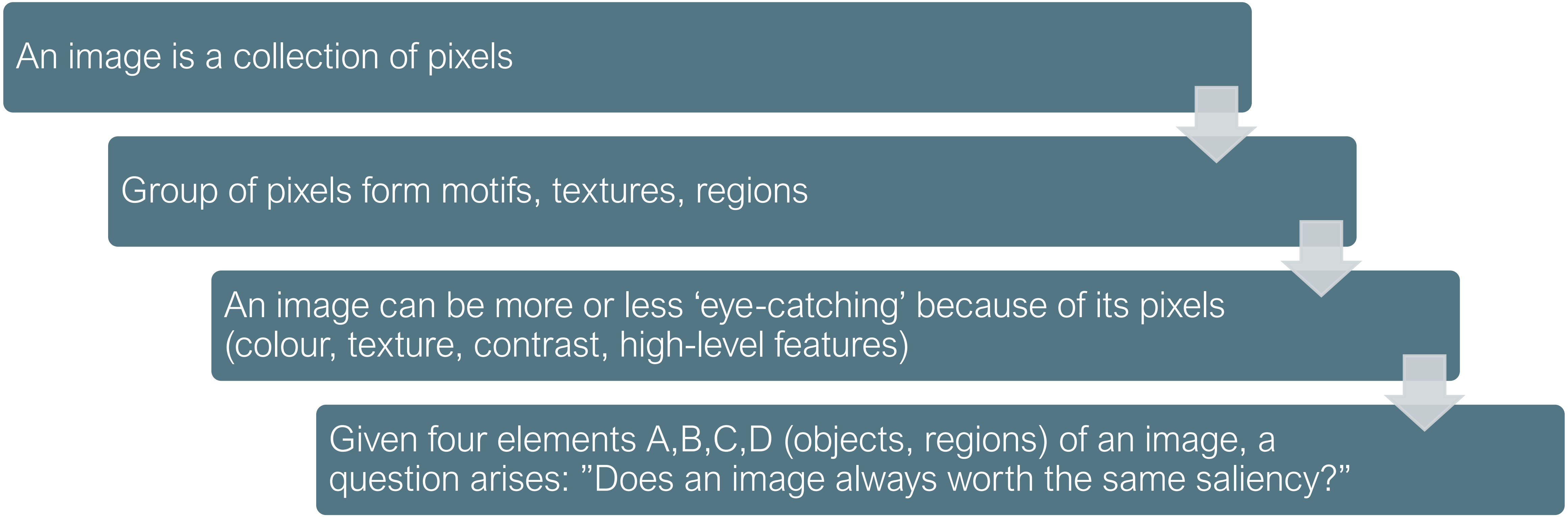
The saliency model is trained using a Fully Convolutional Neural Network trained over ImageNet and evaluated on DAVIS [5], that is a dataset with single visual instances labeled per each frame.

ImageNet is a dataset of over 15 million labelled high-resolution images belonging to roughly 22,000 categories [10]



The picture above is from reference [5]

# AN IMAGE IS MADE UP OF PIXELS



An image is a collection of pixels

Group of pixels form motifs, textures, regions

An image can be more or less ‘eye-catching’ because of its pixels (colour, texture, contrast, high-level features)

Given four elements A,B,C,D (objects, regions) of an image, a question arises: ”Does an image always worth the same saliency?”

# GRID-BASED LAYOUT PERMUTATIONS

**Premise:** For a given image layout, for example a 2 by 2 grid-based layout, a number of 4! spatial permutations are given (it adds up to 24 spatial permutations).

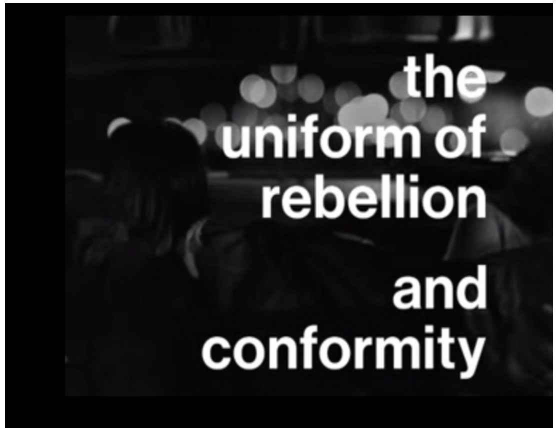
A

C

B

Anti-shock hard drive protection
ASUS Shockshield motion sensor* and anti-shock cushions protect your hard drive from damage
*Subject to model and system configuration

D



24 Permutations

A,B,C,D

A,C,B,D

A,C,D,B

A,D,B,C

A,D,C,B

…..

…..

…..

…..

…..

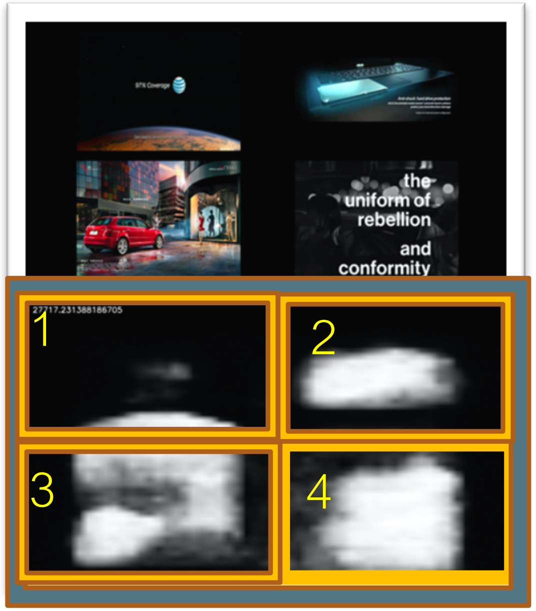
D,C,B,A

# PROPOSED METHOD - PREMISE

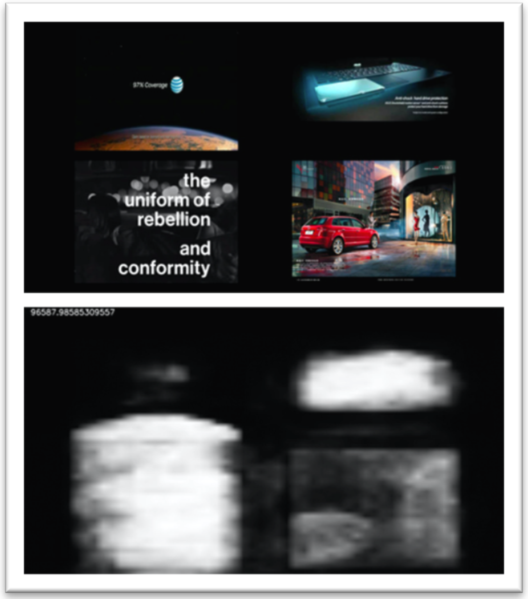
Some regions, such as the one with a red car, show different local saliency maps across different spatial permutations.

Saliency Maps are extracted by using a **deep learning-based solution**[4] evaluated over an object-oriented image and video dataset called **DAVIS** [5].

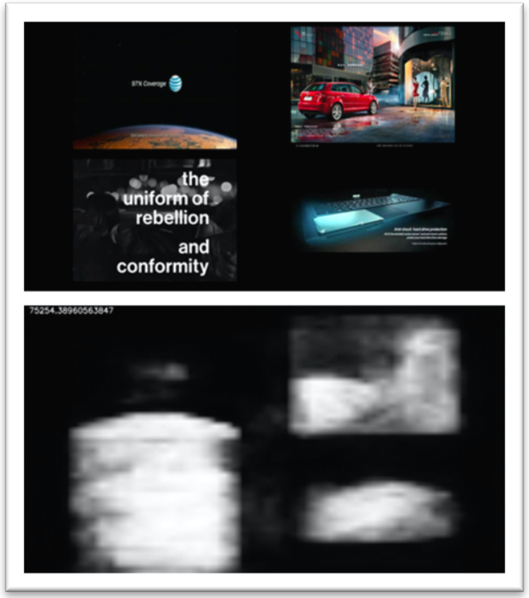
Experiments show different saliency ‘**behaviours**’ of the same regions whose an image consists of.



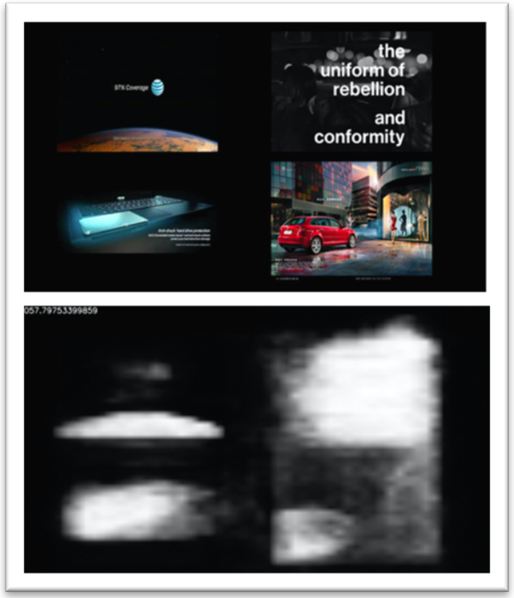
Permutation 1



Permutation 2

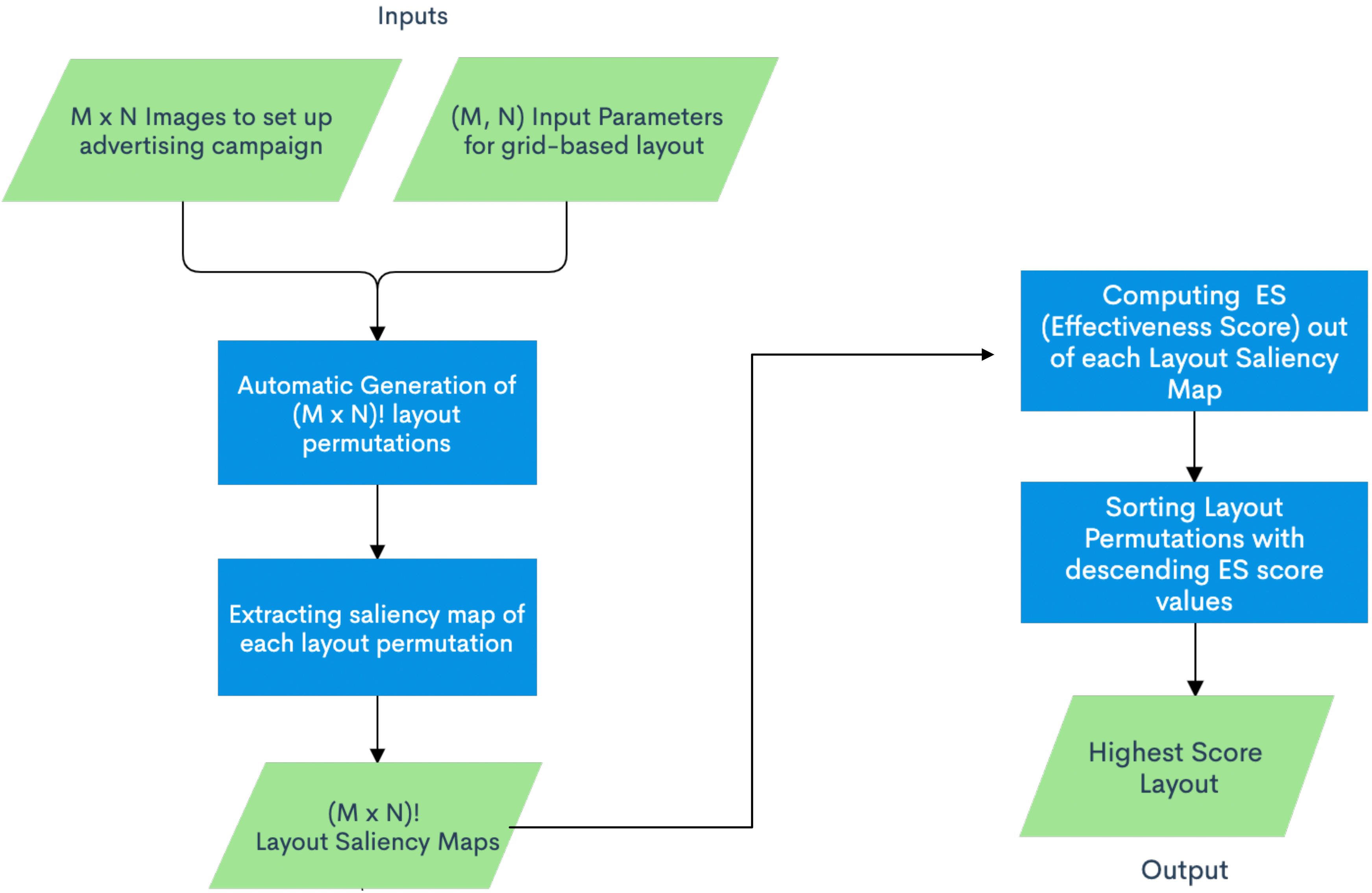


Permutation 3



Permutation 24

# PROPOSED METHOD – ALGORITHM (FLOW-CHART)



Inputs

M x N Images to set up advertising campaign

(M, N) Input Parameters for grid-based layout

Automatic Generation of (M x N)! layout permutations

Extracting saliency map of each layout permutation

(M x N)! Layout Saliency Maps

Computing ES (Effectiveness Score) out of each Layout Saliency Map

Sorting Layout Permutations with descending ES score values

Highest Score Layout

Output

# PROPOSED METHOD

For a given layout made up of *M · N* images, the ’behaviour’ of the overall layout saliency is studied by analysing the varying number of salient pixels on each of the M ·N images.

In greater detail, the **inverse of the relative variance** of local saliency maps is employed as ***ES*** (**Effectiveness Score**) [9].

In equation (1) ***ES*** is the ratio between the absolute mean and variance of *NMSP*$_k$ with k = 1,...,(M · N).

*NMSP*$_k$ stands for Number of Most Salient Pixels of each image in the $k_{th}$ layout content permutation.

$$ES_{(i)} = \frac{\left|\mu(NMSP_k(Layout_{(i)})\right|}{\sigma(NMSP_k(Layout_{(i)}))^2} \qquad (1)$$

$$k = [1, ..., (M \cdot N)] \quad i = \{1, ..., (M \cdot N)!\}$$

# PROPOSED METHOD

For a given layout with $M \cdot N$ images, $NMSP_h$ is the number of the most salient pixels in the local saliency map $LSM_{(h)}$ of the $h^{th}$ image (eq. 2)

$$NMSP_{(h)} = \sum_{i,j \,\epsilon\, Im} LSM_{(h)}(i,j) \geq th \qquad (2)$$

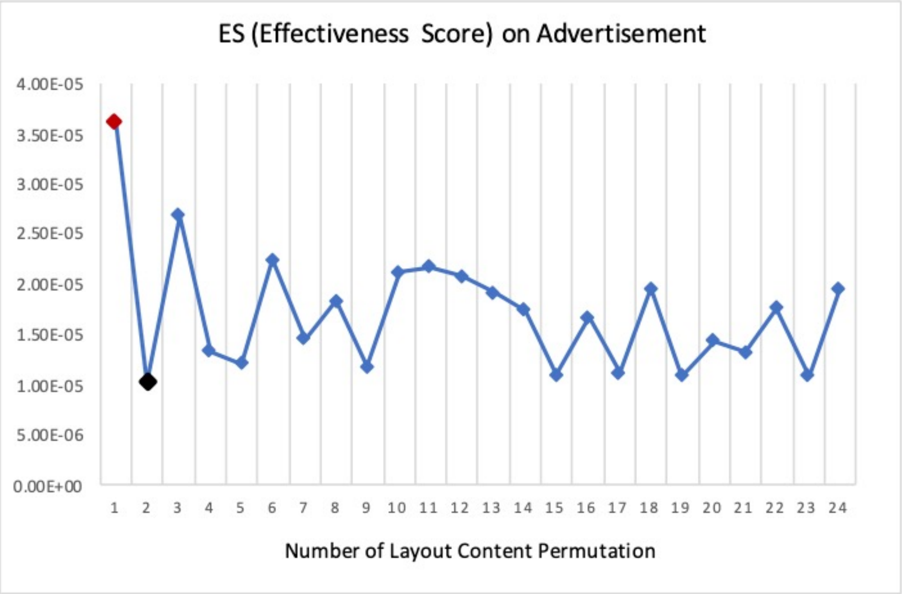
Each Layout content permutation is the union of $M \cdot N$ images $Im_{i'}$ as in equation 3

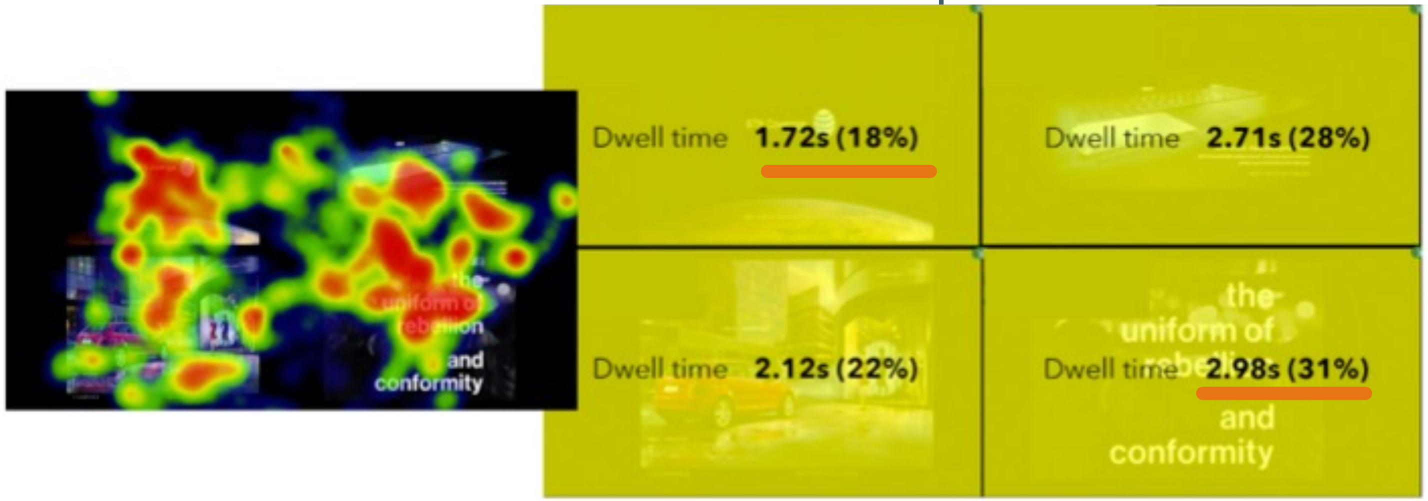
$$Layout = \bigcup_{i'=1}^{M \cdot N} Im_{i'} \qquad (3)$$

The layout showing the **highest score** is the output of the proposed method.

# EXPERIMENTAL RESULTS

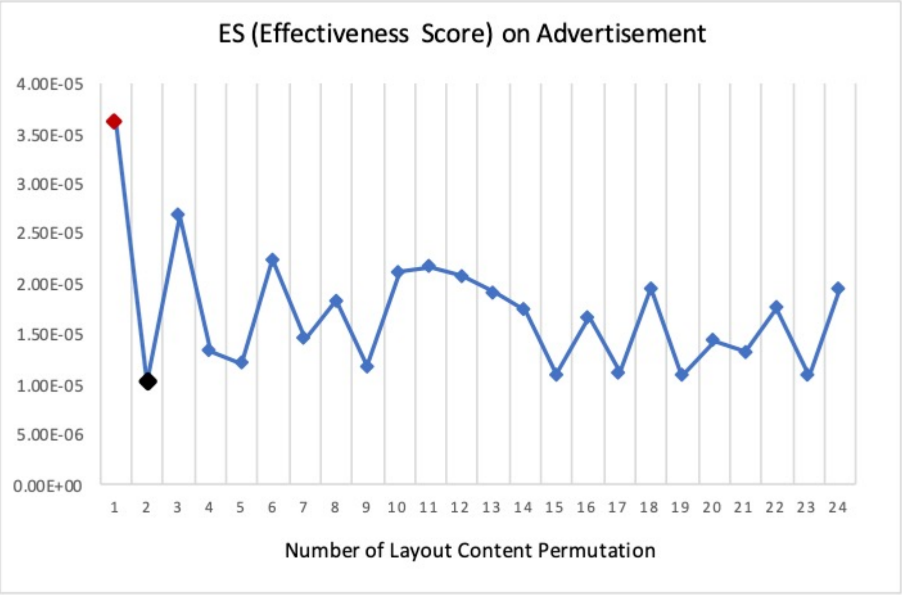
Highest ranking layout configuration

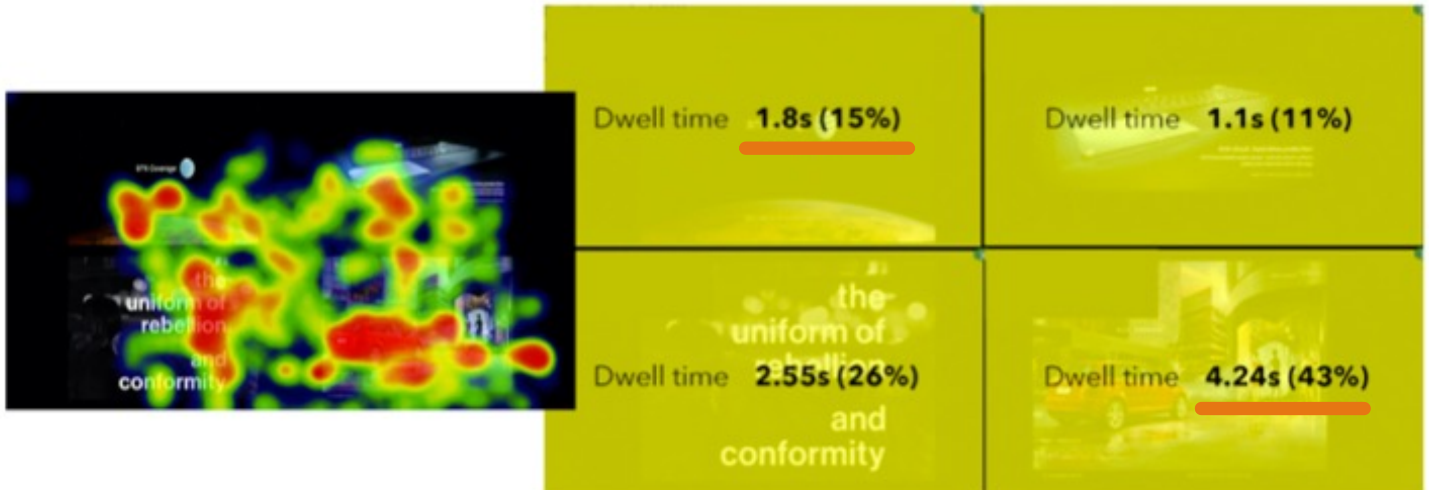




# EXPERIMENTAL RESULTS

Lowest ranking layout configuration





# VALIDATION THROUGH EYE-TRACKING SESSIONS

Gazerecorder[7], a Webtool for webcam-based eye-tracking, was used to carry out the validation of the proposed method

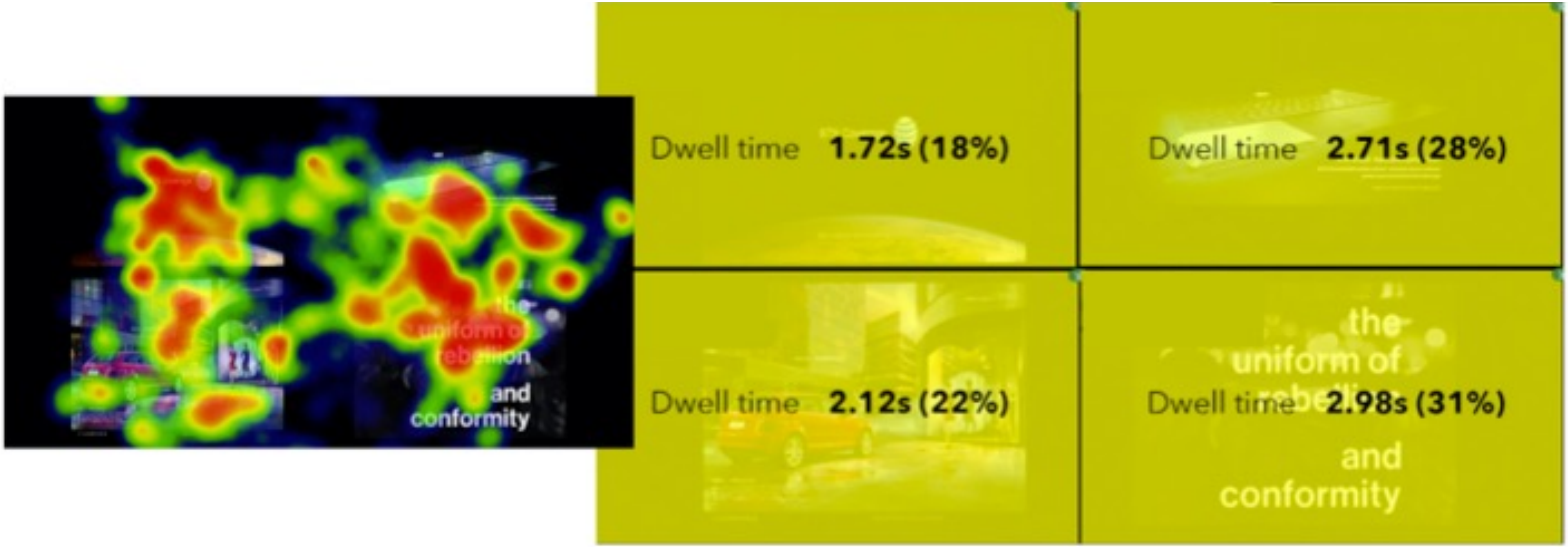
20 participants in the age range [25-40] were shown the layout content permutations with highest and lowest ES value of 5 graphical contents with images out of dataset [6];

Each Image is shown for 10 seconds;

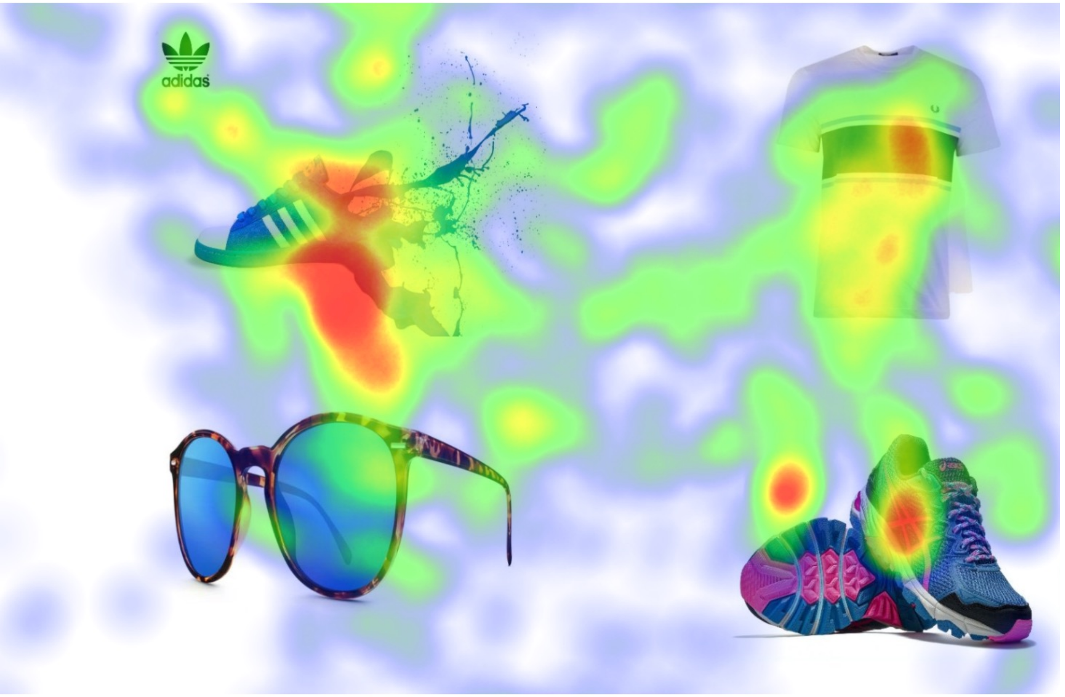
Heatmaps and Dwell times are collected as shown below

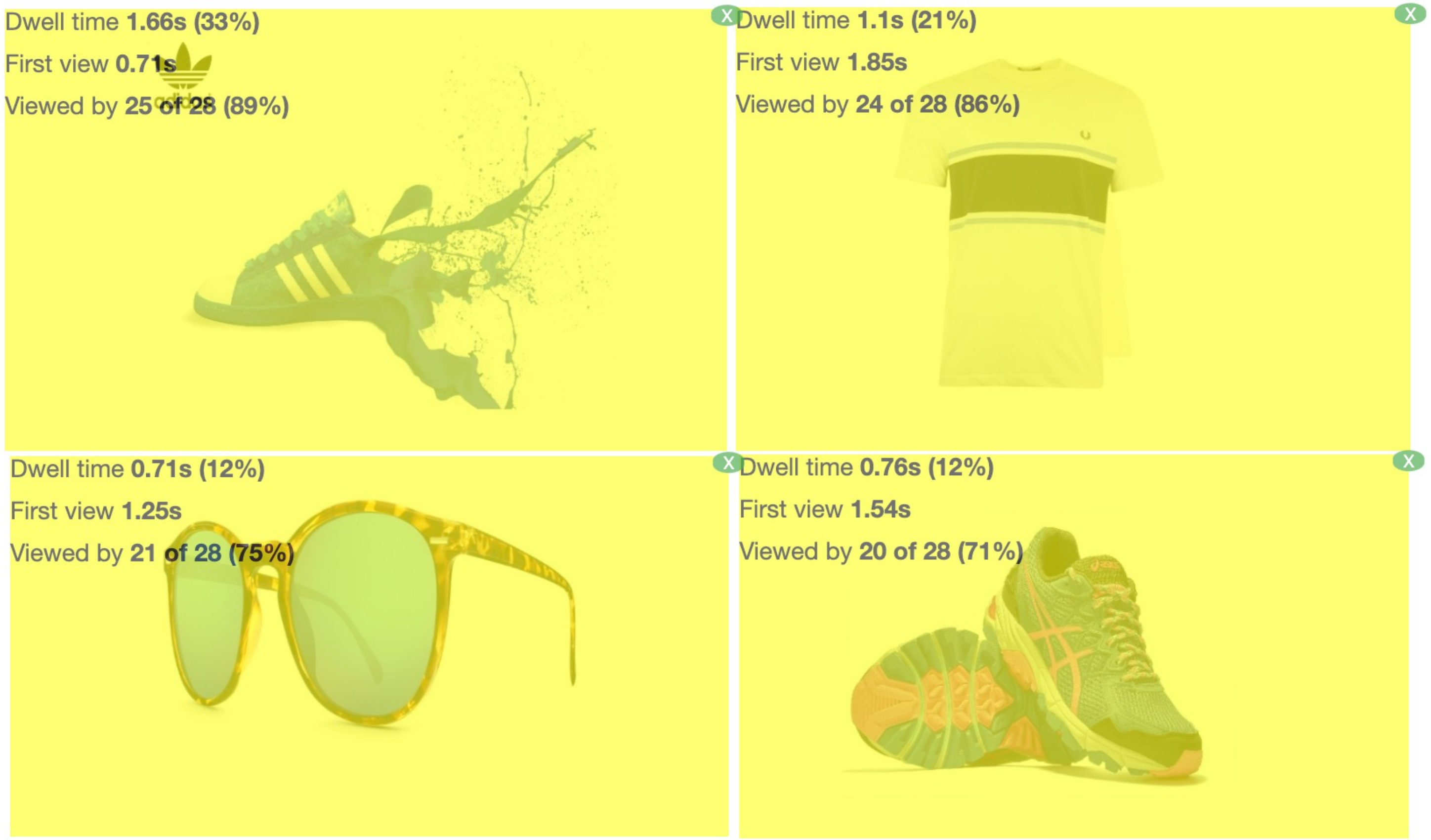
Experiments were conducted to assess consistency between our saliency-based results and webcam eye-tracking session data

# EYE-TRACKING SESSION (CAMPAIGN 1)



# EYE-TRACKING SESSION (CAMPAIGN 2)





Dwell time **1.66s (33%)**
First view **0.71s**
Viewed by **25 of 28 (89%)**

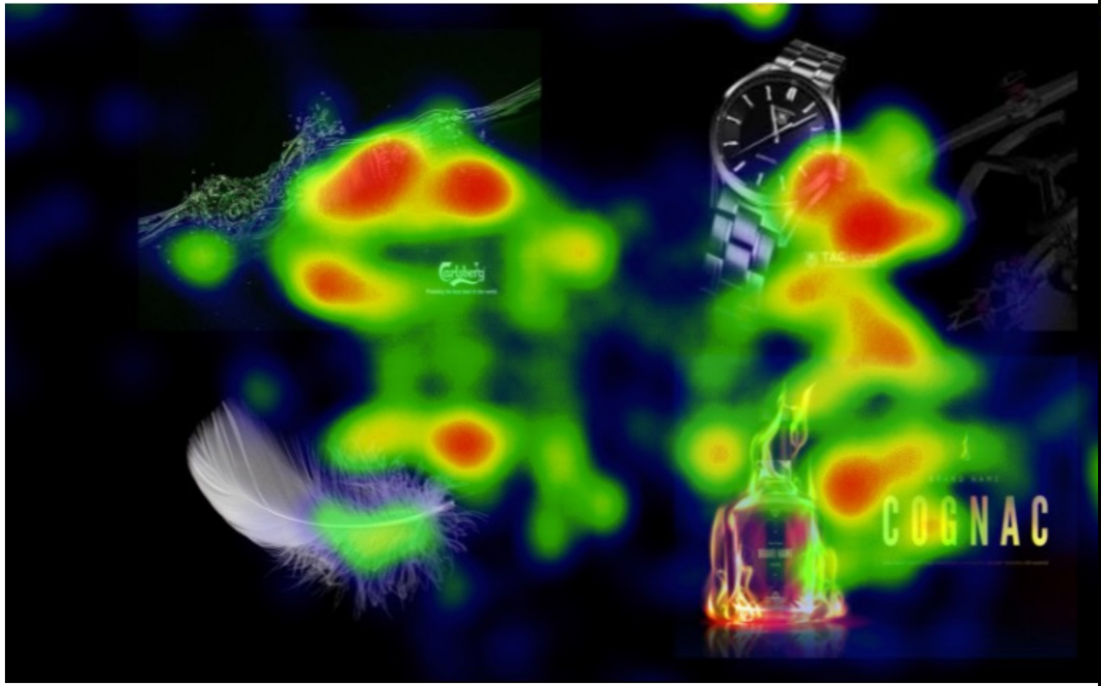
Dwell time **1.1s (21%)**
First view **1.85s**
Viewed by **24 of 28 (86%)**

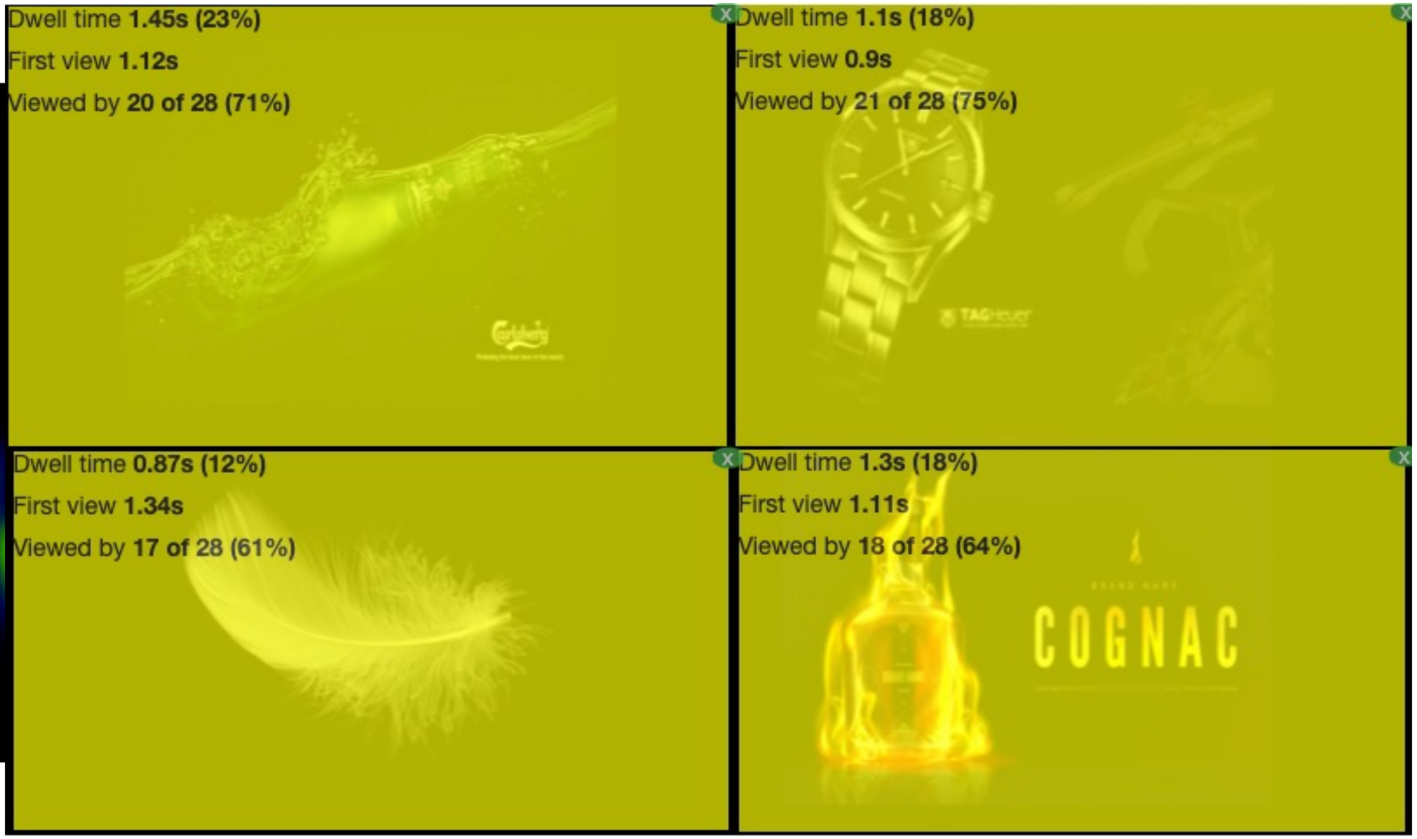
Dwell time **0.71s (12%)**
First view **1.25s**
Viewed by **21 of 28 (75%)**

Dwell time **0.76s (12%)**
First view **1.54s**
Viewed by **20 of 28 (71%)**

One of the 4! Spatial permutations out of the 4 images (the four quadrants)

# EYE-TRACKING SESSION (CAMPAIGN 3)




Dwell time **1.45s (23%)**

First view **1.12s**

Viewed by **20 of 28 (71%)**

Dwell time **1.1s (18%)**

First view **0.9s**

Viewed by **21 of 28 (75%)**

Dwell time **0.87s (12%)**

First view **1.34s**

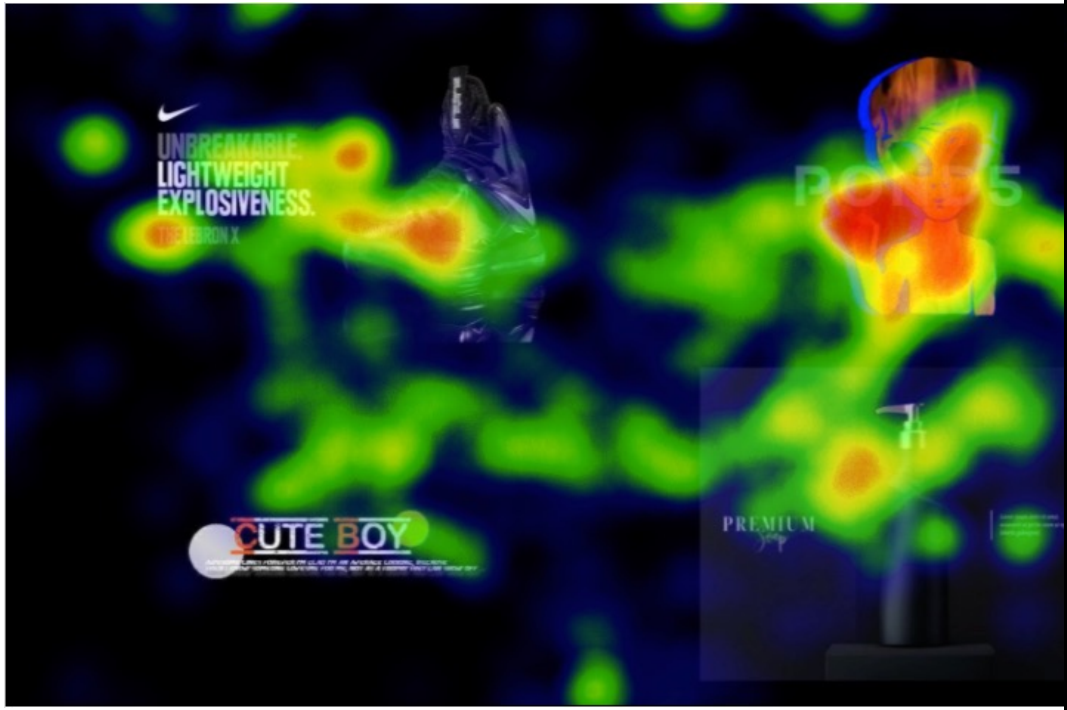
Viewed by **17 of 28 (61%)**

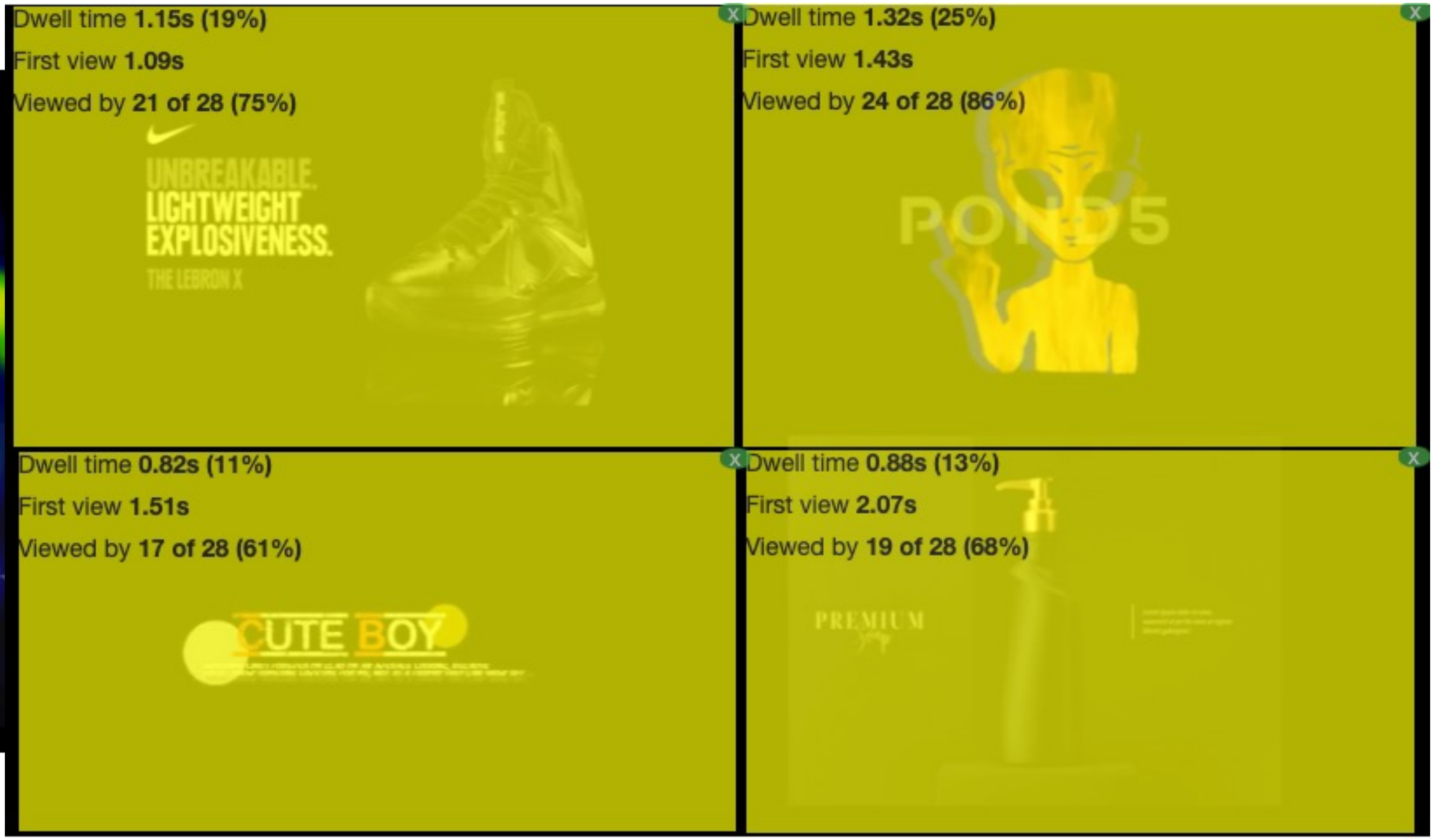
Dwell time **1.3s (18%)**

First view **1.11s**

Viewed by **18 of 28 (64%)**

# EYE-TRACKING SESSION (CAMPAIGN 4)





Dwell time **1.15s (19%)**

First view **1.09s**

Viewed by **21 of 28 (75%)**

Dwell time **1.32s (25%)**

First view **1.43s**

Viewed by **24 of 28 (86%)**

Dwell time **0.82s (11%)**

First view **1.51s**

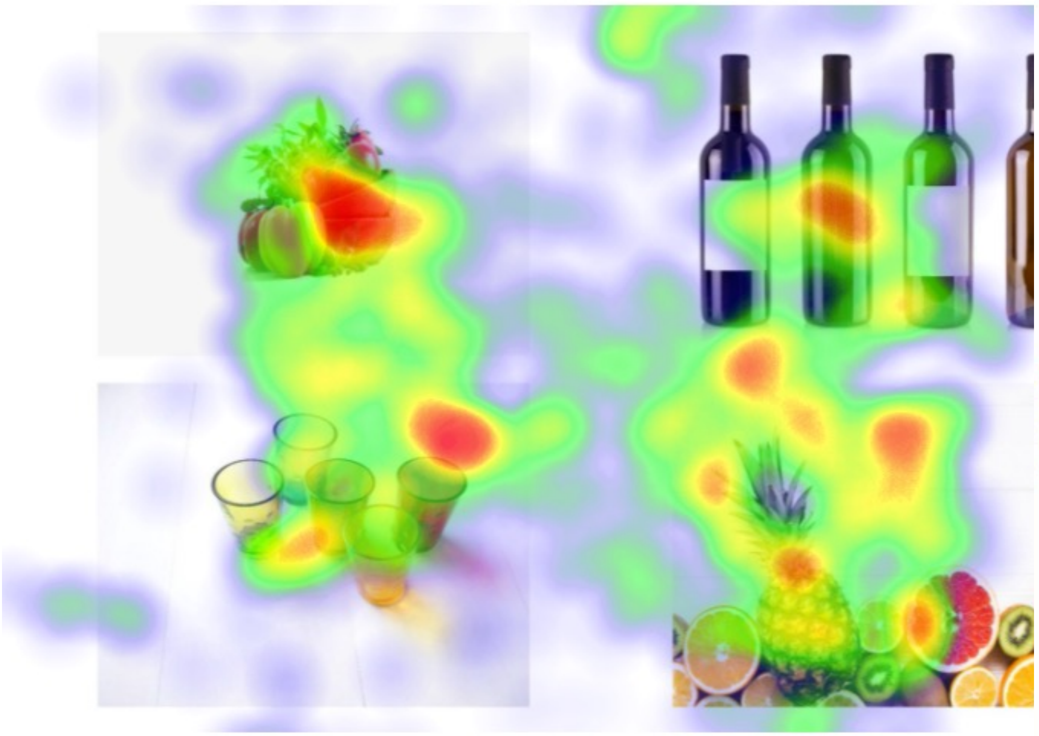
Viewed by **17 of 28 (61%)**

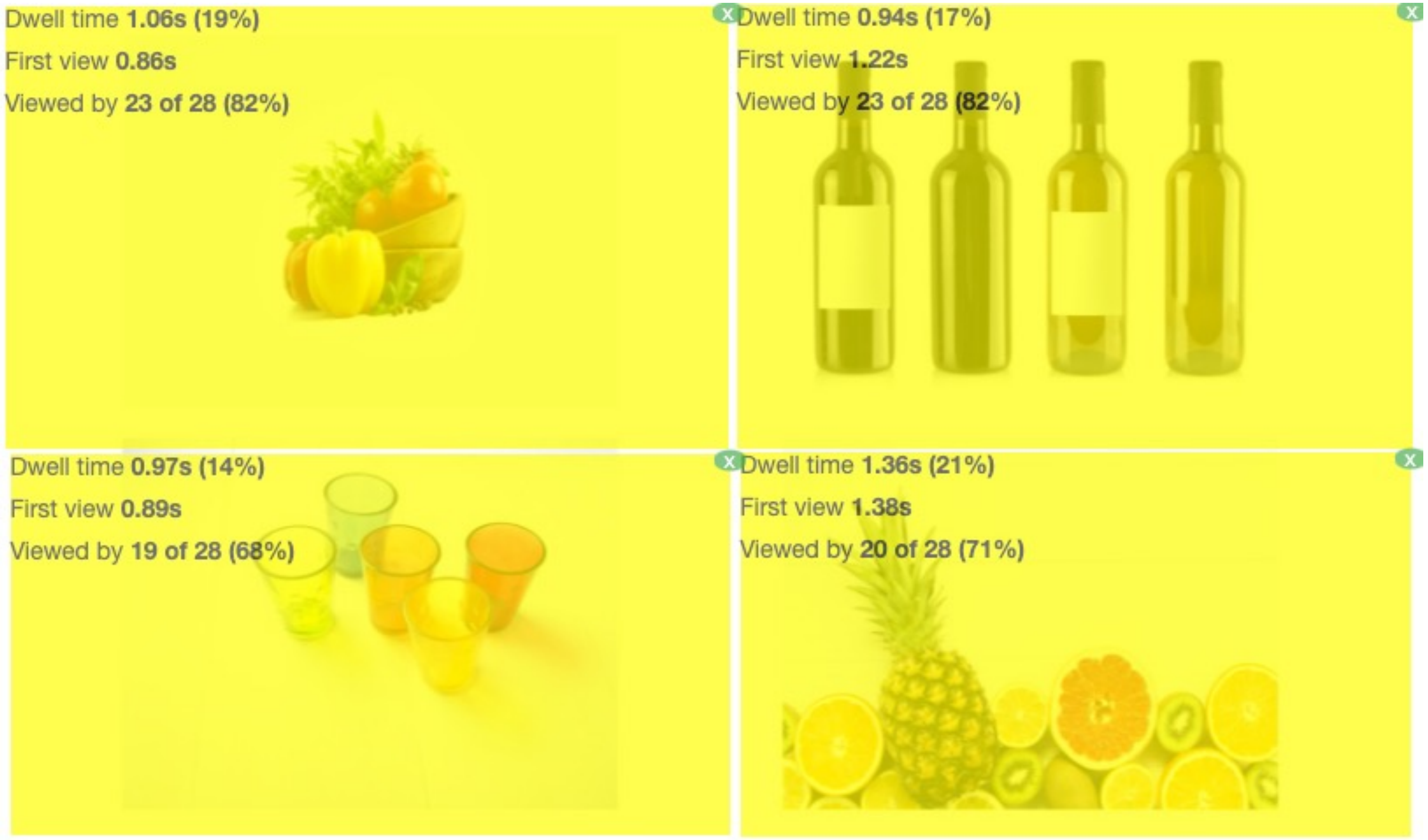
Dwell time **0.88s (13%)**

First view **2.07s**

Viewed by **19 of 28 (68%)**

# EYE-TRACKING SESSION (CAMPAIGN 5)





Dwell time 1.06s (19%)

First view 0.86s

Viewed by 23 of 28 (82%)

Dwell time 0.94s (17%)

First view 1.22s

Viewed by 23 of 28 (82%)

Dwell time 0.97s (14%)

First view 0.89s

Viewed by 19 of 28 (68%)

Dwell time 1.36s (21%)

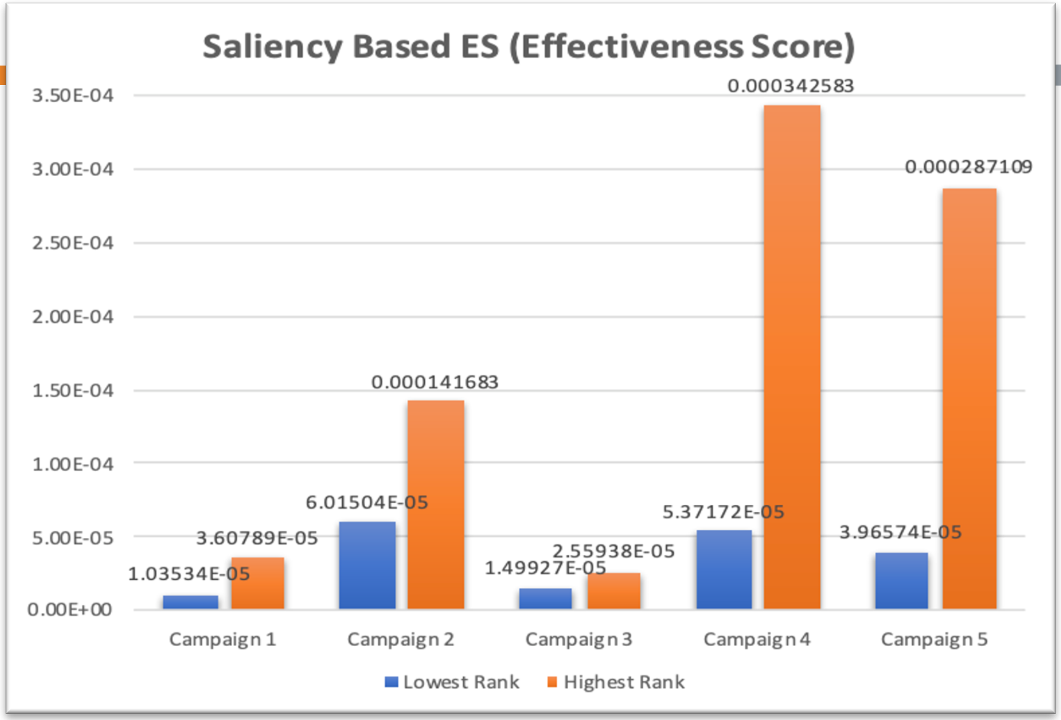
First view 1.38s

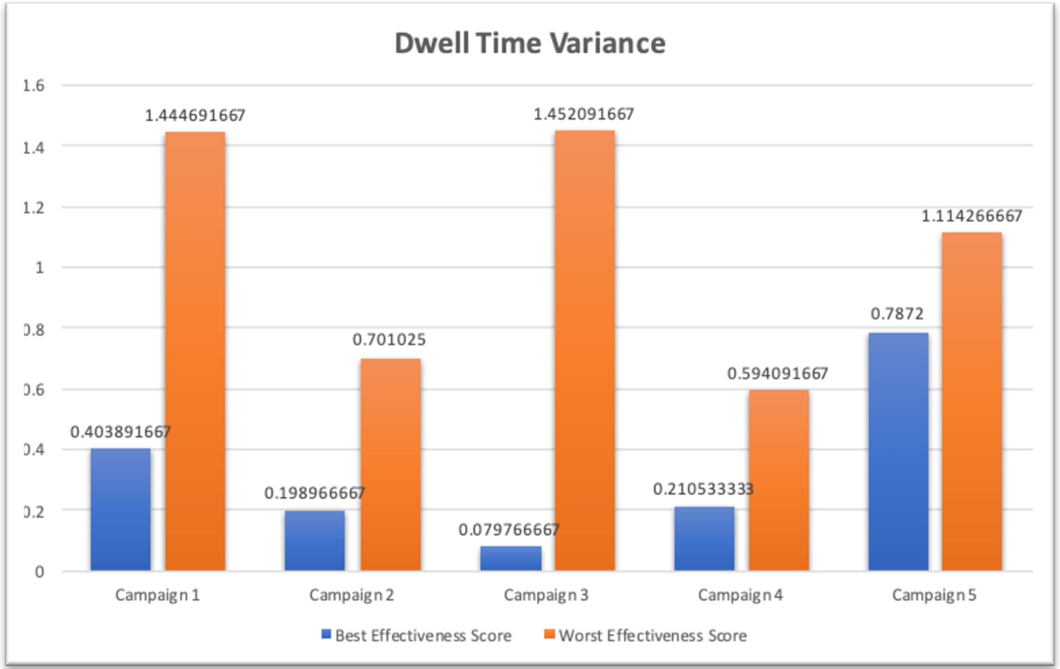
Viewed by 20 of 28 (71%)

# EXPERIMENTAL RESULTS

The graph on the top shows the highest and lowest ES scores for each advertising campaign.

Images scoring best and worst ES values also show more balanced dwell times (graph on the bottom)





# EXPERIMENTAL RESULTS (SETTINGS)

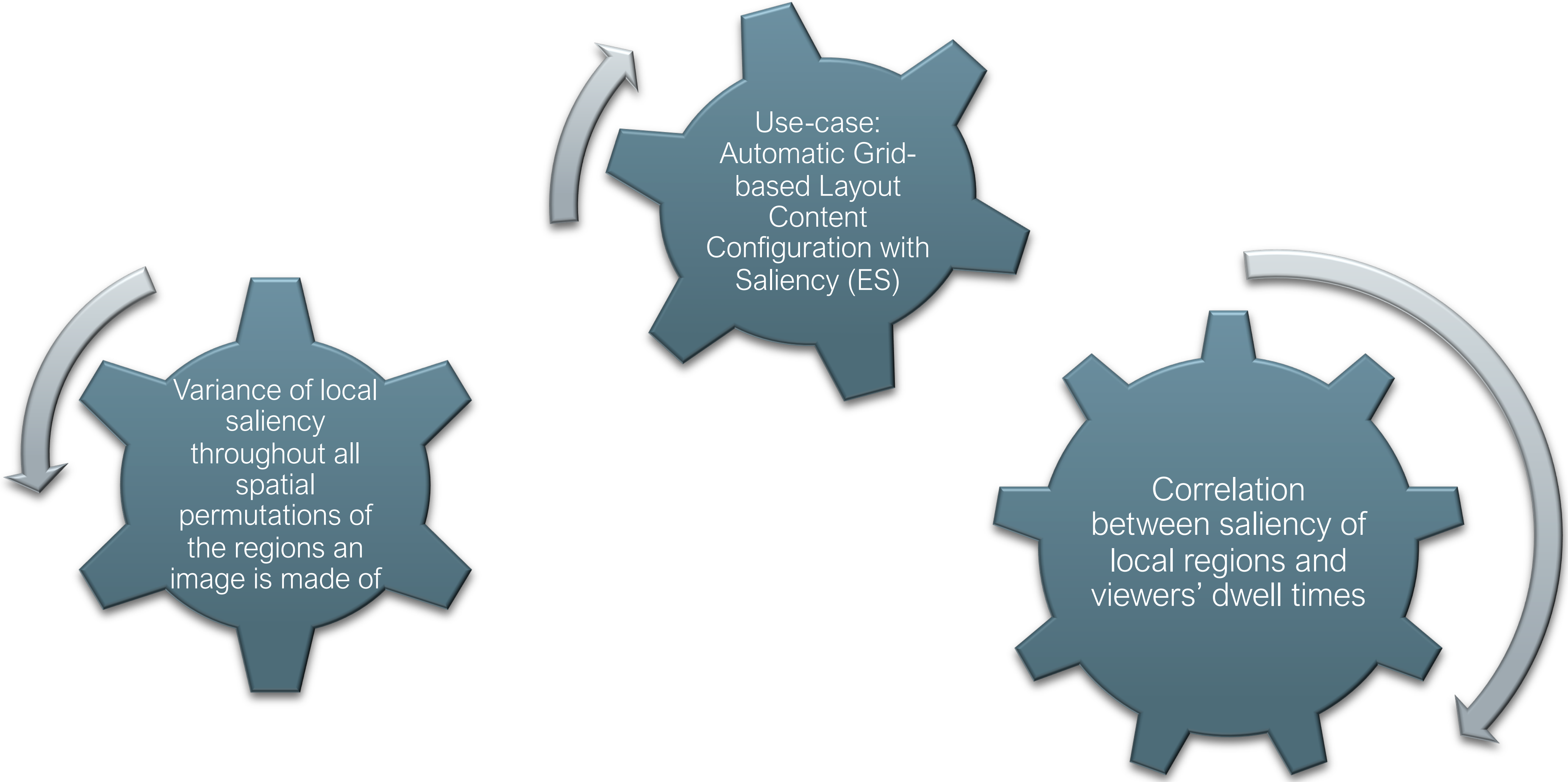
13-inch Mac-book Pro with 16 GB of RAM, 2.4 GHz Quad-Core Intel Core i5, Intel Iris Plus Graphics 655 1536 MB;

Average running time on 2-by-2 grid layouts is 40 seconds;

Python 3.8.0

TensorFlow 2.4.0 – Deep Learning Python Framework

# CONCLUSIONS AND FUTURE WORKS

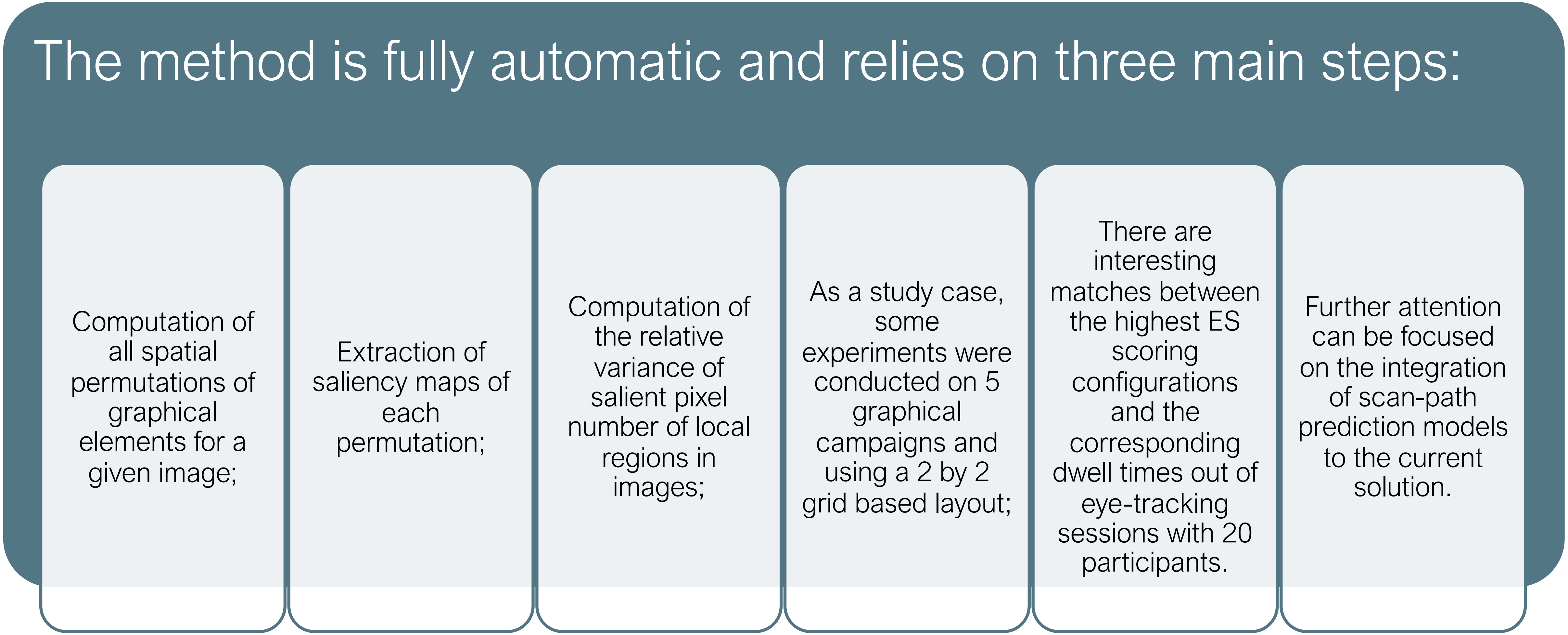


Variance of local saliency throughout all spatial permutations of the regions an image is made of

Use-case: Automatic Grid-based Layout Content Configuration with Saliency (ES)

Correlation between saliency of local regions and viewers’ dwell times

# CONCLUSIONS AND FUTURE WORKS



The method is fully automatic and relies on three main steps:

- Computation of all spatial permutations of graphical elements for a given image;
- Extraction of saliency maps of each permutation;
- Computation of the relative variance of salient pixel number of local regions in images;
- As a study case, some experiments were conducted on 5 graphical campaigns and using a 2 by 2 grid based layout;
- There are interesting matches between the highest ES scoring configurations and the corresponding dwell times out of eye-tracking sessions with 20 participants.
- Further attention can be focused on the integration of scan-path prediction models to the current solution.

# REFERENCES

[1] Dennis, Charles and Brakus, J Jovsko and Gupta, Suraksha and Alamanos, Eleftherios The effect of digital signage on shoppers' behavior: The role of the evoked experience, Journal of Business research, vol. 67, no.11, pages 2250--2257, (2014), Elsevier

[2] Li, Jia and Gao, Wen, Visual saliency computation: A machine learning perspective, vol. 8408, (2014), Springer

[3] Zhang, Jianming and Malmberg, Filip and Sclaroff Stan: Visual Saliency: From Pixel-Level to Object-Level Analysis. (2019). Springer

[4] Wang, W., Shen, J., Shao, L.: Video salient object detection via fully convolutional networks. IEEE Transactions on Image Processing 27(1), 38–49 (2017)

[5] Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine- Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 724–732 (2016)

[6] Hussain, Z., Zhang, M., Zhang, X., Ye, K., Thomas, C., Agha, Z., Ong, N., Ko- vashka, A.: Automatic understanding of image and video advertisements. In: Pro- ceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1705–1715 (2017) Url: http://people.cs.pitt.edu/kovashka/ads/

[7] Deja, S.: Gazerecorder. https://api.gazerecorder.com/

[8] A. Bruno, F. Gugliuzza, R. Pirrone and E. Ardizzone, "A Multi-Scale Colour and Keypoint Density-Based Approach for Visual Saliency Detection," in *IEEE Access*, vol. 8, pp. 121330-121343, 2020, doi: 10.1109/ACCESS.2020.3006700.

[9] Bruno, Alessandro and Lancette, Stéphane and Zhang, Jinglu and Moore, Morgan and Ward, Ville P and Chang, Jian, A Saliency-Based Technique for Advertisement Layout Optimisation to Predict Customers' Behaviour, ICPR Workshops (2), pages 495--507, 2020

[10] ImageNet: https://image-net.org/

The End