Journal of Environmental Psychology 92 (2023) 102110

Contents lists available at ScienceDirect

Journal of Environmental Psychology

journal homepage: www.elsevier.com/locate/jep

Saliency prediction in 360° architectural scenes: Performance and impact of daylight variations

Caroline Karmann [a,b,*], Bahar Aydemir [c], Kynthia Chamilothori [a,d], Seungryong Kim [c,e], Sabine Süsstrunk [c], Marilyne Andersen [a]

[a] Laboratory of Integrated Performance in Design (LIPID), École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[b] Laboratory of Architecture and Intelligent Living (AIL), Karlsruhe Institute of Technology (KIT), Germany
[c] Image and Visual Representation Lab (IVRL), École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[d] Human-Technology Interaction (HTI) Group, Technical University Eindhoven (TU/e), Netherlands
[e] Computer Vision Laboratory (CVLAB), Korea University, South Korea

ARTICLE INFO

Handling Editor: Chiara Burattini

Keywords:
Visual attention
Saliency prediction modelling
Head tracking
Indoor spaces
Daylight

ABSTRACT

Saliency models are image-based prediction models that estimate human visual attention. Such models, when applied to architectural spaces, could pave the way for design decisions where visual attention is taken into account. In this study, we tested the performance of eleven commonly used saliency models that combine traditional and deep learning methods on 126 rendered interior scenes with associated head tracking data. The data was extracted from three experiments conducted in virtual reality between 2016 and 2018. Two of these datasets pertain to the perceptual effects of daylight and include variations of daylighting conditions for a limited set of interior spaces, thereby allowing to test the influence of light conditions on human head movement. Ground truth maps were extracted from the collected head tracking logs, and the prediction accuracy of the models was tested via the correlation coefficient between ground truth and prediction maps. To address the possible inflation of results due to the equator bias, we conducted complementary analyses by restricting the area of investigation to the equatorial image regions. Although limited to immersive virtual environments, the promising performance of some traditional models such as GBVS360eq and *BMS360eq* for colored and textured architectural rendered spaces offers us the prospect of their possible integration into design tools. We also observed a strong correlation in head movements for the same space lit by different types of sky, a finding whose generalization requires further investigations based on datasets more specifically developed to address this question.

1. Introduction

The visual content of our surroundings can influence our perception and behavior in a space. This topic has long preoccupied artists (Balbi et al., 2016), while architects have often speculated about the visual impact of form on human eye movement (Arnheim, 1965, 1977). Through specific arrangements of architectural features (e.g., walls, columns, openings), or by using selected design principles (e.g., Gestalt principles) as well as other representation strategies (e.g., light washing, contrasts, shadow interplay), hierarchies can be created (or assumed) in a spatial composition, anticipating that attention may be drawn to certain elements over others, and that the observer's gaze might be guided. However, these theories mostly rest on assumptions about how particular geometric arrangements are actually perceived (Weber et al., 1995), which remains an open debate.

Visual attention is crucial in defining human experience and behavior in an environment. Both in outdoor (Caduff & Timpf, 2008; Koseoglu & Onder, 2011), and indoor environments (Dong et al., 2020; Wang et al., 2018), landmarks, i.e., prominent spatial features in an environment, have been shown to be crucial in spatial legibility and wayfinding. The integration of visual attention into legibility research holds promise as demonstrated by Wang et al. (2019) whose quantification method directly involves human gaze patterns. Elements of high visual attention also appear to play a critical role in our appraisal of environments. Landscape objects that were found to induce high visual attention through gaze behavior were also those reported by

* Corresponding author. Englerstraße 7Building 20, Karlsruhe, Germany.
E-mail address: caroline.karmann@kit.edu (C. Karmann).

https://doi.org/10.1016/j.jenvp.2023.102110
Received 2 September 2022; Received in revised form 22 April 2023; Accepted 27 June 2023
Available online 5 September 2023

participants when asked what they liked or disliked about the landscape (Cottet et al., 2018). The overall visual information present in a scene has also been suggested to be a defining factor in predicting human preference towards a landscape. Kaplan and Kaplan (1989) suggested four elements that can predict landscape preference, two relating to spatial understanding (legibility and coherence) and two relating to the desire of exploration (mystery and complexity). In the field of restorative environments, Van der Jagt et al. (2017) found that the amount of important visual information in a scene (measured through the speed of scene categorization as natural or built) influenced cognitive restoration, with scenes low in important visual information leading to higher restoration, which according to the authors is in alignment with the concept of soft fascination in the attention-restoration theory (ART) (Kaplan, 1995). The identification of these visually important (or "salient") elements in a scene is therefore of great interest for multiple applications with high societal impact, be it spatial legibility, appraisal or cognitive restoration potential.

Salience is defined as the quality of being particularly noticeable or important. This term has been adopted among vision scientists to describe which visual features within a given scene might attract more of our attention. Saliency modelling, i.e., the prediction of which visual features are salient, has gained a high popularity within the past two decades as a result of increasing capabilities in eye- and head-tracking technologies and advances in vision, neuroscience and computer science, leading to increasing prediction performance (Borji, 2018). The overarching goal of this paper is to test and discuss the reliability and applicability of existing saliency models in daylit architectural design appraisal. For this purpose, we relied on three datasets from architecture and computer vision experiments and tested different types of validated and publicly available models. The specifics of our research objectives are further outlined in section 1.4. In the following sections, we will provide a brief review of the literature regarding saliency prediction modelling, its ties to visual sentiment analysis, and visual attention in architectural design.

1.1. Saliency prediction modelling

Saliency prediction models tell us where people are most likely to devote their perceptual and cognitive resources when they look at a given scene. These models can be classified as.

- "**Traditional**" or "bottom-up" models, which are based on the assumption that human attention follows an exogenous process, and are driven by the recognition of low-level features such as contrasts, colors, directions, orientations. Such models focus on fast, involuntary, signal-driven and task-independent visual attention. Traditional models were the first prediction models developed for perspective images in the 1990s. Example of such models include the Itti, Koch, and Niebur model (IKN) (Itti et al., 1998) and the Graph-Based Visual Saliency model (GBVS) (Harel et al., 2007).
- "**Deep-learning**" (DL) models, which are built from dataset of existing viewing logs (a.k.a., benchmarks) through machine learning methods relying on convolutional neural networks (CNNs). As they rely on training based on existing data, the generalizability of these models when applied to new types of image stimuli, remains an outstanding issue. Examples of such models include DeepGaze I and II (Kümmerer et al., 2014; Kümmerer et al., 2016) and Sal-CEDN (Kroner et al., 2020).

The performance of saliency models is determined by comparing the prediction with ground truth maps.

Ground truth maps are constructed from the aggregated viewing logs of multiple observers, ignoring any temporal fixation information. Both saliency prediction and ground truth are presented as pixelated images (i.e., "maps") with values ranging from 0 to 1, where high pixel values represent a higher probability of fixations. Ground truth maps result from participants' observations of scenes. Participants are usually asked to look in a "free-viewing" manner (without questions about the images) at a given image, a. k.a., the visual stimuli shown on a screen. For perspective images, the viewing time is often very short, ranging from 2 to 5 s (Borji & Itti, 2015; Bylinskii et al., 2015) and common methods to record the participants visual attention usually rely on camera-based **eye- or head-trackers** of varying degrees of precision and intrusiveness (e.g., from webcams to wearable eye trackers). The emergence of immersive virtual environments (IVE) utilizing Head Mounted Displays (HMD) brought new opportunities to both displaying the visual stimuli and to recording participant's attention. For these 360° (or omnidirectional) images, the viewing time is typically around 30 s per scene (Rai et al., 2017; Sitzmann et al., 2018). The recording is based on eye and/or head tracking sensors, usually integrated directly into the HMD.

For omnidirectional images, the ground-truth maps are eventually converted to equirectangular representations. As such, the procedure used to develop the models and to assess their reliability remains based on 2D information (same as for perspective images). This operation however comes with distortions, as straight lines become curves and as the upper and lower extremities of the scenes – which are "points" in a 360° environment – become "lines" once converted to equirectangular maps. These distortions are commonly accounted for in the treatment of visual information. Raw head- or eye-tracking data are then commonly post-processed to distinguish saccades (e.g., rapid head or eye movements) from fixations, with multiple possible approaches for this filtering. In addition, these recording methods by default reduce visual attention to a series of points, but since attention occurs over regions, researchers typically apply two-dimensional Gaussian filters, resulting in smooth pixel maps (Bylinskii et al., 2016). These protocols have proven to be reliable in terms of visual attention compared to real-world conditions (Foulsham et al., 2011).

Visual attention shows known biases for both perspective and 360° images. For perspective images, our gaze tends to be attracted both towards the center of the scene, leading to a so-called "**central bias**" (Harel et al., 2007). For 360° images, our gaze tends to be attracted towards the equator of the scene, leading to a so-called "**equatorial bias**" (Sitzmann et al., 2018). Corrections to account for these biases include Gaussian and Laplacian functions that are usually integrated within the prediction models.

1.2. Saliency and visual sentiment analysis

In order to understand and predict how images induce human emotions, recent research uses visual sentiment analysis, an extension of the sentiment analysis that originally focused on text, to classify the polarity of an image (i.e., positive or negative) (Truong & Lauw, 2017). Fan et al. investigated the link between visual sentiment and saliency with images of higher emotional potential (Fan et al., 2017, 2018) and found that negative sentiments were elicited by the focal region without a notable influence of contextual information, whereas positive sentiments were influenced by both focal and contextual information. In a subsequent study, the authors also found that altering the semantic content of an image (by rotating it, converting it into grayscale or adding blur filters) significantly altered the sentiment recorded, as participants' judgments relied more on low-level features. Zheng et al. (2017) focused on the semantic information captured in the region of interest (instead of the whole picture). They found that images containing outstanding man-made objects or human faces, or that are indoors and closed, tend to express sentiment through their salient objects. She et al. (2020) also focused their analyses on image regions: the images were generally of higher emotional potential (taken from Flickr and Instagram) and the authors used automated image labels to annotate image regions. Their model was able to outperform existing ones. Based on these observations, a question arises: if only certain spatial regions (the most salient ones) can play a role in the observers' sentiments, which would these

regions be in the context of architectural images devoid of people, animals, or other particularly salient objects?

1.3. Visual attention in architectural spaces

Predicting viewing patterns in architecture can provide a powerful aid to manipulate visual distractions, improve way finding and safety, and support the design of compelling spaces. Weber et al. (1995, pp. 57–69) investigated how our visual experience is influenced by various formal-geometric characteristics (e.g., size, contrast, direction, symmetry, closure) and how these factors alter visual attention. Participants were shown photographs of simple architectural arrangements on a computer screen while their head were kept immobile and their eyes recorded with a camera. The authors concluded that elements indicating spatial depth receive special attention, that redundant elements draw less attention than solitary shapes, and that obliquely oriented shapes are more attractive that vertically and horizontally oriented objects. Fifteen years later, Hasse and Weber (2012) displayed façades on a computer screen and used a remote eye tracker with the aim to link aesthetic judgment to viewing behavior. They found that the compositional balance of a facade (that may or may not rely on symmetrical arrangements of shapes and openings) affected judgments of interest, but not judgments of beauty. Hollander et al. (2019), Noland et al. (2017) and Hollander et al. (2020) relied on similar methods to evaluate visual preferences of urban scenes and of ornament in design, and showed a tendency for longer fixations in more complex traditional designs.

The interplay between light, shadow, and architectural features have long been considered central in forming the identity of a space (Corrodi & Spechtenhauser, 2014) and people's experience of it (Köster, 2004; McCarter & Pallasmaa, 2012; Steemers, & Ann Steane, 2012). Architects use tools such as computer renderings or scale models to test the interaction of light and space in their designs and to share these design intentions with their clients (Leslie, 2003). Research in both real (Parpairi et al., 2002) and virtual environments (Rockcastle et al., 2017a) consistently shows that people appreciate variability in the daylight conditions indoors, while studies using projections (Abboushi et al., 2019) and virtual reality (Chamilothori et al., 2019; 2022b) show that the composition of light patterns in a space influences impressions of interest, and that the presence of large sun patches in one's field of view in a social context can even induce physiological responses. Being able to predict visual attention in architectural interiors, particularly in relation to its lighting conditions, would greatly advance our understanding of space features that drive human experience and behavior. Nevertheless, visual attention has scarcely been studied in relation to lighting. Vincent et al. (2009) who noted that light sources are highly visible but rather uninformative, questioned our visual attraction to these elements. In this study, which examined eye movements towards photographs depicting outdoor scenes at dusk with artificial light sources, the authors found that luminance contrast and light sources played a minor role in human fixations and that observers were more likely to look near lights rather than directly at them. They also noted that the visual system commonly neglects highly visible cues in favor of less visible object information. However, this study was conducted using photographs of primarily outdoor scenes at dusk, which raises questions on the transferability of the results in real daylit environments where the human visual system is likely to be influenced by greater light levels. In line with this comment, Sarey Khanie et al. (2017) examined eye movements of participants under real conditions with the sun in their field of vision. In that study, the authors observed that participants were disturbed by glare and that they tended to avoid the brightest area, particularly during visually demanding tasks. Although relevant for visually uncomfortable conditions, these results do not address the effect of lighting on saliency in comfortable conditions, which thus remains to be explored. Furthermore, while lighting is a key aspect of architectural composition for emphasizing certain spatial features, none of the identified studies, with the exception of Sarey Khanie et al. (2017), have examined how light and daylight may affect our visual attention in indoor scenes.

Saliency prediction models have been applied in a number of studies related to the built environment. These include the utilization of bottom-up models to quantify the visual impact of photovoltaics on facades (Xu and Wittkopf 2014), to assess the visual impact of landscapes (Dupont et al., 2016), and to test the impact of biophilic design on people's emotion (Genetics of Design, 2020a; 2020b). In a recent study, Xu et al. compared four models in the context of wayfinding (Xu et al., 2020). The authors concluded on the benefits of using saliency models in design, with the most advanced models (DL) being the best performing. These examples highlight both the growing interest and the large range of possible applications of saliency models for architectural design. However, the limited number of studies and models tested so far also show that the implementation and applicability of saliency prediction in architecture remains largely unexplored. In addition, to the authors' knowledge, visual attention and saliency prediction have not been systematically investigated in scenes depicting interior spaces without people or animals, which are particularly relevant for delineating and understanding the influence of the characteristics of indoor environments on human experience.

1.4. Objectives and hypotheses

Despite its significant presence in computer science research, saliency prediction modelling has not yet truly penetrated the field of architectural design. At the same time, architectural scenes have not been used to develop saliency models. These scenes present particularities compared to natural images: they are typically devoid of people or other objects of strong salience and focus on man-made forms seen under specific lighting conditions. Taking these observations as a starting point, the objectives of this paper are to evaluate the performance of existing saliency prediction models for omnidirectional architectural daylit scene and to determine the effect of daylight on human head movements, used as a proxy for visual attention To better guide our analyses, we formulated three hypotheses.

H1. Traditional models are as reliable as pre-trained DL models when applied to our datasets

By leveraging an end-to-end training with deep convolutional neural networks (CNNs), DL models are known to have overcome inherent limitations of traditional models (Borji, 2018). Nevertheless, a general concern of DL models is their transferability when they are applied to other types of image stimuli than their training dataset. The predictive power of models pre-trained on natural images can thus be limited for rendered architectural scenes that remain primarily based on spatial cues and do not involve very informative objects nor people. Given the focus of traditional models on low-level features (such as direction and contrast), we hypothesized that traditional and pre-trained DL models might be similarly reliable when applied to architectural scenes.

H2. Saliency prediction models are more accurate for overcast sky conditions than for clear sky conditions

Daylight conditions are influenced by the position of the sun and by the weather, and in turn influence the visual content of a given space, most notably through contrasts and the interplay of light and shadow. The same space exposed to different daylight conditions is thus a different visual stimulus. As such, a model is likely to predict different regions of saliency for changing sky conditions. Considering how much the interplay of light and shadow may provide misleading visual cues for saliency prediction, we hypothesized that diffuse conditions might increase the model's accuracy.

H3. There is a correlation between head movements for the same space under different sky conditions (i.e., undergoing daylight dynamics)

Our last hypothesis detaches from the performance of the saliency prediction models to focus on human head movements towards the same space across different daylight conditions, with varying presence of direct sunlight. Following observations on the rather uninformative nature of light [43], we hypothesized that head movements would be similar under these different conditions.

2. Datasets

Our study is based on three datasets of head-tracking logs that were collected based on 360° scenes shown to human participants in an IVE between 2016 and 2018: two datasets, which will be labelled according to their first author's last name i.e. *Chamilothori* and *Rockcastle*, are derived from research on daylight perception in architectural spaces: *Rockcastle* from (Rockcastle et al., 2017), *Chamilothori* from (Chamilothori et al., 2022a; Moscoso et al., 2021). The third dataset, named *Sitzmann* from (Sitzmann et al., 2018), consists of a subset of interior scenes from a publicly available dataset from work on saliency modelling. The three datasets are used to examine saliency prediction in grayscale indoor scenes with no objects (*Rockcastle*, with examples shown in Fig. 4(a)), in colored indoor scenes with relatively few objects (*Chamilothori*, cf. Fig. 5(a)), and in colored indoor scenes with multiple objects and higher levels of detail and texture, but hardly any person present (*Sitzmann*, cf. Fig. 6(a)). Table 1 further summarizes the instruments and experimental protocols used in each study. More details on these datasets are summarized in the Supplementary Material, section 1.

The original objective of the studies behind the *Rockcastle* and *Chamilothori* datasets was actually not saliency prediction but the appraisal of perceptual responses to variation of daylight and spaces. As such, the viewing duration of each scene lasted a few minutes and was made of an initial silent exploration period, the duration of which depended on the participant's readiness in Rockcastle et al. (2017) while it was restricted to 30 s in Chamilothori et al. (2022a). This exploration period was followed by a verbally administered questionnaire, answered while the participant remained immersed in the scene. In the present study, we will thus only examine the first 30 s of viewing for both datasets to remain within the initial exploration period. We should note that as the same questions were asked for each scene (in randomized order), some influence on the free-viewing of the scenes cannot be excluded. For the *Chamilothori* dataset, we also note that only two types of rooms were modeled, with variations in window size, shading geometry patterns, and sky type: despite these variations, there was therefore a certain level of redundancy between the scenes.

Table 1
Description of the 360° datasets used in this study.

	Rockcastle	Chamilothori	Sitzmann
Virtual reality (VR) headset	Oculus Rift CV1 (75 Hz)	Oculus Rift CV1 (75 Hz)	Oculus Rift DK2 (75 Hz)
Software	Oculus and Unity	Oculus and Unity	Oculus and Unity
FOV	100 × 110°	100 × 110°	95 × 106°
Headset resolution	1080 × 1200 px/eye	1080 × 1200 px/eye	960 × 1080 px/eye
Head-tracking (sampling rate)	90 Hz (estimated) Log retrieved every 11ms.	90 Hz (estimated) Log retrieved every 11ms.	120 Hz.
Eye-tracking (sampling rate)	N/A	N/A	PupilLabs stereoscopic installed in HMD (120 Hz)
Type of scenes	Rendering from 3D models	Rendering from 3D models	Rendering from 3D models
HMD projection Resolution of original image	Cubemap (1200 × 1200 px/cubeface)	Equirectangular (4320 × 2160 px)	Equirectangular (8192 × 4096 px)
Rendering engine or camera	Radiance	Radiance	Unknown (most likely to be various engines)
Tone-mapping	Ward 97	Reinhard 02	(unknown)
Scene starting point	Horizon line of the most contrasted spot in the image	Towards the window	4 options per scene (90° changes)
Redundancy (between-subject factor; each participant exposed to the same condition of a factor)	Space: No (8 types) Sky: Yes (2 types) (randomization)	Sky: Yes (3 types) Space: Yes (2 types) Window size: Yes (3 types) Façade geometry: No (6 types) (randomization)	No
Duration of scene shown	>30 s.*	>30 s.*	30 s.
Number of scenes	16	96	14 interior scenes (of 22). Subset of interior scenes with hardly any people on them
Number of participants per image (average)	12	10-47 (it differed across spaces because of the experimental protocol)	44 (inferred)
Participant position	Standing	Seated	Seated

* The actual viewing time was longer. For saliency purposes, we only analyze the head tracking logs of the first 30 s of exposure to the scene.

3. Method

These three datasets, comprising a total of 126 rendered interior scenes associated to headtracking logs, were used to evaluate the prediction accuracy of existing saliency models. Overall, thirteen saliency prediction models as well as three maps based on mathematical signals were applied on these scenes, while the corresponding head tracking logs were used to generate ground truth maps. Both saliency models and head tracking logs were adapted to equirectangular representations, with x representing the yaw and y the pitch. For each scene, we compared the prediction and ground truth maps by using image-based correlation metrics. Fig. 1 summarizes this approach. Details about the selected saliency models, the determination procedure for the ground truth, and the metrics used for the comparisons are further detailed in the following sections.

3.1. Saliency prediction models

In this study, we examined different types of models, including traditional and DL models. The selection procedure started with known models that had already been validated, and then depended on the public availability of the models. This led to a number of widely used traditional models and publicly available DL models. Some models were readily available for 360° scenes and some were only available for 2D scenes. The models tested are summarized in Table 2 and each of them is further detailed in the Supplementary Material, section 2.1. Note that the DL models were all pretrained on other datasets (i.e., we did not train any of them with our own data).

In addition to these models, we generated three maps based on mathematical signals that we used to compare the prediction accuracy of the models output against artificial and random signals. These signals comprised one Laplacian function, which was based on parameters identified by Sitzmann et al. (2018) as the most adequate to describe the equator bias ($\mu = -1.30°$, $\beta = 18.58°$), and two noise signals. These signals were adapted to fit the dimensions of the ground truth maps and are shown in Fig. 2.

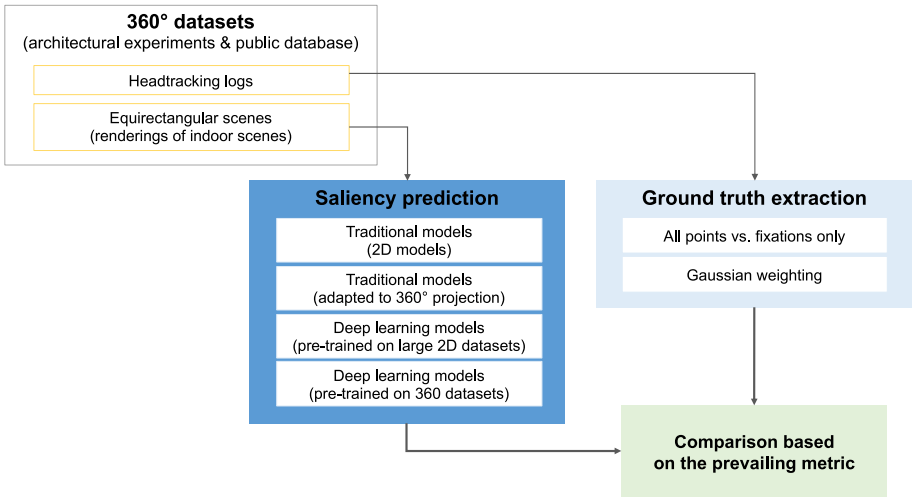

Fig. 1. Summary of the workflow in the present study: we computed the saliency maps from the equirectangular scenes which we compared to the ground truth maps that were extracted from the head tracking logs.

Table 2
Saliency prediction models and baseline images tested in this study.

Model type	Projection type	Name	Reference
Traditional	**Perspective (2D)** (Directly applied)	(1) IKN	(Itti et al., 1998)
		(2) GBVS	(Harel et al., 2007)
		(3) MSSS	(Achanta & Süsstrunk, 2010)
		(4) BMS	(Zhang & Sclaroff, 2013)
	Omnidirectional (Adapted to 360° projection)	(5) BMSeq[a]	(Lebreton & Alexander, 2018)
		(6) BMS360[b]	
		(7) BMS360eq[a, b]	
		(8) GBVS360eq[a, b]	(Lebreton & Alexander, 2018)
Deep learning (DL)	**Perspective (2D)** (Directly applied) Pretrained on 2D datasets	(9) Sal-CEDN	(Kroner et al., 2020)
		(10) UNISAL	(Droste et al., 2020)
		(11) DeepGaze II	(Kümmerer et al., 2016)
	Omnidirectional (Pre-trained/adapted to 360° projection)	(12) SalNet360	(Monroy et al., 2018)
	Omnidirectional (Developed for/pre-trained on 360° stimuli)	(13) SaltiNet	(Assens Reina et al., 2017)
Baseline images (not models)	**NA**	(14) Laplacian	
	NA	(15) Noise-1	
	NA	(16) Noise-2	

[a] We used "eq" to indicate to the inclusion of an equatorial prior in the algorithm. Equatorial prior correction involves the identification of a "line of horizon" of the scene (that can be below/above 0°).

[b] The original models (GBVS (Harel et al., 2007) and BMS (Zhang & Sclaroff, 2013)) were adapted to a 360° projection by (Lebreton & Alexander, 2018).

3.2. Generation of ground truth maps

Ground-truth saliency maps are constructed from the aggregated viewing logs of multiple observers without accounting for temporal information. We reviewed ground truth mapping methods for studies conducted with an HMD (see the Supplementary Material, section 2.2) and decided to only rely on the Gaussian weighting method ($\sigma = 11.7°$) used in Sitzmann et al. (2018). We also corrected for the distortion from the projection by rescaling logs on the poles to avoid distortions, based on the method described in Upenik and Ebrahimi (2017).

3.3. Evaluating saliency prediction

Metrics used to estimate saliency prediction can be classified as either based on location (discrete fixation) or distribution (Bylinskii et al., 2016). The second type was found more suitable in the case of head-tracking logs, where fixation regions may be less precise than with eye-tracking, while still relevant for informing design decisions as the aim is merely to determine which broad regions attract human attention.

For our own study, we used the Pearson correlation coefficient (CC) (linear correlation coefficient), a statistical measure of the strength of association between two variables, following previous work by Sitzmann et al. (2018) and Gutiérrez et al. (2018). CC calculates the cross correlation between the predicted saliency and the ground truth maps after normalizing the maps (*pixel value - mean value of the pixels of the map/SD*) and ranges from −1 (perfectly inversely correlated) to 1 (perfectly correlated). High absolute values indicate that both maps have similar values at the same locations.

As stated previously, our natural head behavior leads to a strong equatorial bias, which can be reinforced by the weight that the HMD adds to the head. Both ground-truth and saliency output maps are likely to show their upper and lower image regions as less visually attractive, which implies "easy-to-correlate" regions. To address this limitation, we decided to apply the CC on both the full image as well as on the central third of the equatorial image region (Fig. 3 right), which represents an angle of 30° above and 30° below the equator. The equatorial region of the image includes nearly all the non-black region of the scene generated through the aforementioned Laplacian fit (Fig. 3, left).

3.4. Statistical analysis

We reported the results by saliency model and/or type of saliency models for each dataset. The main results are reported as average CC which is commonly used in saliency prediction literature (Borji, 2018). In earlier stages, we had considered using the Similarity metric and the Kullback-Leibler divergence but we excluded following initial tests (more details in Supplementary Material, section 2.3). We used statistical tests to assess the difference in average CC obtained across models and group of models (e.g., traditional models vs. DL models). As such, CC is the dependent variable and the models or group of models are the independent variables. We relied on the Shapiro-Wilk test to assess the

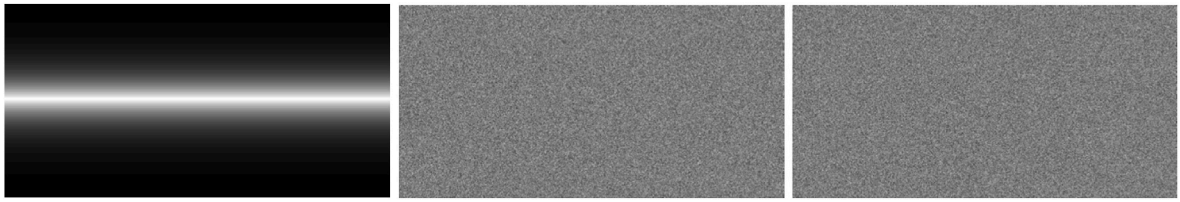

Fig. 2. Laplacian function, noise signal (two types) applied on a 2:1 image format.

Fig. 3. Regions of interest (full image and equatorial region) on which we also apply the CC metric.

normality of the data. For non-parametric data, we used the Mann-Whitney test (comparison involving two groups), the Kruskal Wallis one-way analysis of variance (comparison involving more than two groups), and the r effect size for the Wilcoxon two-sample rank-sum test. For the comparisons between groups, we used a significance level of $\alpha = 0.05$. We relied on Ferguson's thresholds for interpretation of the strength of association (for both CC and r), where 0.2 is considered as a small association, 0.5 a moderate association and 0.8 a strong association (Ferguson, 2009).

4. Results

4.1. Qualitative observations on ground truth and saliency prediction maps

The **ground truth map** (based on the participant's head movements for the chosen extraction model) and the series of **saliency prediction maps** (based on the image stimuli for the different saliency models selected for this study) were produced for a set of 7 scenes (4 different spaces with sky variations), taken from the three datasets. Examples are provided in Figs. 4, Figs. 5 and 6 for selected scenes in the *Rockcastle*, *Chamilothori*, and *Sitzmann* datasets, respectively, and in the Supplementary Material, section 3.

Overall, we note that both the ground truth and the saliency maps feature blurry bright regions on a black background. These so-called regions of interest are often difficult to identify as they spread over different spatial elements. This can be partially explained by the fact that we examine head and not eye movements (hence wider Gaussians). We also usually observe a strong equatorial bias (horizontal central directionality) in the regions of interest, which is a result of human physiognomy and is implemented in both the ground truth map extraction and most of the saliency models. Regions of interest generally tend to include the scenes' brightest areas (e.g., windows or shading), but sometimes include unexpected regions (e.g., focused on bare walls).

Looking more specifically at the **saliency prediction** models, we first note that the earlier models (IKN, MSSS, GBVS and BMS) are generally driven by the brighter and high-contrast regions. BMS and GBVS embed the expected center bias, which can help the prediction but will not be as powerful as an equatorial bias for our data, while MSSS is the only model displaying sharp shapes from the original picture (instead of blurry regions). For all these models, the shades and shading patterns on the floor in the *Chamilothori* dataset are generally predicted as salient regions.

The adaptation of the 2D models GBVS and BMS to the 360° environments has led to notable changes compared to their original maps. By examining the output of GBVS360eq and BMS360eq, we see that the adaptations of these models are generally less pixelated (i.e., have an increased resolution), more contrasted (i.e., have fewer gray regions) and include an equatorial bias embedded into the maps. The scripts available for the adaptation of BMS allowed us to separate the impact of the 360° reprojection and of the added equator prior. BMS360 is closest to the original image but with continuity added between the right and left ends of the image. BMSeq also has this feature but the predictions generally appear slightly more compressed around the equator. Finally, the variations of BMS can sometimes also bring artefacts (cf. Fig. 6(h), (i) and (j) for the *Sitzmann* scene where the white square in the center left of the scenes was not present in the original BMS output Fig. 6(f)).

The DL models developed for the 2D stimuli clearly show a central bias with a reduced region of interest on the edges (including the left and right borders), yet the salient regions (white) remain thin around the equator. These prediction maps overall appear more contrasted (less gray) than the maps from the traditional models. UNISAL leads to a cloudier and blurrier output, followed by Sal-CEDN and DeepGaze II. DeepGaze II can sometimes appear sensitive to objects (e.g., furniture elements are prominent in Fig. 6).

Finally, DL models developed for the 360° stimuli surprisingly display less equatorial bias than many other models. SalNet360 is based on a cubemap extraction and we can see the shape of the cubemap parts (i.e., squares) on all predictions (these are very visible on Fig. 4(n) for instance). In addition, the identified salient regions in each square do not appear to be continuous with those in the adjacent square. SaltiNet was developed directly on 360° images and shows a good continuity of the salient regions. We can identify the location of windows and notable objects on these otherwise very blurred outputs, which spread over two thirds of the height of the image (centered on the equator).

4.2. Overview of prediction accuracy per model

We computed the average CC obtained for each dataset and model. The results are detailed in Table 3 for both the full image and the equatorial image regions. This table is supplemented by Figs. 7 and 8 that represent boxplots of the CC for each model and dataset for the full and equatorial image regions, respectively. Given the difference in testing conditions and sample size for each dataset, we analyze the results for each dataset separately.

The models show a wide spread of average CC, ranging from 0.20 to 0.78 for the full images, and of 0.12–0.70 for the equatorial image regions. Only a few models reached an average CC of 0.5 (moderate correlation) for the centered image region for the *Chamilothori* and *Sitzmann* datasets. None of the models reached 0.8 (high correlation), not even for the full image. The model that performs the best (i.e., that reached the highest CC across the datasets and image region) is BMS360eq, a model originally based on Gestalt principles and which considers colors, is adapted to 360° stimuli and includes an equatorial prior. The lack of color in the *Rockcastle* scenes can explain the poor performance of the model on this data set. BMS360eq is followed by GBVS360eq and BMSeq. In other words, traditional models encapsulating spatial visual cues (such as BMS and GBVS) and further adapted to 360° stimuli are

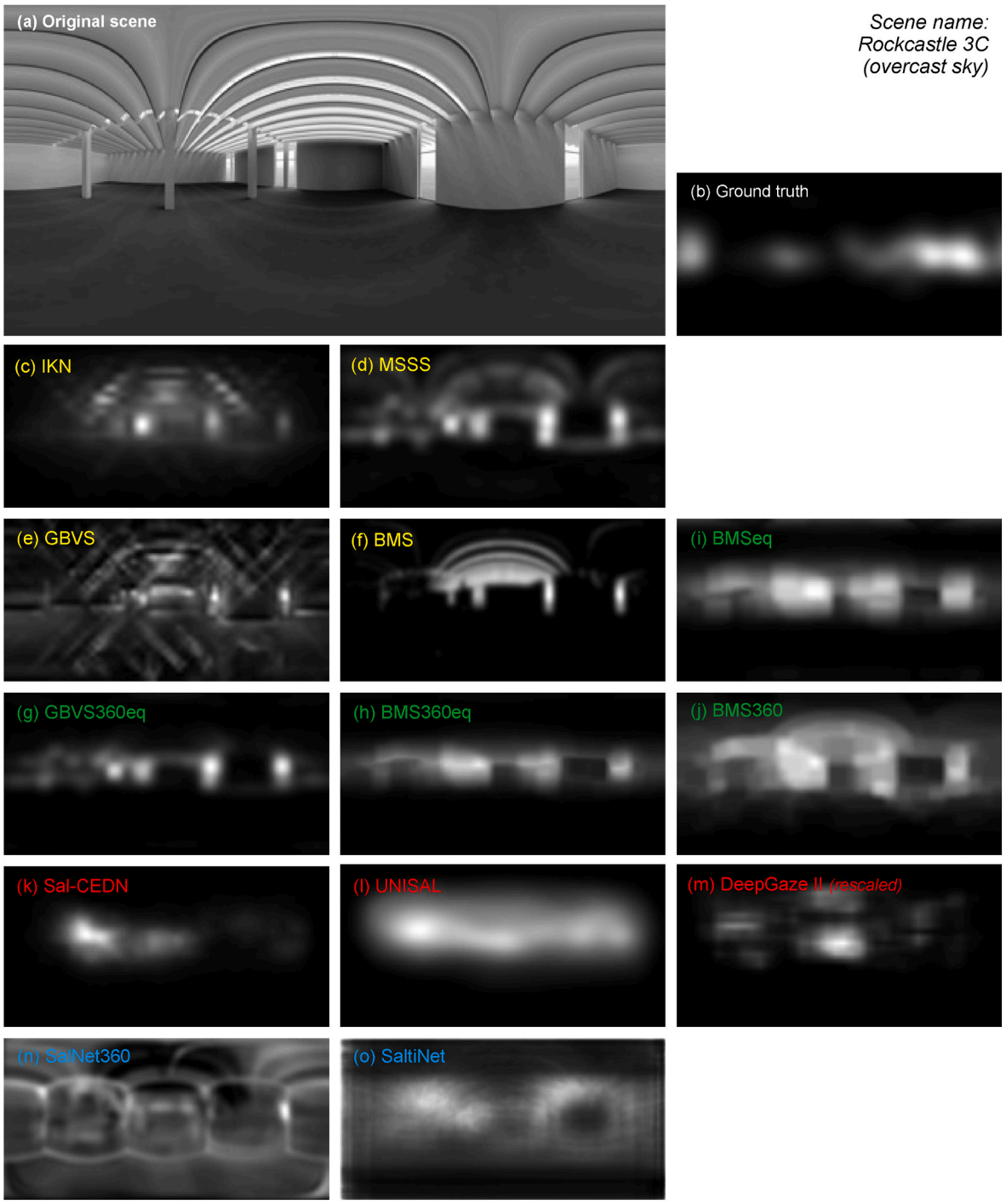

Fig. 4. Original stimuli, the ground truth map and the output of the saliency prediction maps for a scene from the Rockcastle dataset. Saliency prediction model outputs are indicated with colors corresponding to the type of model and projection: traditional 2D models (directly applied) (**yellow**), traditional models adapted to omnidirectional projection (**green**); deep-learning 2D models (directly applied) (**red**) and deep-learning omnidirectional models (**blue**). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

able to obtain the highest CC with the ground truth. The models further show a better performance on the *Chamilothori* and *Sitzmann* datasets, which can be explained by the colors, textures and objects present in the scenes (which are taken into account in most of the saliency models) and, possibly, by the more rigorous free visualization period (thus closer to the procedure used for the creation of the models) than the one behind the *Rockcastle* dataset.

Overall, the saliency models performed better when applied to the full image, which is expected considering the matching of the darker upper/lower image regions between the ground truth and the predictions for many models that incorporate the equatorial bias. Similarly, yet quite surprisingly, we found that the Laplacian signal is able to surpass most models and appears as one of the best performing models when considering the full image regions. This higher correlation is reduced as soon as we only consider the central part of the image. For this reason, and to avoid inflating the conclusions of this analysis, the output of the equatorial image region is the primary one being discussed in the next sections when testing the hypotheses related to the saliency

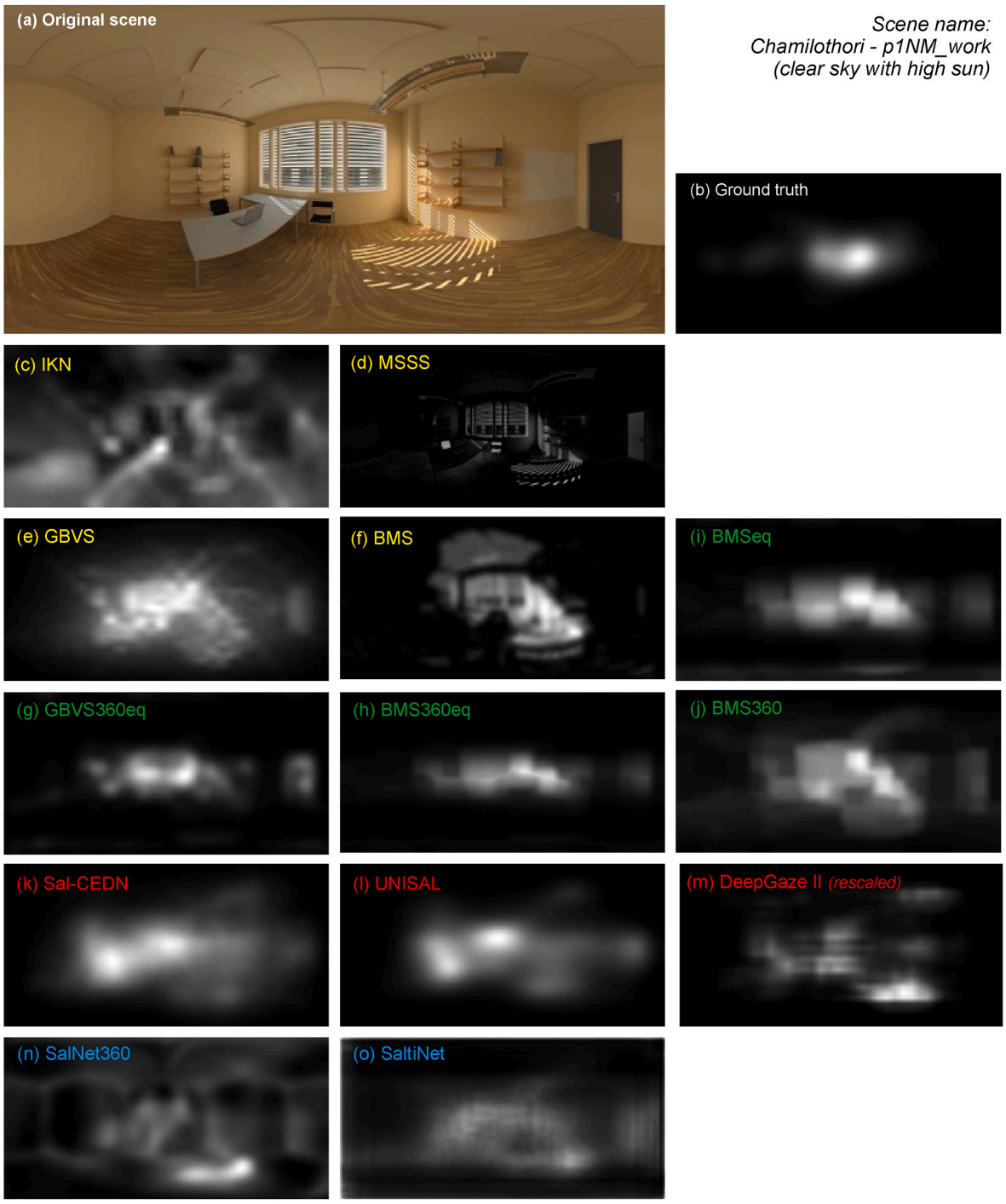

Fig. 5. Original stimuli, the ground truth map and the output of the saliency prediction maps for a scene from the Chamilothori dataset. Saliency prediction model outputs are indicated with colors corresponding to the type of model and projection: traditional 2D models (directly applied) (**yellow**), traditional models adapted to omnidirectional projection (**green**); deep-learning 2D models (directly applied) (**red**) and deep-learning omnidirectional models (**blue**). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

models' predictive performance (Sections 5.3.1 (H1) and 5.3.2 (H2)). That said, both the full image and the equatorial image region were systematically analyzed for each hypothesis to ensure that no unexpected findings would emerge from their comparison. Numerical results are reported for both the full images and the equatorial region images.

4.3. Hypothesis testing

H1. Traditional models are as reliable as pre-trained DL models when applied to our datasets

For this hypothesis, we group the saliency models in two groups: traditional models and DL models (pre-trained on natural image datasets). Considering the difference in sample size and stimuli type, we analyzed each dataset independently. A Shapiro-Wilk test revealed the non-normality of this data, leading us to apply a non-parametric Mann-Whitney test. The difference between the two groups of models was statistically and practically significant with a small effect size only for the *Chamilothori* dataset (see Fig. 9 and Table 4). For the *Sitzmann*

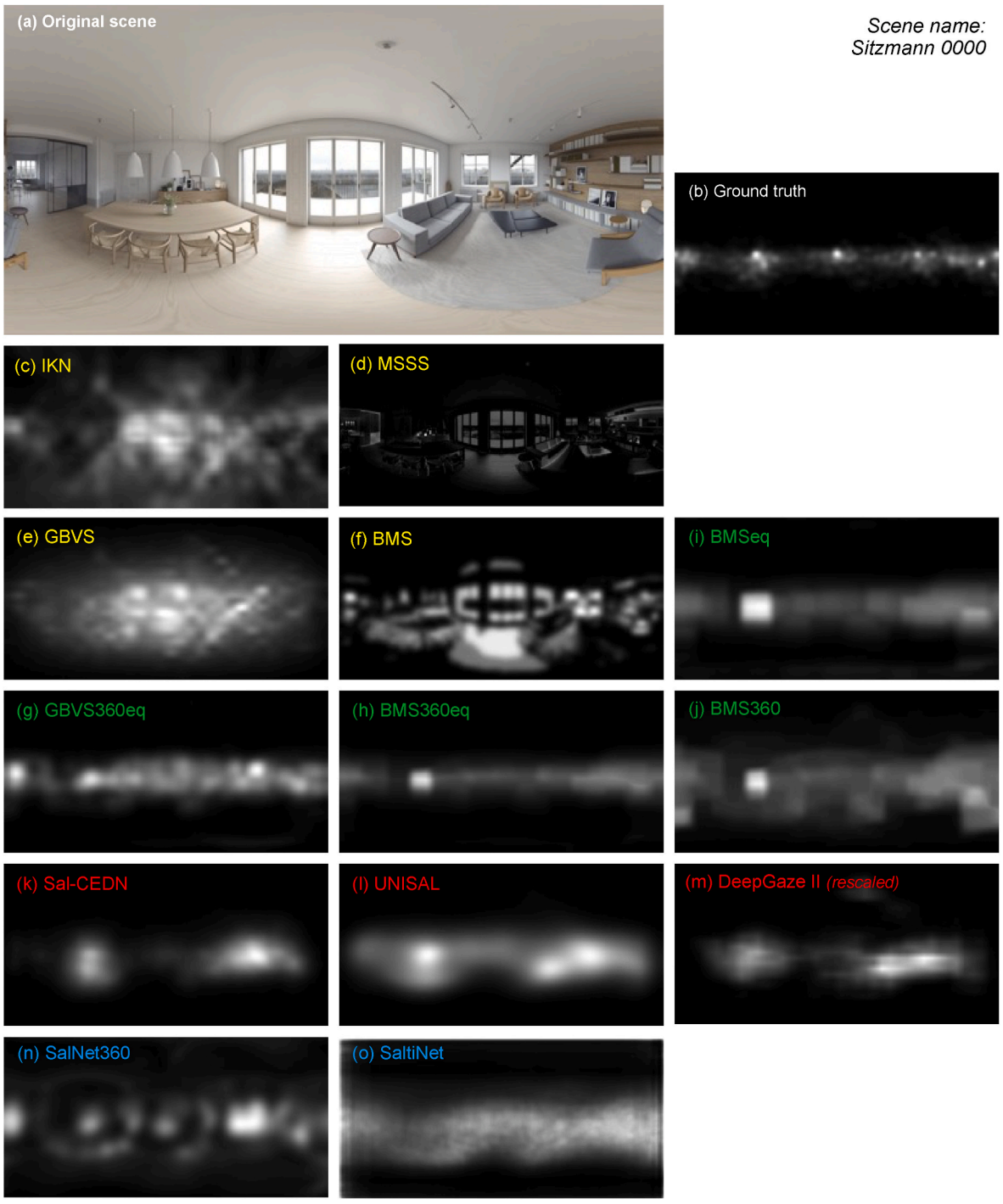

Fig. 6. Original stimuli, the ground truth map and the output of the saliency prediction maps for a scene from the Sitzmann dataset. Saliency prediction model outputs are indicated with colors corresponding to the type of model and projection: traditional 2D models (directly applied) (**yellow**), traditional models adapted to omnidirectional projection (**green**); deep-learning 2D models (directly applied) (**red**) and deep-learning omnidirectional models (**blue**). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

dataset and for the equatorial image regions, results show a significant difference with a small effect size. Interestingly, this difference in prediction performance did not follow the same direction for the two datasets: the *Chamilothori* dataset showed a higher prediction accuracy for traditional models while the *Sitzmann* dataset showed a higher prediction accuracy for DL based models. Following the qualitative observations, spatial objects such as furniture might have been identified as most salient in DL models. Yet, the dominance of the windows and shading patterns shows that a bottom-up traditional model will be more reliable for the *Chamilothori* dataset. DL models highly depend on the type of images they were trained, which can explain their poorer performance on the set of images tested in the present study.

H2. Saliency prediction models are more accurate for overcast conditions than for clear sky types

This hypothesis was tested on the equatorial image region only, using the *Chamilothori* dataset (n = 32 scenes) and the CC between ground-truth and saliency maps for each model, for every sky type (n = 3, see

Table 3
Average correlation coefficients (CC) (absolute values) for each model and dataset for the full images and the equatorial region of the images. Moderate correlations (>0.5) are highlighted for legibility.

Model type	Projection type	Model acronym	*Full images*			*Equatorial region of the images*		
			Rockcastle n = 16	*Chamilothori* n = 96	*Sitzmann* n = 14	*Rockcastle* n = 16	*Chamilothori* n = 96	*Sitzmann* n = 14
Traditional	2D	IKN	0.25	0.18	0.29	0.19	0.25	0.30
		GBVS	0.40	0.65	0.47	0.24	0.64	0.33
		MSSS	0.25	0.30	0.20	0.12	0.29	0.16
		BMS	0.24	0.62	0.26	0.17	0.64	0.21
	360°	BMSeq	0.50	**0.74**	0.61	0.24	0.69	0.48
		BMS360	0.45	0.69	0.57	0.18	0.66	0.46
		BMS360eq	0.53	**0.78**	0.66	0.26	**0.70**	0.53
		GBVS360eq	0.56	**0.75**	0.55	0.24	**0.70**	0.35
Deep learning (DL)	2D	Sal-CEDN	0.44	0.56	0.59	0.28	0.44	0.46
		UNISAL	0.47	0.58	0.62	0.24	0.49	0.50
		DeepGaze II	0.37	0.33	0.51	0.28	0.25	0.46
	360°	SalNet360	0.28	0.37	0.51	0.14	0.40	0.53
		SaltiNet	0.49	0.45	0.49	0.18	0.41	0.39
Baseline images	NA	Laplacian	**0.70**	0.61	0.62	0.45	0.38	0.45
		Noise-1	0.00	0.00	0.00	0.00	0.00	0.00
		Noise-2	0.00	0.00	0.00	0.01	0.00	0.00

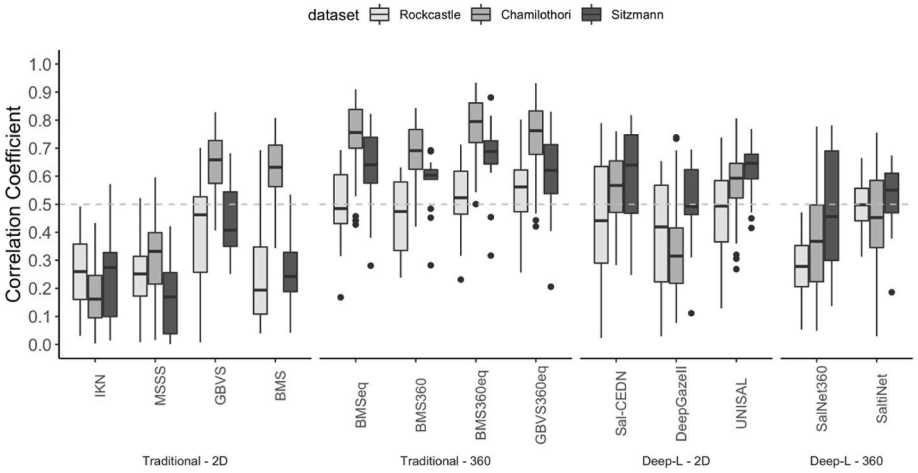

Fig. 7. Boxplot of correlation coefficients (CC) of each model and dataset considering the **full** image.

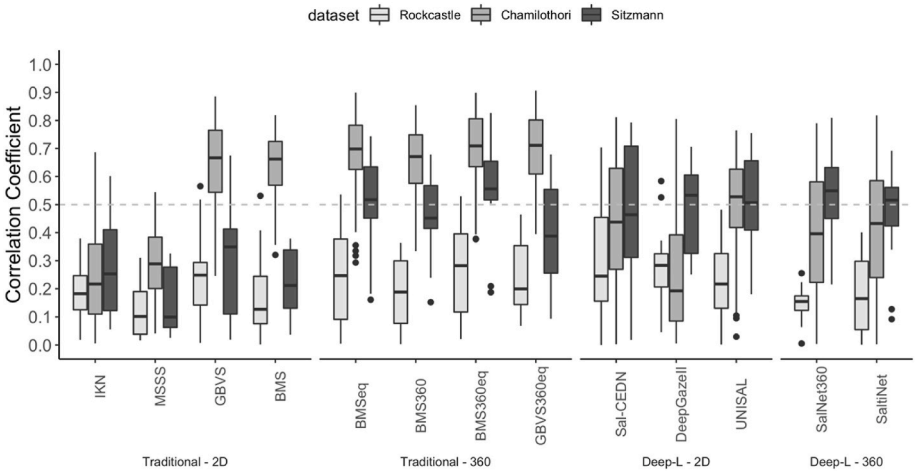

Fig. 8. Boxplot of correlation coefficients (CC) of each model and dataset considering the **equatorial image region**.

Fig. 10). Although the *Rockcastle* dataset also included variations in sky type, the sample size was too limited (n = 8) to be meaningful and was not included in the analysis.

Considering the non-normality of the data and that sky type was a between-subjects factor in the original experiment, the effect of sky type on prediction accuracy (CC) for each model was examined with

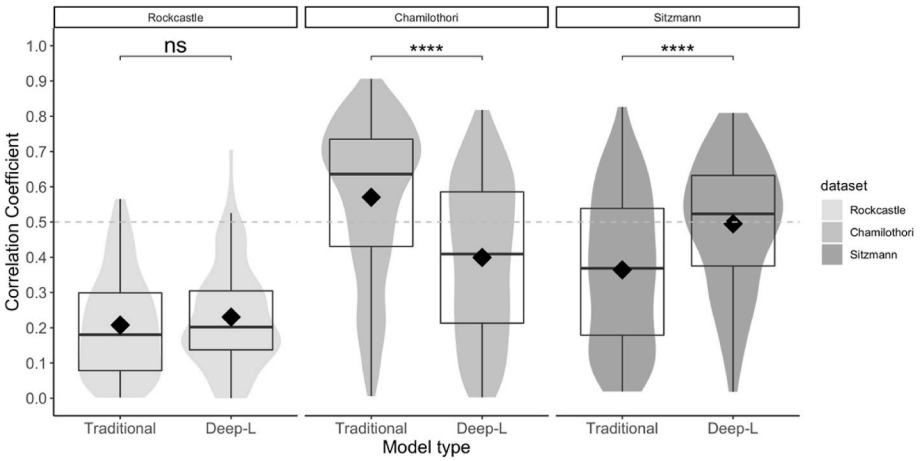

Fig. 9. Boxplot of correlation coefficients between the ground truth and the saliency models grouped by type of the models and datasets considering the **equatorial image region**.

Table 4
Pairwise comparison of correlation coefficients between the ground truth and the saliency models for the traditional vs. deep learning models for the **full** and **equatorial image region**s.

		Traditional models				Deep learning models				Comparison		
		N	**M**	**Mdn**	**SD**	***N***	**M**	**Mdn**	**SD**	**ΔM**	**p-value**	**r (effect size)**
Full images	Rock.	*128*	0.40	0.43	0.19	*80*	0.41	0.44	0.19	−0.01	0.64	0.03 (negl.)
	Cham.	*768*	0.59	0.66	0.24	*480*	0.46	0.48	0.18	0.13	<0.0001	**0.33 (small)**
	Sitz.	*112*	0.45	0.47	0.23	*70*	0.54	0.57	0.17	−0.10	<0.01	0.19 (negl.)
Equat. Region	Rock.	*128*	0.21	0.18	0.14	*80*	0.23	0.20	0.15	−0.02	0.31	0.07 (negl.)
	Cham.	*768*	0.57	0.64	0.22	*480*	0.40	0.41	0.22	0.17	<0.0001	**0.37 (small)**
	Sitz.	*112*	0.36	0.37	0.21	*70*	0.50	0.52	0.19	−0.13	<0.0001	**0.30 (small)**

N: sample size (number of images), M: mean, Mdn: median, SD: standard deviation, ΔM: difference in mean, p-value: statistical significance ($p < 0.001$ highly significant; $p < 0.01$ significant; $p < 0.05$ less significant; ns: not significant), r: effect size.

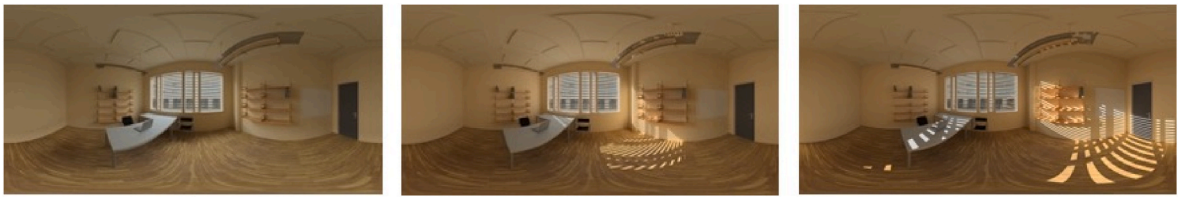

Fig. 10. Rendering of a space from the Chamilothori dataset lit by three sky conditions: overcast (left), clear sky with high sun position (center), and clear sky with low sun position (right).

Kruskall-Wallis tests. Results show that sky condition influenced prediction performance for the traditional models MSSS, BMS, for all the traditional models adapted to 360° projection (which performed best in the *Chamilothori* dataset), and for the DL models UNISAL and SaltiNet (Table 5). Pairwise comparisons between the three sky types per model were conducted with Mann-Whitney U-tests (Fig. 11 and Table 6). We observe a tendency for models to perform better (higher CC) in scenes with an *overcast sky* or a *clear sky with a high sun position* compared to *clear sky with a low sun position*. This difference in prediction performance is statistically significant ($p < 0.001$) or most traditional models, including their adaptation to 360° environments (for GBVS and BMS). For DL models developed for 2D images, the best prediction was found for scenes with *clear sky with high sun angle* for UNISAL (compared to both other conditions) and Sal-CEDN (compared to overcast sky). For DL models developed for 360° images, a small but significant increase in prediction accuracy was found for *overcast sky* compared to *clear sky with low sun angle* for SaltiNet.

Table 5
Results of the Kruskal Wallis test on the effect of sky type on prediction accuracy per saliency model.

Model	Model Type	chi-squared (χ^2)	p-value
IKN	Traditional – 2D	1.922	0.38
MSSS	Traditional – 2D	**27.146**	**< 0.0001**
BMS	Traditional – 2D	**24.739**	**< 0.0001**
GBVS	Traditional – 2D	4.0381	0.13
BMS360	Traditional – 360	**21.192**	**< 0.0001**
BMSeq	Traditional – 360	**32.637**	**< 0.0001**
BMS360eq	Traditional – 360	**34.887**	**< 0.0001**
GBVS360eq	Traditional – 360	**18.108**	**< 0.001**
DeepGazeII	Deep-L-2D	0.69354	0.71
Sal-CEDN	Deep-L-2D	4.0276	0.13
UNISAL	Deep-L-2D	**7.4841**	**< 0.05**
SalNet360	Deep-L-360	1.566	0.46
SaltiNet	Deep-L-360	**6.6679**	**< 0.05**

To conclude, most models performed better in scenes with an *overcast sky* or a *clear sky with high sun position* than with a *clear sky with low sun position,* which depict large shadows and high contrasts that seem to be falsely interpreted by models. These findings partially confirm our

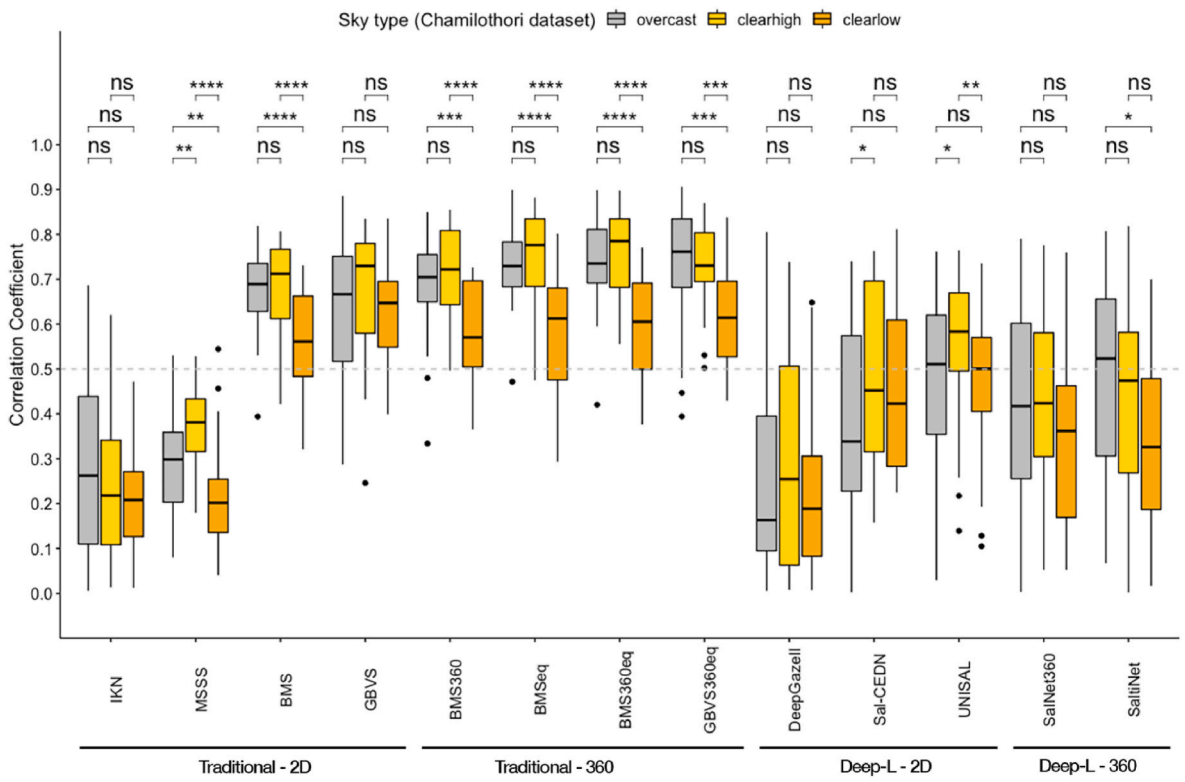

Fig. 11. Boxplot of correlation coefficients between the ground truth and the saliency models for the Chamilothori dataset (n = 32 unique spaces) grouped by type of sky: overcast sky ("overcast"), clear sky with high sun position ('clearhigh') and clear sky with low sun position ("clearlow"). Considering **equatorial image region**. The statistical significance levels shown on the plot are p-values resulting from the pairwise comparison between type of sky, where: **** ($p < 0.0001$), *** ($p < 0.001$) are considered highly significant; ** ($p < 0.01$) significant; * ($p < 0.05$) less significant and "ns" not significant.

hypothesis, as models performed better in scenes with overcast sky conditions compared to clear sky with low sun position (as hypothesized), but compared to scenes with a clear sky and high sun position.

H3. There is a correlation between head movements for the same space under different sky (i.e., daylight) conditions

To test this hypothesis, correlation coefficients (CC) were calculated between ground truth maps across different sky types for each unique space: in this case, a high CC will suggest a strong similarity in head movements between different sky types. The *Rockcastle* dataset included two sky types leading to a comparison between two groups, while the *Chamilothori* dataset included three sky types leading to three comparisons between two groups. The results (here shown for both the full and equatorial image regions) are presented in Table 7 (*Rockcastle* dataset) and Table 8 (*Chamilothori* dataset).

As shown in Table 7, the mean CC was 0.67 with a standard deviation of 0.12 and a minimum correlation of 0.55. However, when looking at the equatorial image regions, the CC decreases substantially, which suggests that the high CC was mainly due to the less salient upper and lower image regions (a typical consequence of the equator bias) which matched across sky types (see example in Fig. 12). The small CC obtained for the equatorial image regions leads us to conclude on rather different head moving patterns across sky types for this dataset.

Examining Table 8, we can see that the mean CC was superior or equal to 0.90 across the three comparisons for the full images and superior or equal to 0.87 for the equatorial image regions. The first quartile remained nearly equally high ($CC \geq 0.88$), suggesting that despite changing light conditions, the façade remained the area of focus. Even looking at equatorial image regions, the correlation coefficient between ground truth maps across different sky types for one unique space remained strong (the first quartile was of $CC \geq 0.84$). These high correlations (see example in Fig. 13) may be explained by the scene variations within the experimental procedure, where the external façade was intended to be dominant: the shadows in the space (resulting from the façade variation) may thus not have altered the head direction as much as expected. To conclude, the strong correlations do suggest highly similar head movements across sky types for the same space in the *Chamilothori* dataset, which would confirm our hypothesis for this dataset.

5. Discussion

5.1. Methodological assumptions

This evaluation required us to take decisions on the pre-processing of the raw data and on the choice of evaluation metrics. Fully immersive 360° environments require scenes to be adapted and projected within the headset. We constructed out ground truth data based on cartesian coordinates (x, y, z), that we transformed to equirectangular coordinates (yaw, pitch), which leads to unavoidable deformations on the poles (the "nadir" of the scene is a point in the 360° environments and is a line in the corresponding equirectangular images). We addressed this deformation by rescaling the logs on the poles to avoid distortions based on the method described in Upenik and Ebrahimi (2017). Additional deformations related to the projection method can also be introduced in the models. We notably observed this inconsistency in SalNet360 (Monroy et al., 2018), where the extraction of data from cubemap projections led to the "stitching" of the cubes together, resulting in observable discontinuities of the salient regions at the edges.

Head movement was used instead of eye movement to determine visual attention, as it was the only view behavior data available. Although the literature has shown good correlations between head and eye movements (Sitzmann et al., 2018), testing further the similarity of the two signals would be beneficial. The obtained ground truth maps,

Table 6
Post-hoc pairwise comparisons with Mann-Whitney U-tests to examine the difference in prediction accuracy between pairs of sky types (overvast, 'clearhigh', i.e., clear sky with high sun angle, and 'clearlow', i.e., clear sky with low sun angle). In this table, we only report statistically significant results following the main analyses in Table 5.

Model	Model Type	group 1	group 2	p-value	Effect size (r)
MSSS	Traditional – 2D	overcast	clearhigh	<0.01	−0.37 (small)
	Traditional – 2D	overcast	clearlow	<0.01	0.36 (small)
	Traditional – 2D	clearhigh	clearlow	<0.0001	0.61 (moderate)
BMS	Traditional – 2D	overcast	clearlow	<0.0001	0.51 (moderate)
	Traditional – 2D	clearhigh	clearlow	<0.0001	0.55 (moderate)
BMS360	Traditional – 360	overcast	clearlow	<0.001	0.45 (small)
	Traditional – 360	clearhigh	clearlow	<0.0001	0.52 (moderate)
BMSeq	Traditional – 360	overcast	clearlow	<0.0001	0.61 (moderate)
	Traditional – 360	clearhigh	clearlow	<0.0001	0.61 (moderate)
BMS360eq	Traditional – 360	overcast	clearlow	<0.0001	0.62 (moderate)
	Traditional – 360	clearhigh	clearlow	<0.0001	0.65 (moderate)
GBVS360eq	Traditional – 360	overcast	clearlow	<0.001	0.46 (small)
	Traditional – 360	clearhigh	clearlow	<0.001	0.44 (small)
UNISAL	Deep-L-2D	overcast	clearhigh	0.041	−0.26 (small)
	Deep-L-2D	clearhigh	clearlow	<0.01	0.33 (small)
SaltiNet	Deep-L-360	overcast	clearlow	0.012	0.31 (small)

which show extreme equator bias, raise questions about whether the changes in pitch (i.e., looking up and down) might have been greater if we had considered eye in addition to the head movements. While trying to determine the most appropriate ground truth mapping method, and given the numerous methods available, we should also note that the search for the best saliency prediction model often has to be accompanied by a search for the best ground truth as well: in other words, optimization has to take place at two levels (Droste et al., 2020; Kümmerer et al., 2016; Reddy et al., 2020, pp. 10241–10247). Obviously, this double optimization effort may lead to additional adaptations of the Gaussian functions and fixation filtering algorithms, which can be questioned because this can also deviate from human attention patterns. One should therefore be aware that the pre-processing of the data inevitably involves biases.

On the other hand, there is no doubt that viewing conditions in VR are different in nature from the conditions found in real environments. Beyond the issues of representation (pixel size, lack of dynamism), HMDs are limited in luminance range and are set to provide a visually comfortable environment, which is not necessarily the case in real environments where we may be disturbed and thus avoid regions of high luminance and glare (Sarey Khanie et al., 2017). Even though the tested scenes from *Rockcastle* and *Chamilothori*, initially created for a realistic perception of daylit spaces (Chamilothori et al., 2018), were rendered with software that has photometric accuracy, included tone-mapping correction and excluded any direct sun in the field of view (unlike Sarey Khanie et al. (2017) e.g.), they had not been developed for examining visual attention under different lighting conditions. At the same time, the saliency protocols commonly relying on 2D images screened on a computer, or on 360 stimuli viewed in VR, have also been neglecting the question of visual discomfort and how this could affect head and eye movements. The question of visual attraction in real conditions (where visual discomfort is not eliminated) thus deserves more attention in future studies.

5.2. Bringing saliency prediction to architectural design

As suggested by Arnheim in the context of artworks, the viewer does not have the simple mechanical role of recording visual stimulation provided by the work of art, but the fundamental task of giving meaning to it (Arnheim, 1965). Free observation is in itself a questionable posture given that humans are unlikely to be free from thoughts when looking at a given scene, especially in the framework of human subject testing. Further, the *Rockcastle* and *Chamilothori* datasets came from previous research on human perception to daylit architectural spaces. The protocol used in the original studies included ratings of spatial qualities (e.g., visual interest, pleasantness, scene complexity) occurring at the end of each exposure. From one scene to another, the questions were randomized but remained the same and we can speculate whether these questions may have guided some of the thoughts (and therefore head movements) of the study participants. Finally, the scenes of the *Chamilothori* dataset came from multiple experiments with within-subject factors: a given space was viewed multiple times by one participant

Table 7
Correlation coefficients of ground truth maps from human head movements when exposed to a space rendered for different sky conditions for the Rockcastle dataset.

Rockcastle	Sky1	Sky2	n[(a)]	Mean	SD	Min.	1Q[(b)]	Mdn[(b)]	3Q[(b)]	Max.
Full image	Overcast	Clear	8	**0.67**	**0.12**	0.55	0.59	0.61	0.68	0.90
Equatorial region	Overcast	Clear	8	**0.27**	**0.13**	0.08	0.20	0.26	0.33	0.48

[(a)] n: number of unique spaces tested for different sky types (note: in the original study "sky type" was a "between participant" variable).
[(b)] SD: standard deviation, Mdn: median, 1Q.: 1st quartile, 3Q.: 3rd quartile.

Table 8
Correlation coefficients of ground truth maps from human head movements when exposed to a space rendered for different sky conditions for the Chamilothori dataset.

Chamilothori	Sky1	Sky2	n[(a)]	Mean	SD	Min.	1Q[(b)]	Mdn[(b)]	3Q[(b)]	Max.
Full image	Overcast	Clearhigh	32	**0.92**	**0.05**	0.79	0.91	0.94	0.96	0.98
	Overcast	Clearlow	32	**0.92**	**0.07**	0.66	0.91	0.94	0.96	0.98
	Clearhigh	Clearlow	32	**0.90**	**0.08**	0.53	0.88	0.92	0.96	0.98
Equatorial region	Overcast	Clearhigh	32	**0.90**	**0.07**	0.71	0.88	0.93	0.94	0.97
	Overcast	Clearlow	32	**0.90**	**0.10**	0.50	0.89	0.92	0.94	0.98
	Clearhigh	Clearlow	32	**0.87**	**0.12**	0.33	0.84	0.90	0.94	0.98

[(a)] n: number of unique spaces (including room type, window size pattern and context) tested for different sky types (note: in the original study "sky type" was a "between participant" variable).
[(b)] SD: standard deviation, Mdn: median, 1Q.: 1st quartile, 3Q.: 3rd quartile.

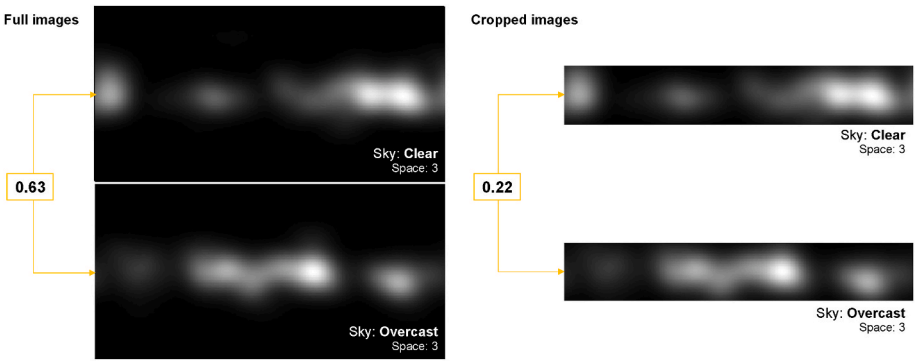

Fig. 12. Example of correlation coefficient between ground truth maps across different sky types for one unique space taken from the Rockcastle's dataset for full and central (or cropped) images regions.

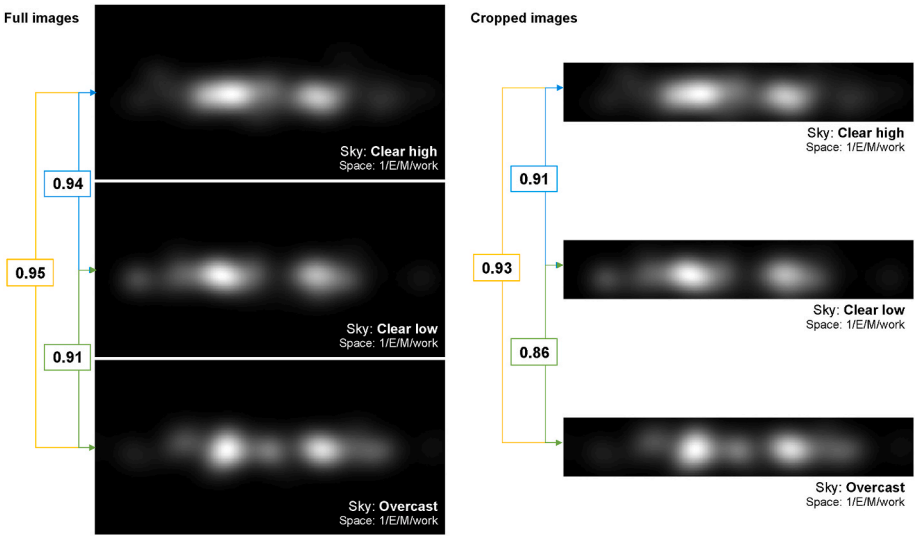

Fig. 13. Example of correlation coefficient between ground truth maps across different sky types for one unique space (including variations of space, window size, pattern, and context) taken from the Chamilothori's dataset for full and central (or cropped) images regions.

with only some elements (e.g., façade geometry or window sizes) that would change across scenes. We can therefore raise the question of whether the visual attention was strongly linked to the elements that changed from one scene to another (i.e., the windows and façade geometry and their corresponding light patterns), hereby bringing one more influencing factor in the "free-viewing" of the scenes.

The modelling of saliency prediction itself does not come from the field of design. Therefore, while the idea of being able to predict the visual attention of humans is both exciting and appealing in the framework of architectural design, such models must be examined critically and used with caution in this new context. First, one must account for which stimuli these models were originally created or trained, especially given that the proper application of saliency models on complex images may be difficult for non-experts. A workaround could be to integrate them into design tools made available to architects. However, this would bring their application even further away from the context in which they were developed, with clearly insufficient evidence right now about their actual applicability and/or relevance for this new context. Finally, as the performance of the models depends on the composition of the scenes, such, as our results show, on the lighting conditions in the scene, different saliency models may be needed for different types of spaces. On a positive note, the fact that some of the traditional models were found to be more effective for bare spaces (i.e., without people or highly informative objects) does offer promise about models with a reasonable prediction accuracy for design purposes in the future.

5.3. *Effect of sky type on head movement*

We found saliency prediction models to be more accurate for overcast and clear sky with high sun positions than for clear sky types with low sun positions. This output is not that surprising given that large sun patches in the space can be detected by algorithms as salient regions (even though they remain limited in the scene information they provide).

With respect to H3 and the correlation in head movements for the same space with different daylight conditions, results showed that the outcomes for the *Rockcastle* dataset were different from the *Chamilothori* dataset, with the former showing low correlation, while the latter showed strong correlations for all considered comparisons (i.e., between the three sky types) for the equatorial image region. As observed in section 4.2, the prediction of saliency models was lowest for the *Rockcastle* dataset, which we attributed partly to the minimalist nature of the scenes (devoid of color, texture and objects) and partly to the quality of the free visualization period which was not part of the original protocol and which depended mainly on the time needed for the participant to be ready to answer the questions. Given these two observations, the dataset of *Chamilothori* and *Rockcastle* are different and certainly the one from Rockcastle is the least suitable for interpretation. By contrast, the scene content and rendering procedure used in the study of Chamilothori et al.

(2018) included colors, textures and objects and was validated for its realism while the silent free-viewing period was rigorously implemented in the experimental protocol. The analysis conducted on this dataset showed a strong correlation between head movements for the same space under different sky types, aligning with the results from Vincent et al. (2009) who concluded that light sources and shadows, although highly visible, are not very informative, thus questioning our attraction to these features.

However, in a different analysis based on a subset of the *Chamilothori* dataset (n = 3 spaces), linear mixed models were used to look at *differences* (and not similarities) in the number of fixations towards the floor for the same space across the three different sky conditions (Karmann et al., 2021) and a statistically significant difference ($p < 0.05$) was found in the percentage of fixations towards the floor when comparing overcast sky conditions to clear sky conditions with a high sun position. No significant difference was found when comparing fixations across other sky types (i.e., clear sky with low sun position, where sun patches take over most of the floor area). This previous analysis was limited in its sample size, and thus additional analyses, ideally with other datasets as well, might be necessary to bring conclusive statements on the impact of shading patterns on visual attention.

Overall, it should be noted that as neither of the *Rockcastle* or *Chamilothori* experimental protocols were designed to investigate correlations between head (or eye) movements and lighting distribution, it would be premature to draw any conclusion about H3, though the analyses of the present study do show the potential of the proposed approach to investigate this in more appropriate datasets. Our findings should be considered as a promising starting point to enable further investigations rather than as absolute conclusions about the influence of sky conditions on view behavior.

5.4. Linking head-tracking with affective evaluations of the space

The studies behind the *Chamilothori* dataset (Chamilothori et al., 2022a; Moscoso et al., 2021) included questions related to participant's subjective perception of the space following Russel's circumplex model of affect (pleasantness, interestingness, excitement, calmness) as well as additional spatial features (complex, bright, spacious). In these studies, the sky type, space function (suggested by the furniture in the scene), and the type of shading patterns employed were the independent variables. While neither the sky type nor the spatial context influenced space impressions, the type of shading patterns influenced both the affective appraisal and the visual appearance of the space, e.g., rendering the same space more pleasant, interesting, or bright. However, in our ground truth data, we found a strong correlation between head movements for the different types of shading patterns (as in the difference between sky types), suggesting that participants were very much fixated on the windows, no matter the shadow casted on the floor. This finding shows that, at least with this analysis, differences in the affective evaluation of the scenes were not reflected in the participants' head movements.

A different approach that is beyond the focus of this paper would be to derive metrics directly from head tracking data (e.g., fixation time, area, saccade, entropy) and link these to the recorded emotions. This method was used in the work of Batool et al. (2022) who studied head movement in relation to different types of view out and found that natural scenes were characterized by lower numbers of fixations and saccades, and longer fixation durations, compared to urban views, but that for both types, the most preferred scenes led to more fixations and saccades. Analyses relying on ground truth maps might not have revealed the subtleties of these outcomes, and more work would be needed in the future to examine the relationship between such derived metrics and subjective responses.

6. Conclusion

Saliency prediction, which provides an estimate of where, in a given scene, people are likely to devote their visual attention, offers a high potential in becoming a useful tool to inform architectural design. In fact, predicting viewing patterns in a space can aid environmental legibility, improve way-finding, support research related to spatial appraisal and cognitive restoration, and thus contribute to the design of more socially responsive spaces. At this stage, we do not know the reliability and applicability of existing saliency prediction models in daylight architectural design, and particularly in interior scenes that do not contain people or animals, which are known to be highly salient. The present study examined the performance of 11 saliency prediction models on rendered architectural scenes viewed in VR. The models were chosen based on their diversity and availability, and consisted of both traditional and deep-learning (DL) methods that had been developed and/or were adapted to perspective images and 360° visual stimuli. Three datasets of rendered scenes were used: the *Rockcastle* dataset, containing interior spaces rendered in black and white with neither objects nor texture, the *Chamilothori* dataset, containing interior spaces rendered in color with some furniture and texture, and the *Sitzmann* dataset, containing color rendered spaces with higher levels of detail and texture, but hardly any people visible for the chosen subset of scenes. We utilized the Pearson Correlation Coefficient (CC) between saliency prediction maps and ground truth maps as a measure of prediction accuracy, where the ground truth maps were derived from head movement data collected from participants in VR. In order to account for equator bias, we examined both the full 360° scenes and the equatorial image region. The *Rockcastle* and *Chamilothori* datasets, initially developed for daylight research, included multiple space variations rendered for different sky conditions, which also allowed us, for the first time, to examine the effect of daylight conditions on saliency prediction performance. One overarching – and unsurprising – finding was that the equatorial bias was very present in both the participants' head movements and in the saliency models. To address the possible inflation of results based on this bias, we chose to focus mostly on the equatorial image region (± 30° around the central horizon line) in further analyses.

Saliency models developed from or adapted to 360° scenes commonly outperformed models developed for perspective images when applied to 360° scenes. The performance of traditional models, which focus on low-level bottom-up features such as direction and luminance contrasts, and of DL models, which are able to predict attraction towards high-level information such as objects and people, strongly depended on the information present in the scenes. We notably found that BMS360eq and GBVS360eq, two traditional models adapted to 360°, were the most promising for the *Chamilothori* dataset. For this dataset, saliency prediction models also showed a significantly higher performance for the scenes rendered with an *overcast sky* or with a *clear sky with high sun position* compared to the same scenes rendered with a *clear sky with low sun position*, suggesting that the large shadows and high contrasts resulting from this sky type appear to be misinterpreted in saliency prediction, especially by traditional models.

When we compared the truth maps for one same space illuminated by different types of sky, we found a strong correlation between sky types for the *Chamilothori* dataset, even when considering only the equatorial part of the image. This result suggests that participants did not substantially alter their head movements while viewing spaces under different types of sky, a finding whose generalization would require further investigations using datasets specifically generated to address these questions.

Saliency modelling requires some knowledge in computer science, and, depending on the complexity of the models, significant computing power. Traditional models are the easiest and least costly. The findings of the present study show promising performance for some of these models for 360° colored architectural spaces with basic furniture and

textures, and suggest the prospect of their possible integration into design tools, which could further support human-centric design in architecture.

Conflicts of interest

Potential conflict of interest exists:

We wish to draw the attention of the Editor to the following facts, which may be considered as potential conflicts of interest, and to significant financial contributions to this work:

The nature of potential conflict of interest is described below:

No conflict of interest exists.

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Funding

Funding was received for this work.

All of the sources of funding for the work described in this publication are acknowledged below:

[List funding sources and their role in study design, data analysis, and result interpretation].

This work was supported by the École Polytechnique Fédérale de Lausanne (all authors) and by the Swiss National Science Foundation via the Sinergia grant [CRSII5-18035] (Bahar Aydemir). The Chamilothori and Rockcastle datasets have been collected with the additional support of the VELUX Stiftung grants [1022] and [936]. This work was also conducted in part at the Karlsruhe Institute of Technology (Caroline Karmann), the Eindhoven University of Technology (Kynthia Chamilothori) and the Korea University (Seungryong Kim).

No funding was received for this work.

Intellectual property

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

Research ethics

We further confirm that any aspect of the work covered in this manuscript that has involved human patients has been conducted with the ethical approval of all relevant bodies and that such approvals are acknowledged within the manuscript.

IRB approval was obtained (required for studies and series of 3 or more cases).

Written consent to publish potentially identifying information, such as details or the case and photographs, was obtained from the patient(s) or their legal guardian(s).

Authorship

The International Committee of Medical Journal Editors (ICMJE) recommends that authorship be based on the following four criteria.

1. Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; AND
2. Drafting the work or revising it critically for important intellectual content; AND
3. Final approval of the version to be published; AND
4. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

All those designated as authors should meet all four criteria for authorship, and all who meet the four criteria should be identified as authors. For more information on authorship, please see http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html#two.

All listed authors meet the ICMJE criteria.

We attest that all authors contributed significantly to the creation of this manuscript, each having fulfilled criteria as established by the ICMJE.

One or more listed authors do (es) not meet the ICMJE criteria.

We believe these individuals should be listed as authors because:

[Please elaborate below]

We confirm that the manuscript has been read and approved by all named authors.

We confirm that the order of authors listed in the manuscript has been approved by all named authors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Asst Prof. Siobhan Rockcastle (University of Oregon) as well as Prof. Barbara Matusiak and Dr. Claudia Moscoso (Norwegian University of Science and Technology) for agreeing to share, respectively, the Rockcastle and Chamilothori datasets for the purposes of the present study.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jenvp.2023.102110.

References

Abboushi, B., Elzeyadi, I., Taylor, R., & Sereno, M. (2019). Fractals in architecture: The visual interest, preference, and mood response to projected fractal light patterns in interior spaces. *Journal of Environmental Psychology, 61*, 57–70. https://doi.org/10.1016/j.jenvp.2018.12.005

Achanta, R., & Süsstrunk, S. (2010). Saliency detection using maximum symmetric surround. In *In 2010 IEEE international Conference on image processing* (pp. 2653–2656). IEEE. https://doi.org/10.1109/ICIP.2010.5652636.

Arnheim, R. (1965). *Art and visual perception: A psychology of the creative eye*. Univ of California Press. https://doi.org/10.2307/426441

Arnheim, R. (1977). *The dynamics of architectural form*. Univ of California Press. https://doi.org/10.2307/989519

Balbi, B., Protti, F., & Montanari, R. (2016). *Driven by caravaggio through his painting*. Cognitive. https://www.iaria.org/conferences2016/COGNITIVE16.html.

Batool, A., Rutherford, P., McGraw, P., Ledgeway, T., & Altomonte, S. (2022). Gaze correlates of view preference: Comparing natural and urban scenes. *Lighting Research and Technology, 54*(6), 576–594. https://doi.org/10.1177/14771535211055703

Borji, A. (2018). *Saliency prediction in the deep learning era: An empirical investigation*. https://doi.org/10.48550/arXiv.1810.03716

Borji, A., & Itti, L. (2015). *Cat 2000: A large scale fixation dataset for boosting saliency research*. https://doi.org/10.48550/arXiv.1505.03581

Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability. *Vision Research, 116*, 165–178. https://doi.org/10.1016/j.visres.2015.03.005

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2016). *"What do different evaluation metrics tell us about saliency models?"*. https://doi.org/10.48550/arXiv.1604.03605

Caduff, D., & Timpf, S. (2008). On the assessment of landmark salience for human navigation. *Cognitive Processing, 9*, 249–267. https://doi.org/10.1007/s10339-007-0199-2

Chamilothori, K., Chinazzo, G., Rodrigues, J., Dan-Glauser, E., Wienold, J., & Andersen, M. (2019). *Subjective and physiological responses to façade and sunlight pattern geometry in virtual reality*. Building and Environment. https://doi.org/10.1016/j.buildenv.2019.01.009

Chamilothori, K., Wienold, J., & Andersen, M. (2018). Adequacy of immersive virtual reality for the perception of daylit spaces: Comparison of real and virtual environments. *Leukos*, 1–24. https://doi.org/10.1080/15502724.2017.1404918

Chamilothori, K., Wienold, J., Moscoso, C., Matusiak, B., & Andersen, M. (2022a). *Regional differences in the perception of daylit scenes across europe using virtual reality. Part II: Effects of façade and daylight pattern geometry*. https://doi.org/10.1080/15502724.2021.1999257. *LEUKOS* 1–25.

Chamilothori, K., Wienold, J., Moscoso, C., Matusiak, B., & Andersen, M. (2022b). Subjective and physiological responses towards daylit spaces with contemporary façade patterns in virtual reality: Influence of sky type, space function, and latitude. *Journal of Environmental Psychology, 82*, Article 101839. https://doi.org/10.1016/j.jenvp.2022.101839

Corrodi, M., & Spechtenhauser, K. (2014). Illuminating. In *Illuminating. Birkhäuser*. https://doi.org/10.1515/9783038216414

Cottet, M., Vaudor, L., Tronchère, H., Roux-Michollet, D., Augendre, M., & Brault, V. (2018). Using gaze behavior to gain insights into the impacts of naturalness on city dwellers' perceptions and valuation of a landscape. *Journal of Environmental Psychology, 60*, 9–20. https://doi.org/10.1016/j.jenvp.2018.09.001

Dong, W., Tong, Q., Liao, H., Liu, Y., & Liu, J. (2020). Comparing the roles of landmark visual salience and semantic salience in visual guidance during indoor wayfinding. *Cartography and Geographic Information Science, 47*(3), 229–243. https://doi.org/10.1080/15230406.2019.1697965

Droste, R., Jiao, J., & Alison Noble, J. (2020). Unified image and video saliency modeling, 419–35. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer vision – eccv 2020*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-58558-7_25.

Dupont, L., Kristien Ooms, Antrop, M., & Van Eetvelde, V. (2016). Comparing saliency maps and eye-tracking focus maps: The potential use in visual impact assessment based on landscape photographs. *Landscape and Urban Planning, 148*, 17–26. https://doi.org/10.1016/j.landurbplan.2015.12.007

Fan, S., Jiang, M., Shen, Z., Koenig, B. L., Kankanhalli, M. S., & Qi, Z. (2017). The role of visual attention in sentiment prediction. In *Proceedings of the 25th ACM international conference on Multimedia* (pp. 217–225). https://doi.org/10.1145/3123266.3123445

Fan, S., Shen, Z., Jiang, M., Koenig, B. L., Xu, J., Kankanhalli, M. S., & Qi, Z. (2018). Emotional attention: A study of image sentiment and visual attention, 7521–31. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*. https://doi.org/10.1109/CVPR.2018.00785.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*(5), 532–538. https://psycnet.apa.org/doi/10.1037/a0015808.

Foulsham, T., Walker, E., & Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Research, 51*(17), 1920–1931. https://doi.org/10.1016/j.visres.2011.07.002

Genetics of Design. (2020a). Empathy in design: Measuring how faces make places. *The Genetics of Design*. Retrieved https://geneticsofdesign.com/2020/06/21/empathy-in-design-measuring-how-faces-make-places/. (Accessed 7 July 2020).

Genetics of Design. (2020b). "Empathy in design: Measuring the impact of biophilia.". *The Genetics of Design*. Retrieved https://geneticsofdesign.com/2020/05/26/empathetic-design-measuring-the-impact-of-biophilia/. (Accessed 7 July 2020).

Gutiérrez, J., David, E., Rai, Y., & Le Callet, P. (2018). Toolbox and dataset for the development of saliency and scanpath models for omnidirectional/360° still images. *Signal Processing: Image Communication, 69*, 35–42. https://doi.org/10.1016/j.image.2018.05.003

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. In *Advances in neural information processing systems* (pp. 545–552). https://doi.org/10.7551/mitpress/7503.003.0073

Hasse, C., & Weber, R. (2012). Eye movements on facades: The subjective perception of balance in architecture and its link to aesthetic judgment. *Empirical Studies of the Arts, 30*(1), 7–22. https://doi.org/10.2190/EM.30.1.c

Hollander, J. B., Purdy, A., Wiley, A., Foster, V., Jacob, R. J. K., Taylor, H. A., & Brunyé, T. T. (2019). Seeing the city: Using eye-tracking Technology to explore cognitive responses to the built environment. *J. Urbanism: Int. Res. Placemaking and Urban Sustain., 12*(2), 156–171. https://doi.org/10.1080/17549175.2018.1531908

Hollander, J. B., Sussman, A., Levering, A. P., & Foster-Karim, C. (2020). Using eye-tracking to understand human responses to traditional neighborhood designs. *Planning Practice and Research*, 1–25. https://doi.org/10.1080/17549175.2018.1531908

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259. https://doi.org/10.1109/34.730558

Kaplan, S. (1995). The restorative benefits of nature: Toward an integrative framework. *Journal of Environmental Psychology, 15*(3), 169–182. https://doi.org/10.1037/030621

Kaplan, R., & Kaplan, S. (1989). *The experience of nature: A psychological perspective*. CUP Archive. https://doi.org/10.1016/0272-4944(95)90001-2

Karmann, C., Chamilothori, K., Schoenmakers, S., Aydemir, B., & Andersen, M. (2021). Virtual reality to assess visual attraction and perceived interest to daylit scene variations. In *Anfa 2021: Quantified buildings, quantified self. La Jolla, CA; USA*. https://www.researchgate.net/publication/354862437_Virtual_reality_to_assess_visual_attraction_and_perceived_interest_to_daylit_scene_variations.

Khanie, S., Mandana, J. S., Einhäuser, W., Wienold, J., & Andersen, M. (2017). Gaze and discomfort glare, Part 1: Development of a gaze-driven photometry. *Lighting Research and Technology, 49*(7), 845–865. https://doi.org/10.1177/1477153516649016

Koseoglu, E., & Onder, D. E. (2011). Subjective and objective dimensions of spatial legibility. *Proc. - Soc. Behav. Sci., 30*, 1191–1195. https://doi.org/10.1016/j.sbspro.2011.10.231

Köster, H. (2004). *Dynamic daylighting architecture: Basics, systems, projects*. Springer Science & Business Media. https://books.google.de/books?id=zSNs3qIEE7MC&printsec=frontcover&redir_esc=y#v=onepage&q&f=false.

Kroner, A., Senden, M., Kurt, D., & Goebel, R. (2020). Contextual encoder-decoder network for visual saliency prediction. *Neural Networks*. https://doi.org/10.1016/j.neunet.2020.05.004

Kümmerer, M., Theis, L., & Bethge, M. (2014). *"Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet."*. https://doi.org/10.48550/arXiv.1411.1045. *ArXiv Preprint ArXiv:1411.1045*.

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2016). *"DeepGaze II: Reading fixations from deep features trained on object recognition."*. https://doi.org/10.48550/arXiv.1610.01563. *ArXiv Preprint ArXiv:1610.01563*.

Lebreton, P., & Alexander, R. (2018). GBVS360, BMS360, ProSal: Extending existing saliency prediction models from 2D to omnidirectional images. *Signal Processing: Image Communication, 69*, 69–78. https://doi.org/10.1016/j.image.2018.03.006

Leslie, R. P. (2003). Capturing the daylight dividend in buildings: Why and how? *Building and Environment, 38*(2), 381–385. https://doi.org/10.1016/S0360-1323(02)00118-X

McCarter, R., & Pallasmaa, J. (2012). *Understanding architecture: A primer on architecture as experience*. Phaidon Press Limited. https://www.phaidon.com/store/architecture/understanding-architecture-9780714848099/.

Monroy, R., Lutz, S., Chalasani, T., & Aljosa Smolic. (2018). SalNet360: Saliency maps for omni-directional images with CNN. In *Signal processing: Image communication*. https://doi.org/10.1016/j.image.2018.05.005

Moscoso, C., Chamilothori, K., Wienold, J., Andersen, M., & Matusiak, B. (2021). Regional differences in the perception of daylit scenes across europe using virtual reality. Part I: Effects of window size. *Leukos*, 1–22. https://doi.org/10.1080/15502724.2020.1854779

Noland, R. B., Weiner, M. D., Gao, D., Cook, M. P., & Nelessen, A. (2017). Eye-tracking Technology, visual preference surveys, and urban design: Preliminary evidence of an effective methodology. *J. Urbanism: Int. Res. Placemaking and Urban Sustain., 10*(1), 98–110. https://doi.org/10.1080/17549175.2016.1187197

Parpairi, K., Baker, N. V., Steemers, K. A., & Compagnon, R. (2002). The luminance differences index: A new indicator of user preferences in daylit spaces. *Lighting Research and Technology, 34*(1), 53–66. https://doi.org/10.1191/1365782802li030oa

Rai, Y., Gutiérrez, J., & Le Callet, P. (2017). A dataset of head and eye movements for 360 degree images. In *Proceedings of the 8th ACM on multimedia systems conference* (pp. 205–210). ACM. https://doi.org/10.1145/3083187.3083218.

Reddy, N., Jain, S., Yarlagadda, P., & Gandhi, V. (2020). *Tidying deep saliency prediction architectures*. IEEE. https://doi.org/10.1109/IROS45743.2020.9341574

Reina, A., Marc, X. G.-i-N., McGuinness, K., Noel, E., & O'Connor. (2017). Saltinet: Scan-Path prediction on 360 degree images using saliency volumes. In *Proceedings of the IEEE international conference on computer vision* (pp. 2331–2338).

Rockcastle, S., Ámundadóttir, M. L., & Andersen, M. (2017). Contrast measures for predicting perceptual effects of daylight in architectural renderings. *Lighting Research and Technology, 49*(7), 882–903. https://doi.org/10.1177/1477153516644292

Rockcastle, S. F., Chamilothori, K., & Andersen, M. (2017a). An experiment in virtual reality to measure daylight-driven interest in rendered architectural scenes. *Proceedings of Building Simulation 2017: 15th Conference of IBPSA*. https://doi.org/10.26868/25222708.2017.828

She, D., Yang, J., Cheng, M.-M., Lai, Y.-K., Rosin, P. L., & Wang, L. (2020). WSCNet: Weakly supervised coupled networks for visual sentiment classification and detection. *IEEE Transactions on Multimedia, 22*(5), 1358–1371. https://doi.org/10.1109/TMM.2019.2939744

Sitzmann, V., Ana Serrano, Pavel, A., Agrawala, M., Gutierrez, D., Masia, B., & Gordon, W. (2018). Saliency in VR: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics, 24*(4), 1633–1642. https://doi.org/10.1109/TVCG.2018.2793599

Steemers, K., & Ann Steane, M. (2012). *Environmental diversity in architecture*. London and New York: Routledge. https://doi.org/10.4324/9780203561270

Truong, Q.-T., & Lauw, H. W. (2017). Visual sentiment analysis for review images with item-oriented and user-oriented CNN. In *Proceedings of the 25th ACM international conference on Multimedia* (pp. 1274–1282). https://doi.org/10.1145/3123266.3123374

Upenik, E., & Ebrahimi, T. (2017). A simple method to obtain visual attention data in head mounted virtual reality. In *2017 IEEE international conference on multimedia expo workshops (ICMEW)* (pp. 73–78). https://doi.org/10.1109/ICMEW.2017.8026231

Van der Jagt, Alexander, P. N., Craig, T., Brewer, M. J., & Pearson, D. G. (2017). A view not to Be missed: Salient scene content interferes with cognitive restoration. *PLoS One, 12*(7), Article e0169997. https://doi.org/10.1371/journal.pone.0169997

Vincent, B. T., Baddeley, R., Correani, A., Tom, T., & Leonards, U. (2009). Do we look at lights? Using mixture modelling to distinguish between low-and high-level factors in natural image viewing. *Visual Cognition, 17*(6–7), 856–879. https://doi.org/10.1080/13506280902916691

Wang, C., Chen, Y., Zheng, S., & Liao, H. (2018). Gender and age differences in using indoor maps for wayfinding in real environments. *ISPRS International Journal of Geo-Information, 8*(1), 11. https://doi.org/10.3390/ijgi8010011

Wang, Z., Liang, Q., Duarte, F., Zhang, F., Charron, L., Johnsen, L., Cai, B., & Ratti, C. (2019). Quantifying legibility of indoor spaces using deep convolutional neural

networks: Case studies in train stations. *Building and Environment, 160*, Article 106099. https://doi.org/10.1016/j.buildenv.2019.04.035

Weber, R., Choi, Y., & Stark, L. (1995). *The impact of formal properties on eye movement during the perception of architecture*. https://doi.org/10.35483/ACSA.Intl.1995.32. Lisbon.

Xu, R., & Stephen, W. (2014). Visual assessment of BIPV retrofit design proposals for selected historical buildings using the saliency map method. *Journal of Facade Design and Engineering, 2*(3–4), 235–254. https://doi.org/10.7480/jfde.2014.3-4.911

Xu, R., Xia, H., & Tian, M. (2020). Wayfinding design in transportation architecture – are saliency models or designer visual attention a good predictor of passenger visual attention? *Front. Arch. Res.*. https://doi.org/10.1016/j.foar.2020.05.005

Zhang, J., & Sclaroff, S. (2013). Saliency detection: A boolean map approach. In *Proceedings of the IEEE international conference on computer vision* (pp. 153–160). https://doi.org/10.1109/ICCV.2013.26

Zheng, H., Chen, T., You, Q., & Luo, J. (2017). When saliency meets sentiment: Understanding how image content invokes emotion and sentiment. In *2017 IEEE international conference on image processing (ICIP)* (pp. 630–634). IEEE. https://doi.org/10.1109/ICIP.2017.8296357.