Journal of Building Engineering 97 (2024) 110678

Contents lists available at ScienceDirect

Journal of Building Engineering

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

Comparative analysis of saliency map algorithms in capturing visual priorities for building inspections

Muhammad Rakeh Saleem [a], Rebecca Napolitano [a,*]

[a] Department of Architectural Engineering, The Pennsylvania State University, University Park, 16802, PA, United States

ARTICLE INFO

Keywords:
Eye tracking
Fixation maps
Visual attention
Computer vision
Building inspection
Artificial intelligence
Human-building interactions

ABSTRACT

This study investigates the efficacy of saliency mapping algorithms in capturing the visual priorities of building inspectors for structural damage assessment. Our work established a ground truth dataset by implementing eye-tracking technology to capture the gaze patterns of building inspectors. Further, it enables a detailed evaluation of the saliency models' ability to reflect experts' visual attention during inspection tasks. Our comparative analysis assesses the performance of three saliency models— EnDec, DeepGaze, and SALICON— against this ground truth data, using conventional saliency metrics such as Area under the Curve, Similarity, Normalized Scanpath Saliency, Correlation Coefficient, and Kullback-Leibler Divergence. Our findings reveal that while the SALICON model demonstrates a marginally better performance and highlights areas where these models fall short, particularly in accurately reflecting the critical visual properties of inspectors, this insight is crucial for advancing the field. By highlighting these limitations, we have drawn attention to the need for developing more specialized saliency models tailored to the unique demands of building inspection tasks. Thus, the study not only fulfills its objectives of comparative analysis but also contributes to the broader discourse on improving automated structural inspection systems. This study highlights the need to develop specialized computer vision models to address specific building inspection challenges. By identifying strengths and improvement areas, this research contributes valuable insights and highlights the potential and current limitations of applying computer vision techniques to real-world building inspection tasks.

1. Introduction

Preliminary damage assessment and structural inspection are crucial for routine building inspections or evaluating the damage caused by natural disasters [1]. For damage assessment or initial response following post-disaster, it is critical to ensure the safety of the built environment [2,3]. Traditional structural inspections are typically conducted by experienced and knowledgeable engineers, who make assessments based on visual observations of the damage to a structure, following strict procedures and guidelines. While most inspection processes are conducted by manual visual inspection, human resources are often limited following the post-disaster recovery process due to safety concerns and the associated cost [4,5].

In recent years, unmanned aerial vehicles (UAV) have been used widely for damage assessment and structural inspections [6–8]. Previous works have explored the application of UAVs to enhance infrastructure inspection tasks such as building monitoring [9], railway inspection [10], structural component recognition [11], and LIDAR-based bridge inspection [12,13]. Further, artificial

* Corresponding author.
E-mail addresses: rakeh@psu.edu (M.R. Saleem), nap@psu.edu (R. Napolitano).

https://doi.org/10.1016/j.jobe.2024.110678
Received 15 February 2024; Received in revised form 20 August 2024; Accepted 4 September 2024
Available online 10 September 2024

intelligence (AI) based methods have increasingly gained popularity for the past two decades [14–17]. But where should a UAV look for damage, and how can it guarantee accurate damage reconnaissance? To achieve this, we need to transfer the expert knowledge of human inspectors to UAVs, and the UAV should be capable of identifying the salient damage from a structure.

Convolutional neural networks (CNNs) have succeeded in fields like image recognition and computer vision [4]. Nevertheless, the opacity of the CNN model raises concerns about their generalizability and which features contribute to their decisions. Contemporary methods and analyses must address the challenge of machine learning interpretability more thoroughly. Current research concentrates on a limited array of issues, resulting in sparse practical guidance [18]. Some studies [19] emphasize defining taxonomies and understanding interpretability mechanisms, offering critical reviews of machine learning algorithms. Lipton [20] outlined the deficiencies that should be addressed to ensure that the algorithms perform predictably. Some evaluative aspects have been discussed in Refs. [18,21]; however, they do not extensively cover the interpretability of a deep neural network (DNN) model.

Building upon the foundation of using CNNs for damage detection, recent strides have been made in leveraging saliency maps as a means of enhancing detection capabilities [22,23]. Saliency maps provide a visual representation of the most relevant regions within an image, thereby offering insights into where the network focuses its attention during the decision-making process. While saliency maps have shown promise in various applications, their utilization, specifically for damage detection, remains largely unexplored. A normal saliency map provides a visual representation of how much each point of an input image stands out with respect to its neighboring points. The extent to which saliency maps accurately represent the areas impacting network decisions is still being determined [24]. By examining if existing saliency map algorithms can identify the visual features most important to building inspectors, this research seeks to enhance the efficiency of building inspections.

To examine the efficacy of saliency maps in accurately capturing the critical visual features for building inspection [25], eye-tracking technology can serve as a valuable benchmark. By tracking the gaze patterns of experienced human inspectors as they assess structural damage, we can establish a ground truth dataset that reflects the regions of interest deemed significant by human experts. This ground truth dataset can then be utilized to evaluate the correspondence between the areas highlighted by saliency maps and those attended to by human inspectors. Such comparative analysis not only provides a means to quantify the accuracy of saliency maps but also offers insights into potential discrepancies or areas for improvement. In eye tracking research, heat maps and attention maps show the average of the continuous fixation locations where participants focus more on their attention time and are viewed the most. Fig. 1 shows a typical saliency predictor with a test image highlighting the saliency map. This work seems to do an in-depth exploration to understand the gap between saliency mapping predictors and ground truth fixation data to capture inspectors' visual priorities for building damage assessment. This exploration holds the potential to not only improve the accuracy of damage reconnaissance but also to shed light on the intricate features and patterns indicative of structural damage, thus advancing the field of automated structural inspection and disaster response.

This research is a novel contribution as it addresses the performance of various saliency map algorithms within the context of building inspections leveraging eye tracking technology. The primary research question centers on comparing the ability of these algorithms to identify and represent visual features deemed important by building inspectors during assessments. Findings suggest that conventional saliency models do not adequately address the complex requirements of building inspections [26], highlighting the necessity for specialized data and models tailored to this domain. This research contributes valuable insights into applying computer vision methods in real-world inspection scenarios, with implications for enhancing inspection accuracy and efficiency.

2. Related work

The current state-of-the-art analysis methods do not comprehensively address machine learning interpretability. Instead, they focus on a narrow subset of issues, so only limited guidance can be extracted. An attempt to introduce model interpretability and to use it to optimize performance was made by Ref. [27]. The approach was successful in reconstructing and visualizing features of the input image that had been identified by the intermediate layers of a network. Saliency maps are probably the most popular technique for visually explaining CNN's decision [28]. In the last decade, saliency prediction has been widely studied. As presented in Ref. [29], the visualizations provided by this approach help explain the failure of CNNs, identify biases present in the datasets, and prepare models

Fig. 1. Comparison between a normal fixation map and conventional saliency map.

that are robust against adversarial attacks. Therefore, they offer an improved development process and greater generalization of trained models.

Several eye tracking datasets have been recently constructed and shared in the community to understand visual attention and to build computational saliency models. An eye-tracking dataset includes natural images as the visual stimuli and eye movement data recorded using eye tracking devices. Most datasets have their own distinguishing features in image selection. For example, POET [30], the largest dataset we know by far, contains 6270 images and is only viewed by five subjects. The MIT dataset [31] is more general due to its relatively large size, i.e., 1003 images, and the generality of the image source. The OSIE dataset [6] features multiple dominant objects in an image to facilitate object and semantic saliency. The experimental requirements inherently limit the scale of the current datasets. The present work leverages the SALICON eye tracking dataset, which was not explicitly captured for building damage analysis, to train and compute the saliency maps. The ground truth dataset, derived from human expertise while doing building assessments, will serve as a benchmark against which the accuracy of existing saliency datasets and models can be evaluated.

Methods for comparing saliency maps and ground truth are presented in Ref. [32]. In this study, the authors performed two types of randomization tests. The first focused on the randomization of a model, and the second randomized labels in a training dataset to check the performance of saliency map algorithms on a correctly labeled test dataset. With the advent of models using deep neural networks [28,33–38], the saliency prediction has been improved remarkably. The first model [39] to use ensembles of the deep network (eDN) trained from scratch to predict saliency cannot scale to outperform the current state-of-the-art due to limited data. Kummerer et al. [38] addressed this issue by reusing existing neural networks trained for image classification to predict fixation maps. Subsequently, it was found that DNN trained on object recognition (AlexNet [40] trained on VGG-16 [41]) could significantly outperform training from scratch [36]. Liu et al. [42] presented a multi-resolution CNN trained from image regions centered on fixation and non-fixation locations at multi-scales. The SALICON model [43] fine-tunes a mixture of deep features from AlexNet, ImageNet [20], and GoogleNet [21] for saliency prediction using the SALICON and OSIE datasets.

3. Methodology

3.1. Comparison metrics for saliency maps and ground truth datasets

In this paper, we study saliency metrics functions that take two inputs representing eye fixations— ground truth fixation map based on task-specific eye tracking data and predicted saliency map based on task-agnostic eye tracking data, and output a number assessing their similarity or dissimilarity. Given ground truth gaze fixations, these comparison metrics are used to define scoring functions, which take a saliency map prediction as input and return the score, assessing the accuracy of the prediction. We consider the five most common saliency evaluation metrics, as shown in Table 1. While some metrics have been designed specifically for saliency evaluation (AUC [44], normalized scanpath saliency [45]), others have been adapted from signal detection (variants of AUC [46]), image matching and retrieval (Similarity [47]), information theory (KL-divergence [47]) and statistics (Pearson's correlation coefficient [48]). Because of their original intended applications, these metrics expect different input formats: Kullback-Leibler divergence expects valid probability distributions as input, and Similarity can operate on unnormalized densities and histograms. At the same time, Pearson's Correlation Coefficient (CC) treats its inputs as random variables.

The MIT Saliency Benchmark [31] interprets metric scores and different methods. It accepts saliency maps as intensity maps without restricting input to any particular form (probabilistic or otherwise). If a metric expects valid probability distributions, we simply normalize the input saliency maps without additional modifications or optimizations. Different metrics use different formats of ground truth for evaluating saliency models. Location-based metrics consider saliency map values at discrete fixation locations, while distribution-based metrics treat ground truth fixation maps and saliency maps as continuous distributions. To evaluate the performance of the saliency metric, Eq. (1) shows the relation that involves comparing predicted saliency maps and ground truth eye fixations. The fundamental concept behind saliency metrics is to quatitatively assess how well a predicted saliency map corresponds to actual human visual attention, as represented by the ground truth eye fixations. Good saliency models should have high values for similarity metrics and low values for dissimilarity metrics.

$$Metric = \mathrm{F}(P, GT) \tag{1}$$

where $_F$ is a function that takes P as a saliency map and GT as a ground truth and predicts the score for the metric under consideration. In this paper, we analyze these five metrics in isolation from the input format and report on the Eye tracking dataset [9]. The only distinction we make in terms of the input that these metrics operate on is whether the ground truth is represented as discrete fixation locations or a continuous fixation map. Accordingly, we categorize metrics as location-based or distribution-based [49]. In this section, we discuss the particular advantages and disadvantages of each metric and present visualizations of the metric computations.

Table 1
Metrics used for evaluating saliency maps.

Metrics	Location-based	Distribution-based
Similarity	Area under ROC curve [44,46,50]	Similarity [47]
	Normalized scanpath saliency [44,48,50–52]	Pearson's correlation coefficient [44,46,48]
Dissimilarity	–	Kullback-leibler divergence [47,51]

3.2. Location-based metrics

3.2.1. Area under ROC curve (AUC)

Given the goal of predicting where viewers will focus their attention on an image, a saliency map can distinguish between pixels that attract gaze and those that do not, effectively acting as a pixel-level classification tool. This concept introduces a metric for assessing the effectiveness of saliency maps. According to signal detection theory, the Receiver Operating Characteristic (ROC) curve is a tool that evaluates the balance between true positives and false positives across different levels of sensitivity [53]. The area under this ROC curve, AUC, is the predominant measure for evaluating saliency maps. Essentially, the saliency map is analyzed as a binary classifier that identifies fixations at different threshold levels, generating a ROC curve by comparing the rate of true positives and false positives for each threshold level. The way true and false positives are computed varies across different AUC methodologies. Alternatively, AUC can be viewed as an indicator of a model's accuracy on a 2AFC task, where the model must choose between two potential points on an image, identifying which one is more likely to be the fixation [54].

3.2.2. Normalized scanpath saliency (NSS)

Normalized Scanpath Saliency is a measure between a saliency map and a set of fixations computed as the average normalized saliency at fixation locations along a subject's scan path [55]. The absolute saliency values are part of the normalization calculations. Given a saliency map P and a binary map of fixation locations Q^B:

$$NSS\left(P, Q^B\right) = \frac{1}{N}\sum_i \overline{P_i} \times Q_i^B \tag{2}$$

where $\overline{P} = \frac{P - \mu(P)}{\sigma(P)}$ *and* $N = \sum_i Q_i^B$

where i indexes the ith pixel and N is the total number of fixated pixels. A positive value indicates correspondence between maps above chance, and a negative NSS indicates anti-correspondence. For instance, a unity score corresponds to fixations falling on portions of the saliency map with a saliency value of one standard deviation above average.

3.3. Distribution-based metrics

3.3.1. Similarity (SIM)

The similarity metric measures the similarity between two distributions, viewed as histograms. It has gained popularity in the saliency community as a simple comparison between pairs of saliency maps. After normalizing the input maps, SIM is computed as the sum of the minimum values at each pixel. For a saliency map P and a continuous fixation map Q^D:

$$SIM\left(P, Q^D\right) = \sum_i min\left(P_i, Q_i^D\right) \tag{3}$$

iterating over discrete pixel locations i. Note that the model with the sparser saliency map has a lower histogram intersection with the ground truth map. SIM is very sensitive to missing values, penalizing predictions that fail to account for all ground truth density. The downside of a distribution metric like SIM is that the choice of the Gaussian sigma (or blur) in constructing the fixation and saliency maps affects model evaluation.

3.3.2. Pearson's correlation coefficient (CC)

Pearson's Correlation Coefficient, also called linear correlation coefficient, is a statistical method used generally in the sciences to measure correlated or dependent two variables. It is one of the most common methods used for numerical variables, and its values are between -1 and $+1$, where -1 means negative correlation, $+1$ means positive correlation, and 0 means no correlation. CC can be used to interpret saliency and fixation maps, P and Q^D, as random variables to measure their linear relationship [56].

$$CC\left(P, Q^D\right) = \frac{\sigma\left(P, Q^D\right)}{\sigma(P) \times \sigma\left(Q^D\right)} \tag{4}$$

where $\sigma(P, Q^D)$ is the covariance of P and Q^D. CC is symmetric and penalizes false positives and negatives equally. Large positive CC values occur at locations where the saliency map and ground truth fixation map have similar magnitudes.

3.3.3. Kullback-leibler divergence (KLD)

Unlike SIM, Kullback-Leibler is a statistical measurement from information theory that measures and quantifies the differences between two probability distributions. Although KL divergence is a popular method, choosing a statistical distance check can sometimes be challenging. In saliency literature, depending on how the saliency predictions and ground truth fixations are interpreted as distributions, different KLD computations are possible. Analogous to our other distribution-based metrics, our KLD metric takes as input a saliency map P and a ground truth fixation map Q^D and evaluates the loss of information when P is used to approximate Q^D:

$$KLD\left(P, Q^D\right) = \sum_i Q_i^D \, log\left(\epsilon + \frac{Q_i^D}{\epsilon + P_i}\right) \tag{5}$$

where ϵ is a regularization constant. KL-Judd is an asymmetric dissimilarity metric, with a lower score indicating a better approximation of the ground truth by the saliency map. The pixels where the ground truth value Q_i^D is non-zero, but P_i is close to or equal to

zero, a large quantity is added to the KLD score, making the regions brighter in the KLD visualization. However, KLD is so sensitive to zero-values that a sparse set of predictions is penalized harshly, significantly worse than chance.

3.4. Qualitative evaluation of saliency

Most saliency papers include side-by-side comparisons of different saliency maps computed for the same images (as in Fig. 3). Visualizations of saliency maps are often used to highlight improvements over previous models. A few anecdotal images might be utilized to showcase model strengths and weaknesses. Bruce et al. [57] discussed the problems with visualizing saliency maps, particularly the strong effect of contrast on the perception of saliency models. We propose supplementing saliency map examples with visualizations of metric computations (as shown in Figs. 5 and 6) to provide an additional means of comparison that is more tightly linked to the underlying model performance than the saliency maps themselves.

4. Evaluation setup

4.1. Data collection

We use an eye tracking dataset for building façade inspection as the ground truth dataset for this paper [58]. This dataset contains eye tracking data of ten participants for two building structures, recorded using Tobii Pro Glasses 3 [59] with a 100 Hz sampling rate, 1920 x 1080 resolution @ 25 fps, and four infrared cameras. Participants were allowed to view and walk around the structure freely to capture dynamic gaze data. The free viewing task is most commonly used for saliency modeling, requiring a few additional assumptions [49]. Pro Glasses 3 allows one-point calibration of gaze patterns and attention filters to compensate for dynamic eye movements due to the nature of the data collection setup (wearable eye tracking).

The gaze data collected will serve as the ground truth fixation map. An important step was to generate discrete and continuous fixation maps for saliency metrics since some of them calculate the similarity score based on fixation location while others calculate

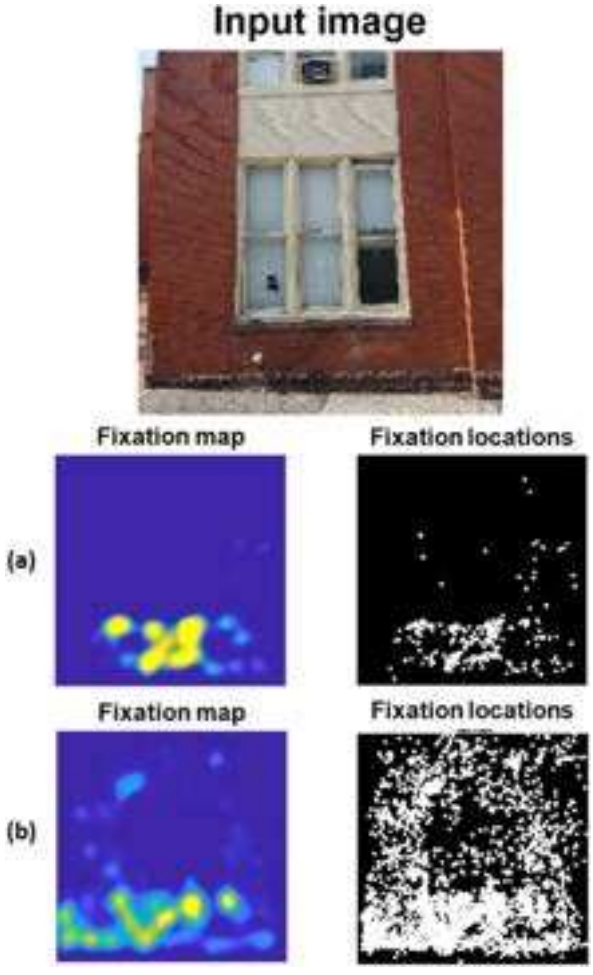

Fig. 2. Saliency metrics are compared on how well they approximate ground truth eye movements, represented as discrete fixation locations or a continuous fixation map for (a) $n_{participant} = 1$ and (b) $n_{participant} = 10$.

based on fixation distribution. Fig. 2 shows an example of a fixation map (distribution-based) and fixation locations (location-based). To generate fixation points, the collected gaze data was first denormalized and converted to image coordinates such that they corresponded to the input image. Finally, the results are compared with the existing saliency maps to compare the differences with the ground truth data and compute the scores based on the fixation location.

4.2. Methods for comparing saliency maps

To visualize saliency evaluation metrics and highlight their differences in metric behaviors, we used different saliency models for which code is available online. The results were reproduced from these models trained on the SALICON dataset [60]. These models include 1) a Contextual encoder-decoder network for visual saliency prediction [61], 2) head pose estimation using CNN and adaptive gradient methods [62], and 3) Saliency in context [43]. These models were chosen due to their simplicity and were most common in generating saliency maps without requiring additional training but fine-tuning the models to generate saliency maps.

4.2.1. Encoder decoder model

Kroner et al. [61] proposed a CNN architecture based on semantic segmentation with modified modules to predict fixation density maps of the input image. Their approach leverages object-specific features to replicate human behavior under free viewing conditions. They adapted from the popular VGG16 architecture [63] as an image encoder by reusing the pre-trained convolutional layers to extract complex features. To restore the original image resolution on the decoder end, extracted features were upsampled and processed through a series of convolution layers. The actual implementation of the EnDec model and training was done on SALICON before fine-tuning the weights towards fixation prediction on either of the datasets MIT1003 [64] or CAT2000 [65] with the same optimization parameters. For our evaluation, the pre-trained model was fine-tuned and tested using SALICON weights. The results and discussion are provided in section 5 in detail.

4.2.2. DeepGaze model

Patacchiola et al. [62] originally proposed a CNN-based model for head pose estimation using adaptive gradient methods. They implement an object detection framework for the face detector and a CNN network for the head pose estimator. The graphical representation of their CNN models has two convolution layers, two subsampling layers, and two fully connected layers. Originally, their work involved training on the AFLW dataset for face detection and was tested using various methods. They had a major challenge regarding increasing pose estimation error for the face detector when it returned a frame that was not well centered on the subject's face.

4.2.3. SALICON model

Jiang et al. [43] proposed a visual attention method by introducing a novel method for collecting extensive human attention data during natural image exploration. Unlike existing datasets focusing on images and task-specific annotations, SALICON emphasizes capturing the dynamics of human attention shifts. Employing a mouse-contingent multi-resolution approach inspired by studies of peripheral vision, it facilitates the simulation of natural viewing behaviors using a standard mouse, thus allowing for the collection of data on an unprecedented scale. This approach has been validated in both laboratory and online settings, resulting in a proof-of-concept dataset from the COCO image dataset. The dataset's potential to improve visual understanding and saliency model training has been demonstrated through its application in saliency prediction, proving to be a valuable ground truth resource for algorithm evaluation. With ongoing data collection efforts, SALICON is poised to significantly contribute to the fields of visual understanding and computer vision by providing a comprehensive resource for studying and modeling human visual attention.

5. Results and findings

We evaluate the performance of these models on existing saliency metrics for assessing building damage using eye tracking data, which is built upon the existing work of Saleem et al. [9]. To this end, we perform 1) fine-tuning and training of saliency models with different pre-trained weights, 2) compare the efficiency of individual models among other saliency models, and 3) evaluate the results for two different building sites.

These experiments aim to quantify the efficacy of conventional saliency models in detecting salient damage features and to

Fig. 3. Saliency maps corresponding to three different models.

understand if future studies are needed to acquire more specific data and models for building industry-specific tasks. Most saliency models include a side-by-side comparison of saliency maps and benchmark the results in tabular form. Visualization of saliency models is often used to highlight improvements over previous models. This work will highlight the saliency map for visualization and provide metrics' quantitative measures to understand the models' performances better. For instance, Fig. 2 shows an input image and how well it approximates human ground truth eye movements. Fig. 2a indicates the gaze information for one participant showing their fixations on the lower part of the building with few outliers on the upper part; Fig. 2b indicates the cumulative, discrete fixation location distributed entirely over the entire façade. However, the fixation map highlights the region of interest in the lower part of the building below the window frame.

The present work will not discuss individual participants' gaze data since the prior work of Saleem et al. [9]. discussed the individual behavior and analysis of gaze patterns comprehensively. This study will focus on how different models generate saliency maps and how different saliency metrics will impact the evaluation overall. A Python package based on TensorFlow was used to compare saliency mapping algorithms, and an out-of-the-box implementation was provided. Fig. 3 compares the three different models, i.e., EnDec, DeepGaze, and SALICON.

This paper seeks to compare the effectiveness of models trained with task-agnostic eye tracking data to ground truth fixation maps based on task-specific eye tracking data. The task-agnostic models were fine-tuned with SALICON pe-trained weights. SALICON is currently the largest public dataset for saliency prediction and contains 10,000 training images and 5000 validation and testing images taken from the COCO dataset. The dataset contains images from the building site with various scenes of different lighting conditions. Although the eye tracking information was collected, the scene-viewing task was unrelated to damage assessment. A comparison of the fine-tuned saliency map on different pre-trained weights, such as the SALICON [60], MIT300 [31], and OSIE [66] datasets, is shown in Fig. 4.

5.1. Location-based metrics analysis

The AUC metric evaluates a saliency map's predictive power by how many ground truth fixations it captures in successive level sets. To compute AUC, the saliency map is treated as a binary classifier of fixation at various threshold values, and an ROC curve is swept out. Thresholding the saliency map produces the level sets in the rightmost column. For each level set, the actual positive rate is the proportion of fixations landing in the level set (green points in the rightmost column). The false positive rate is the proportion of image pixels in the level set not covered in fixations. Five level sets corresponding to points on the ROC curve were included, and the AUC score for the saliency map is the area under the ROC curve.

Judd et al. [41] proposed their AUC variant, called AUC-Borji [14], depicted in Fig. 5. A saliency map visualization and its corresponding AUC curve for evaluating saliency as a fixations classifier is indicated in the top row where *TP* and *FP* indicate true positives and false positives, respectively. In this context, the TP rate, also known as sensitivity or recall, represents the proportion of actual positives the model correctly identifies. The FP rate represents the proportion of actual negatives that are incorrectly identified as positives by the model. The ROC curve is a graphical representation of the trade-off between the TP and FP rates at various threshold settings of a binary classifier.

The AUC score for the EnDec model is 0.54 (45° angle), which means the model cannot identify salient features and perform class separation between true positives and true negatives [45]. For area under the curve metrics, the ROC curve is very important. The closer it is to the top left corner, the higher the test accuracy because the top left corner has the highest sensitivity with a false positive rate of 0. The natural distribution of fixations on an image tends to have higher density near the center, and therefore, models that incorporate center bias into their predictions would achieve a high AUC score. A model predicting the center portion of an image achieves a lower score of 0.5 and would likely perform worse since center bias is present in prediction and the model can not differentiate between true positives and false positives. Similarly, the bottom rows indicate the level sets and their corresponding threshold levels. Here, the green dots represent the true positive fixations, and the red dots represent the false positive fixations on individual levels set to illustrate their behavior. Different AUC implementations differ in calculating a true positive or false positive [46].

For the DeepGaze and SALICON model, the AUC curve is shown in Fig. 6, along with the TP and FP rate, where the AUC score for DeepGaze is 0.45 and SALICON is 0.62. Similarly, the curve for DeepGaze falls below 0.5, meaning that DeepGaze has failed and negatively correlates with the fixation map. On the other hand, SALICON models has a score of 0.62, with salient features highlighted

Fig. 4. Saliency maps generated based on (a) SALICON, (b) MIT1003, and (c) OSIE pretrained weights.

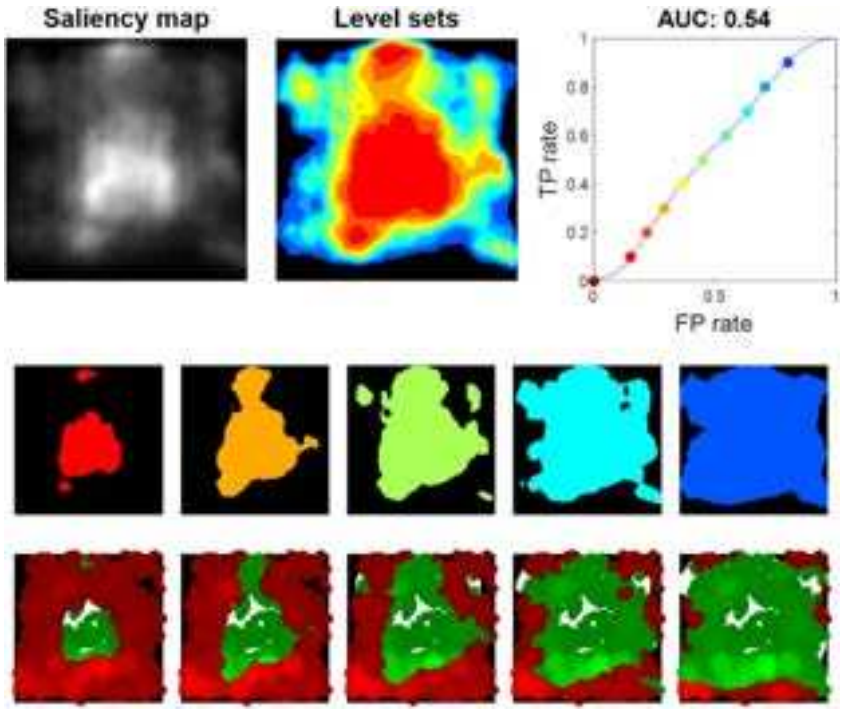

Fig. 5. AUC visualization of EnDec model with different level set.

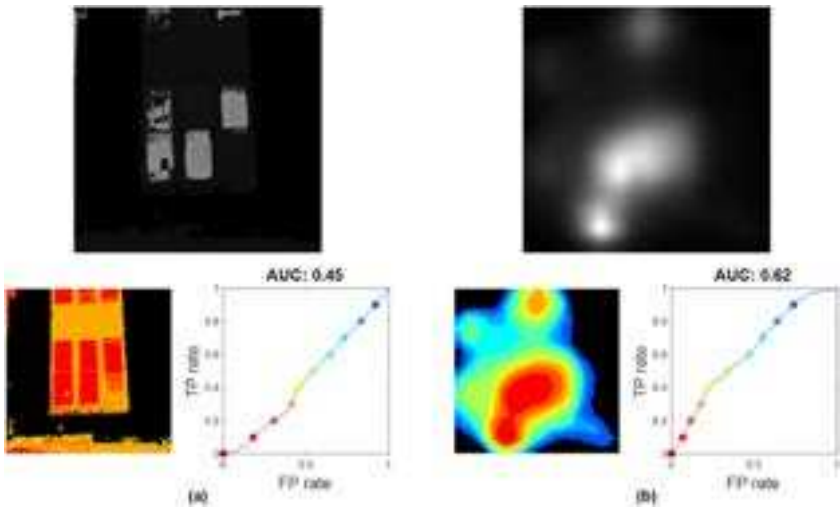

Fig. 6. Comparison of AUC curves for (a) DeepGaze and (b) SALICON model.

close to the fixation map. It is interesting to note that although the SALICON model is trained based on the MIT300 dataset with gaze data, the task-free nature of data is inefficient for our case study. Therefore, the current saliency models perform poorly for our building inspection tasks even though the score is closer to $+1$.

NSS measures similarity based on fixation location between a saliency map and fixation map (human ground truth). Fig. 7 visually illustrates the NSS technique comparing EnDec, DeepGaze, and SALICON models. Like AUC, the higher the NSS score, the higher the similarity, and vice versa. NSS usually normalizes a saliency map by the standard deviation of the saliency values. $NSS_{En_{D}ec}$ has center bias with more similarity in the center according to the threshold scale, but in addition to this, the fixations are quite dispersed overall. $NSS_{DeepGaze}$ has highly similar fixation points, but there is no such relation with the fixation map since the model highlights the salient features on the window frame of the actual image. $NSS_{SALICON}$ shows a good similarity between the saliency map and ground truth fixations, highlighting the damaged regions on the test structure. Also, $NSS_{SALICON}$ has an average positive score of 0.3477 compared to the other two models with negative scores. Table 2 shows comprehensive quantitative results and compares all the metrics under

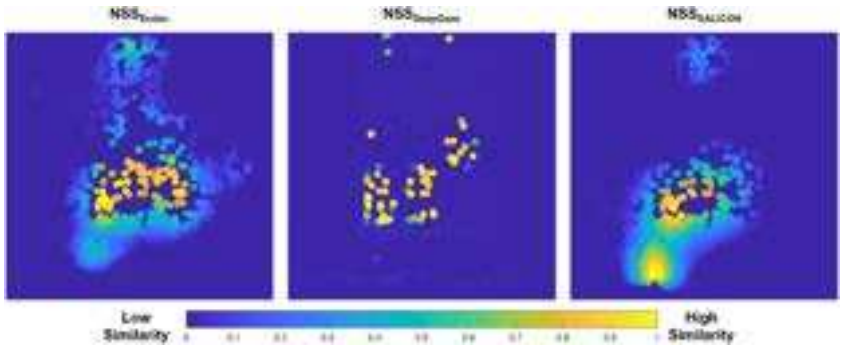

Fig. 7. Normalized Scanpath Saliency map ($n_{participants} = 10$).

Table 2
Performance score of location-based saliency metrics for different models.

Sample Space	Saliency Models	Building-1		Building-2	
		NSS ↑	AUC-Borji ↑	NSS ↑	AUC-Borji ↑
Individual participant	EnDec	−0.0679	0.4419	**0.9606**	**0.6577**
	DeepGaze	−0.2991	0.4748	0.6295	0.5715
	SALICON	**0.1646**	**0.4970**	0.1075	0.4954
All participants	EnDec	−0.0088	0.4753	**0.8036**	**0.6284**
	DeepGaze	−0.2328	0.4791	0.4601	0.5475
	SALICON	**0.1810**	**0.5009**	0.1862	0.4999

consideration.

Table 2 compares saliency models for two building sites, and the sample size is changed from individual participants to all participants. Individual participants indicate average scores for the location-based metrics compared to the total sample size. It should be

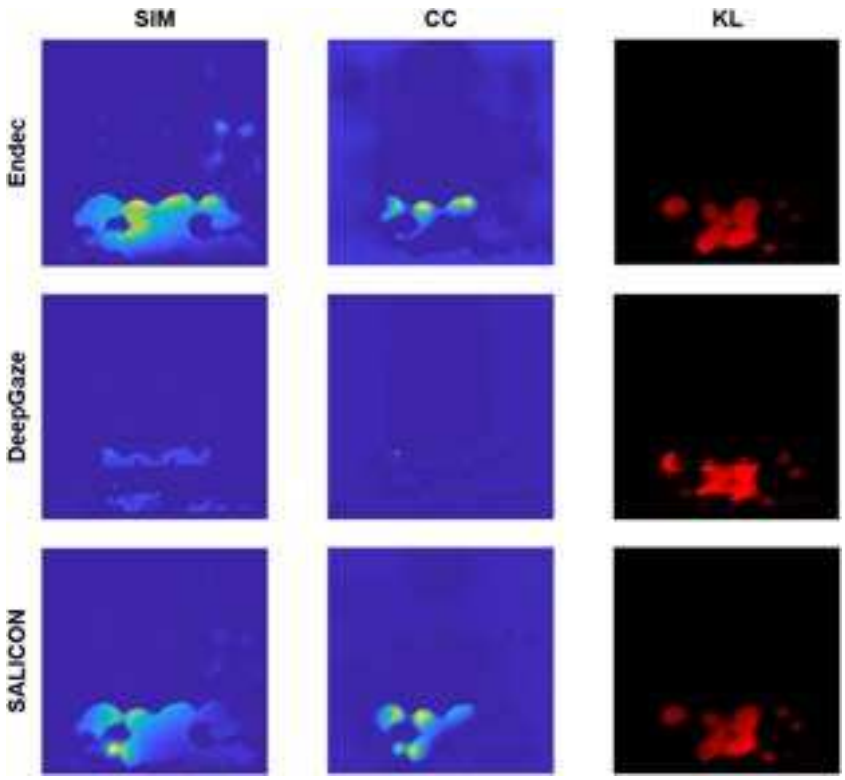

Fig. 8. Comparison of distribution-based metrics.

noted that for location-based metrics, the score must be high, which corresponds to higher correlation and similarity between two entities. From our results, we noticed that AUC-Dorji has an average score of 0.5, while the NSS score is lower for building 1 and higher for building 2. A score of greater than 0.8 means higher similarity and correlation; lower than that means the results are insignificant and there is no strong similarity. From our results, our argument validates that conventional saliency maps cannot be used to train a model for building inspection since there is no similarity among salient features generated by available models for damage assessment. Further, the results clearly show the SALICON model's performance over the other two. Although the metric score is higher for SALICON based compared to the EnDec and DeepGaze model, the score is close to 0 for NSS, while AUC-Borji has 0.5, which states it is not a strong correlation between the saliency map and the fixation map.

5.2. Distribution-based analysis

Location-based metrics described so far score saliency models on how accurately they predict discrete fixation locations. Suppose the ground truth fixation locations are interpreted as a possible sample from some underlying probability distribution. In that case, another approach is to predict the underlying distribution directly instead of the fixation locations. Although it is challenging to predict ground truth distribution, Gaussian blurring the fixation locations into a fixation map often approximates it. This section describes distribution-based metrics that score saliency models on how accurately they approximate the continuous fixation map. Taking the comparison further, we evaluated and compared the differences for metrics such as SIM, CC, and KL divergence.

SIM has gained popularity in the saliency community as a simple comparison between pairs of saliency maps and computed the sum of pixel minimums between the predicted saliency map and the ground truth human fixations. A similarity score 1 indicates that the predicted map is identical to the ground truth data. The CC metric measures the linear coefficient between the saliency and fixations maps, with a score between -1 and $+1$. CC treats false positives and negatives symmetrically, but SIM places less emphasis on false positives than false negatives. As a result, all three saliency maps have low SIM and CC scores, resulting in a negative correlation between the ground truth fixation map and respective saliency models. Unlike SIM and CC, KLD measures the dissimilarity between the saliency map and the ground truth data and is much more sensitive to false negatives than SIM or CC. Fig. 8 visually illustrates the comparison among the three metrics (SIM, CC, and KLD) and how they compute similarity and dissimilarity among different models. SIM and CC metrics measure the similarity between the saliency map and the ground truth fixation map. SIM measures the histogram intersection between two maps, while CC measures cross-correlation.

Fig. 8 shows the behavior of SIM and CC and how they are affected by false negatives and false positives. SIM penalizes false negatives significantly more than false positives, but CC treats both symmetrically. Due to its symmetric computation, CC can not distinguish whether the differences between maps are due to false positives or false negatives. The corresponding score also highlights that CC performs better and scores better than SIM, which is consistently higher for the SALICON model. The results in Table 3 suggest that although the SALICON model has higher correlation and similarity compared to the other two models and performs better than expected for location-based metrics, the score is still lower or close to 0 which validates our argument that the saliency map generated using these models cannot be relied upon for fixation. There is no similarity of saliency maps with the actual fixation maps.

6. Conclusion and future work

This paper provides a comparative analysis of saliency mapping algorithms and the effectiveness of models trained with task-agnostic eye tracking data to ground truth fixation maps in capturing visual priorities for building inspection. We compared the performance of three different models, i.e., EnDec, DeepGaze, and SALICON, with the ground truth data and provided a visual representation of the most relevant regions within an image. By tracking the gaze patterns of experienced human inspectors as they assess structural damage, we can establish a ground truth dataset that reflects the regions of interest deemed significant by human experts. This ground truth dataset can then be utilized to evaluate the correspondence between the areas highlighted by saliency maps and those attended to by human inspectors. Despite looking at many saliency metrics, we compare the performance only for five common methods: the area under the curve, similarity, scan path saliency, correlation coefficient, and kullback-leibler. The results suggest that although the SALICON model has higher correlation and similarity compared to the other two models and performs better than expected for location-based metrics, the score is still lower or close to 0 which validates our argument that the saliency map generated using these models cannot be relied upon for fixation. There is no similarity of saliency maps with the actual fixation maps. This proves our argument that conventional saliency models are not suitable for our application in identifying damage and generating a saliency map.

This result underscores the imperative for developing advanced, domain-specific saliency models tailored to meet the unique requirements of building inspection tasks. Furthermore, the findings suggest the creation of specialized datasets that more accurately reflect the complexity of structural assessments, thereby enhancing the fidelity of saliency models in practical applications. The implications of this study extend beyond the structural inspection and civil engineering domain, offering valuable insights into the potential for human-machine collaboration in the broader field of automated disaster response and infrastructure maintenance. By bridging the gap between computational predictions and expert human judgment, we can significantly advance our capabilities in early damage detection, risk assessment, and the prioritization of repair efforts, ultimately contributing to the resilience and safety of the built environment.

CRediT authorship contribution statement

Muhammad Rakeh Saleem: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Rebecca Napolitano:** Writing – review & editing, Supervision, Resources, Project

Table 3
Performance score of distribution-based saliency metrics for different models.

Sample Space	Saliency Models	Building-1			Building-2		
		SIM ↑	CC ↑	KLD ↓	SIM ↑	CC ↑	KLD ↓
Individual participant	EnDec	2.10e-04	−0.0010	8.6842	7.19e-04	0.0021	8.5134
	DeepGaze	7.53e-05	−0.0045	23.2541	**9.47e-04**	0.0125	15.7875
	SALICON	**2.84e-04**	**0.0025**	**8.5805**	4.67e-04	**0.0191**	**7.4024**
All participants	EnDec	0.0029	−4.73e-04	**6.1088**	0.0038	0.0089	7.4163
	DeepGaze	0.0014	−0.0126	19.5211	**0.0046**	0.0220	14.6189
	SALICON	**0.0038**	**0.0098**	6.6659	0.0030	**0.0384**	**5.7741**

administration, Funding acquisition.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this manuscript, the authors used ChatGPT in the introduction and conclusion part to improve the readability and grammar of the manuscript. After using these tools, the authors reviewed and edited the content as needed and takes full responsibility for the content of the publication.

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.

Data availability

Data will be made available on request.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. BCS 2121909 and IIS 2123343. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors would also like to thank all the participants who obtained the data for this study and provided insights based on their experience and knowledge.

References

[1] "Preliminary damage assessments | FEMA.gov.". https://www.fema.gov/disaster/how-declared/preliminary-damage-assessments#report-guide. (Accessed 12 February 2024).
[2] "ATC-20.". https://www.atcouncil.org/atc-20. (Accessed 12 February 2024).
[3] C.-M. Chang, T.-K. Lin, F. Moreu, D.K. Singh, V. Hoskere, "Post disaster damage assessment using ultra-high-resolution aerial imagery with semi-supervised transformers,", Sensors 2023 23 (19) (Oct. 2023) 8235, https://doi.org/10.3390/S23198235, 23, Page 8235.
[4] S. Varghese, V. Hoskere, Unpaired image-to-image translation of structural damage, Adv. Eng. Inf. 56 (Apr. 2023) 101940, https://doi.org/10.1016/J.AEI.2023.101940.
[5] K. Malek, A. Mohammadkhorasani, F. Moreu, Methodology to integrate augmented reality and pattern recognition for crack detection, Comput. Aided Civ. Infrastruct. Eng. 38 (8) (May 2023) 1000–1019, https://doi.org/10.1111/MICE.12932.
[6] H. Kim, Y. Narazaki, B.F. Spencer, Automated bridge component recognition using close-range images from unmanned aerial vehicles, Eng. Struct. 274 (Jan. 2023) 115184, https://doi.org/10.1016/J.ENGSTRUCT.2022.115184.
[7] R. Nasimi, F. Moreu, M. Nasimi, R. Wood, Developing enhanced unmanned aerial vehicle sensing system for practical bridge inspections using field experiments, Transp Res Rec 2676 (6) (Jun. 2022) 514–522, https://doi.org/10.1177/03611981221075618/ASSET/IMAGES/LARGE/10.1177_03611981221075618-FIG7.JPEG.
[8] Y. Zhao, B. Lu, M. Alipour, UAS-Based automated structural inspection path planning via visual data analytics and optimization [Online]. Available: https://arxiv.org/abs/2312.15109v1, Dec. 2023. (Accessed 14 February 2024).
[9] M.R. Saleem, R. Mayne, R. Napolitano, Analysis of gaze patterns during facade inspection to understand inspector sense-making processes, Sci. Rep. 13 (1) (Feb. 2023) 1–11, https://doi.org/10.1038/s41598-023-29950-w, 2023 13:1.
[10] K. Máthé, L. Buşoniu, "Vision and control for UAVs: a survey of general methods and of inexpensive platforms for infrastructure inspection,", Sensors 2015 15 (7) (Jun. 2015) 14887–14916, https://doi.org/10.3390/S150714887, 15, Pages 14887-14916.
[11] Y. Narazaki, V. Hoskere, G. Chowdhary, B.F. Spencer, Vision-based navigation planning for autonomous post-earthquake inspection of reinforced concrete railway viaducts using unmanned aerial vehicles, Autom Constr 137 (May 2022) 104214, https://doi.org/10.1016/J.AUTCON.2022.104214.
[12] N. Bolourian, A. Hammad, LiDAR-equipped UAV path planning considering potential locations of defects for bridge inspection, Autom Constr 117 (Sep. 2020) 103250, https://doi.org/10.1016/J.AUTCON.2020.103250.
[13] K. Mirzaei, M. Arashpour, E. Asadi, H. Feng, S.R. Mohandes, M. Bazli, Automatic compliance inspection and monitoring of building structural members using multi-temporal point clouds, J. Build. Eng. 72 (Aug. 2023) 106570, https://doi.org/10.1016/J.JOBE.2023.106570.
[14] M.R. Saleem, J.W. Park, J.H. Lee, H.J. Jung, M.Z. Sarwar, Instant bridge visual inspection using an unmanned aerial vehicle by image capturing and geo-tagging system and deep convolutional neural network, Struct. Health Monit. 20 (4) (Jul. 2021) 1760–1777, https://doi.org/10.1177/1475921720932384/ASSET/IMAGES/LARGE/10.1177_1475921720932384-FIG16.JPEG.
[15] M. Alipour, D.K. Harris, G.R. Miller, Robust pixel-level crack detection using deep fully convolutional neural networks, J. Comput. Civ. Eng. 33 (6) (Nov. 2019) 04019040, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000854/ASSET/8AD9FE1B-32F0-4B01-8D48-C82CFFBFEF9E/ASSETS/IMAGES/LARGE/FIGURE18.JPG.

[16] D.J. Atha, M.R. Jahanshahi, Evaluation of deep learning approaches based on convolutional neural networks for corrosion detection, Struct. Health Monit. 17 (5) (Sep. 2018) 1110–1128, https://doi.org/10.1177/1475921717737051/ASSET/IMAGES/LARGE/10.1177_1475921717737051-FIG8.JPEG.
[17] Y.J. Cha, W. Choi, O. Büyüköztürk, Deep learning-based crack damage detection using convolutional neural networks, Comput. Aided Civ. Infrastruct. Eng. 32 (5) (May 2017) 361–378, https://doi.org/10.1111/MICE.12263.
[18] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: an overview of interpretability of machine learning. International Conference on Data Science and Advanced Analytics, Jul. 2018, pp. 80–89, https://doi.org/10.1109/DSAA.2018.00018.
[19] S. Chakraborty, et al., "Interpretability of Deep Learning Models: A Survey of Results," 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (*SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Jun. 2017, pp. 1–6, https://doi.org/10.1109/UIC-ATC.2017.8397411.
[20] Z.C. Lipton, The mythos of model interpretability, Queue 16 (3) (Jun. 2018) 31–57, https://doi.org/10.1145/3236386.3241340.
[21] Y. Gao, K.M. Mosalam, Deep learning visual interpretation of structural damage images, J. Build. Eng. 60 (Nov. 2022) 105144, https://doi.org/10.1016/J.JOBE.2022.105144.
[22] M.A.A.K. Jalwana, N. Akhtar, M. Bennamoun, A. Mian, CAMERAS: enhanced resolution and sanity preserving class activation mapping for image saliency. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Jun. 2021, pp. 16322–16331, https://doi.org/10.1109/CVPR46437.2021.01606.
[23] K. Chen, G. Reichard, X. Xu, A. Akanmu, Automated crack segmentation in close-range building façade inspection images using deep learning techniques, J. Build. Eng. 43 (Nov. 2021) 102913, https://doi.org/10.1016/J.JOBE.2021.102913.
[24] A. Silva, J. de Brito, Do we need a buildings' inspection, diagnosis and service life prediction software? J. Build. Eng. 22 (Mar. 2019) 335–348, https://doi.org/10.1016/J.JOBE.2018.12.019.
[25] M. Fu, R. Liu, Q. Liu, How individuals sense environments during indoor emergency wayfinding: an eye-tracking investigation, J. Build. Eng. 79 (Nov. 2023) 107854, https://doi.org/10.1016/J.JOBE.2023.107854.
[26] M. Choi, S. Kim, S. Kim, Semi-automated visualization method for visual inspection of buildings on BIM using 3D point cloud, J. Build. Eng. 81 (Jan. 2024) 108017, https://doi.org/10.1016/J.JOBE.2023.108017.
[27] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks arXiv:1311.2901v3 [cs.CV] 28 nov 2013, Computer Vision–ECCV 2014 8689 (PART 1) (2014) 818–833, https://doi.org/10.1007/978-3-319-10590-1_53.
[28] K. Szczepankiewicz, et al., Ground truth based comparison of saliency maps algorithms, Sci. Rep. 13 (1) (Dec. 2023), https://doi.org/10.1038/S41598-023-42946-W.
[29] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. 128 (2) (Feb. 2020) 336–359, https://doi.org/10.1007/S11263-019-01228-7/FIGURES/21.
[30] D.P. Papadopoulos, A.D.F. Clarke, F. Keller, V. Ferrari, Training object class detectors from eye tracking data, Lect. Notes Comput. Sci. 8693 (PART 5) (2014) 361–376, https://doi.org/10.1007/978-3-319-10602-1_24/COVER. LNCS.
[31] MIT/Tuebingen saliency benchmark. https://saliency.tuebingen.ai/. (Accessed 6 February 2024).
[32] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, Adv. Neural Inf. Process. Syst. (Oct. 2018) 9505–9515, 2018-December, https://arxiv.org/abs/1810.03292v3. (Accessed 7 February 2024).
[33] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259, https://doi.org/10.1109/34.730558.
[34] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, Hum. Neurobiol. 4 (4) (1985) 219–227, https://doi.org/10.1007/978-94-009-3833-5_5/COVER.
[35] J. Tilke, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look. IEEE International Conference on Computer Vision, 2009, pp. 2106–2113, https://doi.org/10.1109/ICCV.2009.5459462.
[36] M. Kümmerer, L. Theis, M. Bethge, Deep gaze I: boosting saliency prediction with feature maps trained on ImageNet. International Conference on Learning Representations, 2014.
[37] M. Cornia, L. Baraldi, G. Serra, R. Cucchiara, "A deep multi-level network for saliency prediction,", in: Proceedings - International Conference on Pattern Recognition vol. 0, Jan. 2016, pp. 3488–3493, https://doi.org/10.1109/ICPR.2016.7900174.
[38] M. Kummerer, T.S.A. Wallis, L.A. Gatys, M. Bethge, Understanding low- and high-level contributions to fixation prediction, Proceedings of the IEEE International Conference on Computer Vision (Dec. 2017) 4799–4808, https://doi.org/10.1109/ICCV.2017.513, 2017-October.
[39] E. Vig, M. Dorr, D. Cox, Large-scale optimization of hierarchical features for saliency prediction in natural images. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Sep. 2014, pp. 2798–2805, https://doi.org/10.1109/CVPR.2014.358.
[40] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012) [Online]. Available: http://code.google.com/p/cuda-convnet/. (Accessed 7 February 2024).
[41] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 3rd international conference on learning representations, ICLR 2015 - conference track proceedings. https://arxiv.org/abs/1409.1556v6, Sep. 2014. (Accessed 7 February 2024).
[42] N. Liu, J. Han, D. Zhang, S. Wen, T. Liu, Predicting eye fixations using convolutional neural networks, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Oct. 2015, pp. 362–370, https://doi.org/10.1109/CVPR.2015.7298633, 07-12-June-2015.
[43] M. Jiang, S. Huang, J. Duan, Q. Zhao, "SALICON: saliency in context,", in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Oct. 2015, pp. 1072–1080, https://doi.org/10.1109/CVPR.2015.7298710, 07-12-June-2015.
[44] A. Borji, H.R. Tavakoli, D.N. Sihite, L. Itti, Analysis of scores, datasets, and models in visual saliency prediction, Proceedings of the IEEE International Conference on Computer Vision (2013) 921–928, https://doi.org/10.1109/ICCV.2013.118.
[45] O. Le Meur, T. Baccino, Methods for comparing scanpaths and saliency maps: strengths and weaknesses, Behav. Res. Methods 45 (1) (Jul. 2013) 251–266, https://doi.org/10.3758/S13428-012-0226-9/TABLES/2.
[46] U. Engelke, et al., Comparative study of fixation density maps, IEEE Trans. Image Process. 22 (3) (2013) 1121–1133, https://doi.org/10.1109/TIP.2012.2227767.
[47] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, T. Dutoit, Saliency and human fixations: state-of-the-art and study of comparison metrics, Proceedings of the IEEE International Conference on Computer Vision (2013) 1153–1160, https://doi.org/10.1109/ICCV.2013.147.
[48] A. Borji, D.N. Sihite, L. Itti, Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study, IEEE Trans. Image Process. 22 (1) (2013) 55–69, https://doi.org/10.1109/TIP.2012.2210727.
[49] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, F. Durand, What do different evaluation metrics tell us about saliency models? IEEE Trans. Pattern Anal. Mach. Intell. 41 (3) (Mar. 2019) 740–757, https://doi.org/10.1109/TPAMI.2018.2815601.
[50] O. Le Meur, T. Baccino, Methods for comparing scanpaths and saliency maps: strengths and weaknesses, Behav. Res. Methods 45 (1) (Jul. 2012) 1–16, https://doi.org/10.3758/S13428-012-0226-9.
[51] N. Wilming, T. Betz, T.C. Kietzmann, P. König, Measures and limits of models of fixation selection, PLoS One 6 (9) (Sep. 2011) e24038, https://doi.org/10.1371/JOURNAL.PONE.0024038.
[52] Q. Zhao, C. Koch, Learning a saliency map using fixated locations in natural scenes, J. Vis. 11 (3) (Mar. 2011) 9, https://doi.org/10.1167/11.3.9, 9.
[53] M. Green, J.A. Swets, S. Detection Theory, J.A. Nevin, Signal detection theory and operant behavior: a review of david M. Green and john A. Swets' signal detection theory and Psychophysics.1, J. Exp. Anal. Behav. 12 (3) (May 1969) 475–480, https://doi.org/10.1901/JEAB.1969.12-475.
[54] M. Kümmerer, T.S.A. Wallis, M. Bethge, Information-theoretic model comparison unifies saliency metrics, Proc Natl Acad Sci U S A 112 (52) (Dec. 2015) 16054–16059, https://doi.org/10.1073/PNAS.1510393112/SUPPL_FILE/PNAS.201510393SI.PDF.
[55] R.J. Peters, A. Iyer, L. Itti, C. Koch, Components of bottom-up gaze allocation in natural images, Vision Res 45 (18) (Aug. 2005) 2397–2416, https://doi.org/10.1016/J.VISRES.2005.03.019.

[56] O. Le Meur, P. Le Callet, D. Barba, Predicting visual fixations on video based on low-level visual features, Vision Res 47 (19) (Sep. 2007) 2483–2498, https://doi.org/10.1016/J.VISRES.2007.06.015.
[57] N.D.B. Bruce, C. Wloka, N. Frosst, S. Rahman, J.K. Tsotsos, On computational modeling of visual saliency: examining what's right, and what's left, Vision Res 116 (Nov. 2015) 95–112, https://doi.org/10.1016/J.VISRES.2015.01.010.
[58] "An Eye Tracking Dataset for Building Façade Inspection", doi: 10.5281/ZENODO.7125956.
[59] "Tobii Pro Glasses 3 | Latest in Wearable Eye Tracking - Tobii." Accessed: Feb. 06, 2024. [Online]. Available: https://www.tobii.com/products/eye-trackers/wearables/tobii-pro-glasses-3?creative=639361858008&keyword=tobii%20glasses&matchtype=p&network=g&device=c&utm_source=google&utm_medium=cpc&utm_campaign=&utm_term=tobii%20glasses&gad_source=1&gclid=CjwKCAiA8YyuBhBSEiwA5R3-EzROe5QhfR2VQIGigK4-XaAZphXYIimpL0a9XY7idQGVxwxeV8K8oBoC3NYQAvD_BwE.
[60] Salicon. http://salicon.net/. (Accessed 6 February 2024).
[61] A. Kroner, M. Senden, K. Driessens, R. Goebel, Contextual encoder–decoder network for visual saliency prediction, Neural Network. 129 (Sep. 2020) 261–270, https://doi.org/10.1016/J.NEUNET.2020.05.004.
[62] M. Patacchiola, A. Cangelosi, M. Patacchiola, A. Cangelosi, Head pose estimation in the wild using Convolutional Neural Networks and adaptive gradient methods, PatRe 71 (Nov. 2017) 132–143, https://doi.org/10.1016/J.PATCOG.2017.06.009.
[63] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations, 2014.
[64] T. Judd, F. Durand, A. Torralba, A benchmark of computational models of saliency to predict human fixations [Online]. Available: https://dspace.mit.edu/handle/1721.1/68590, Jan. 2012. (Accessed 6 February 2024).
[65] A. Borji, L. Itti, "CAT2000: a large scale fixation dataset for boosting saliency research,". https://arxiv.org/abs/1505.03581v1, May 2015. (Accessed 6 February 2024).
[66] J. Xu, M. Jiang, S. Wang, M.S. Kankanhalli, Q. Zhao, "Predicting human gaze beyond pixels,", J. Vis. 14 (1) (Jan. 2014) 28, https://doi.org/10.1167/14.1.28, 28.