Title Page


**TITLE: Understanding Human Visual Attention: A Statistical Comparison of Deep Learning-Based Saliency Models Across Diverse Scene Complexities.**

**Author:** David Erika Topos

**Degree:** Bsc. Computer Science

**Supervisor:** Hantao Liu

**Institution:** School of Computer Science and Informatics, Cardiff University

**Date:** 08/ May/ 2025

## Abstract

Human visual attention is a complex process shaped by both low-level features such as colour, contrast, edges, and texture, and high-level features including scene context and object semantics. Understanding the mechanisms underlying gaze behaviour is essential for advancing fields like computer vision, pattern recognition, and human-computer interaction. Although recent developments in computational saliency models particularly those incorporating Convolutional Neural Networks (CNNs) have significantly improved fixation prediction, challenges remain in building models that accurately replicate human visual behaviour across diverse viewing conditions.

This study presents a comprehensive statistical analysis of human attention data and saliency maps predicted by three state-of-the-art models with different input modalities and design philosophies: **DeepGaze IIE** (CNN-based, probabilistic fixation prediction model), Saliency Unification through Mamba (**SUM)** (domain-adaptive Mamba-based model trained with both eye-tracking and mouse tracking data), and **SALICON** (a large-scale model trained using mouse-contingent behavioural data). The core aim is to evaluate how these models perform across a range of stimuli with **varying scene complexities**, and to understand how model architecture and input modality influence saliency prediction.

By analysing model performance across low and high-complexity scenes and using evaluation metrics such as **Normalized Scanpath Saliency (NSS)**, **Area Under the Curve (AUC-Judd)**, and **Correlation Coefficient (CC)**, this study identifies not only the **strengths and weaknesses of each model** but also highlights the influence of **scene complexity and input modality** on saliency accuracy. The results help determine which models are best suited for particular image domains and reveal how architectural design choices and training paradigms affect the ability to replicate human gaze patterns. Furthermore, this analysis offers **recommendations for improving computational saliency models**, including suggestions for better generalization, multimodal integration, and adaptive prediction strategies. These findings aim to contribute to the development of future models that more closely mimic human attention mechanisms across real-world contexts.

## Acknowledgement

I would like to extend my heartfelt thanks to all those who have supported and encouraged me throughout the course of this project. I would also like to acknowledge the teaching and administrative staff at Cardiff University, School of Computer Science & Informatics, for providing the academic foundation, tools, and resources that enabled me to undertake and complete this research project.

Importantly, I would also like to take a moment to acknowledge myself for my perseverance, discipline, and resilience in the face of challenges. Completing this project has required sustained effort, critical thinking, and a willingness to learn from setbacks, and I am proud of the growth I have achieved through this experience.

Finally, I am grateful for access to open-source resources, pre-trained models, and publicly available datasets, which made it possible to conduct meaningful analysis within the constraints of time and resources. This report also benefited from grammar and spelling corrections provided by openaI's large language model.

## Table of Contents

Contents

## 1. Introduction

Human visual attention is a fundamental cognitive process that enables individuals to selectively focus on the most relevant elements within a visual scene while filtering out less important information (Hofheimer & Lester, 2008). This mechanism is influenced by both low-level visual features such as colour, contrast, and texture and high-level factors including object semantics and contextual understanding. The effort to understand and replicate this attentional behaviour has given rise to the field of visual saliency research. Visual saliency refers to the perceptual quality that makes certain elements of a scene stand out and automatically attract human attention, often in a bottom-up, pre-attentive manner (Itti & Koch, 2001). Saliency modelling has become increasingly relevant in domains such as computer vision, image processing, and human-computer interaction, where mimicking human-like attention mechanisms can enhance system performance and user experience.

Visual saliency is commonly understood through two primary components, bottom-up and top-down attention. Bottom-up saliency is stimulus-driven, relying on low-level visual features such as colour, contrast, brightness, and orientation to capture attention in an automatic and involuntary manner (Dhara & Kumar, 2023). For instance, a bright red object on a green background instinctively draws attention due to its strong colour contrast. In contrast, top-down saliency is guided by cognitive factors such as goals, prior knowledge, and task relevance (Ramanishka, 2016). An example is the intentional search for a friend in a crowd, where attention is directed based on memory and purpose rather than purely on visual distinctiveness.

The biological basis of visual attention has inspired the development of computational models that aim to predict which regions of an image or video are most likely to attract human gaze. This area of research, known as saliency modelling, seeks to replicate human attentional patterns through algorithmic approaches. Figure 1 illustrates an example of salient object detection, highlighting how certain regions of an image naturally attract greater visual focus.

**Figure 1: An example of salient object(s) in image(s) and their corresponding ground truth.**

These models have practical applications in fields such as advertising, medical imaging, and mobile interface design. For example, Leiva et al. (2020) highlight the value of gaze prediction in mobile applications, where it allows designers to evaluate and optimise layouts without extensive user testing. Saliency data can also be used to intentionally guide user attention, improving usability by strategically positioning key elements where users are most likely to focus.

A variety of saliency models have been developed to emulate human gaze behaviour. Traditional models, such as the Itti and Koch model (1998) and Graph-Based Visual Saliency (GBVS), use bottom-up approaches that rely primarily on low-level visual features to generate saliency maps. While effective for basic stimuli, these models often lack the capacity to incorporate semantic understanding or contextual relevance. In contrast, modern computational saliency models leverage deep learning techniques and are trained on large-scale eye-tracking datasets. Notable examples include the Saliency Attentive Model (SAM), SALICON, and DeepGaze IIE. These models integrate both bottom-up and top-down cues, resulting in more accurate and context-aware predictions. Zhao et al. (2021) found that deep learning-based models significantly outperform traditional methods, especially when dealing with distorted or complex images. This improvement is largely due to their ability to learn hierarchical visual features and adapt to diverse visual environments. Figure 2 demonstrates a comparative analysis between traditional and deep learning models on distorted images, underscoring the robustness of the latter.

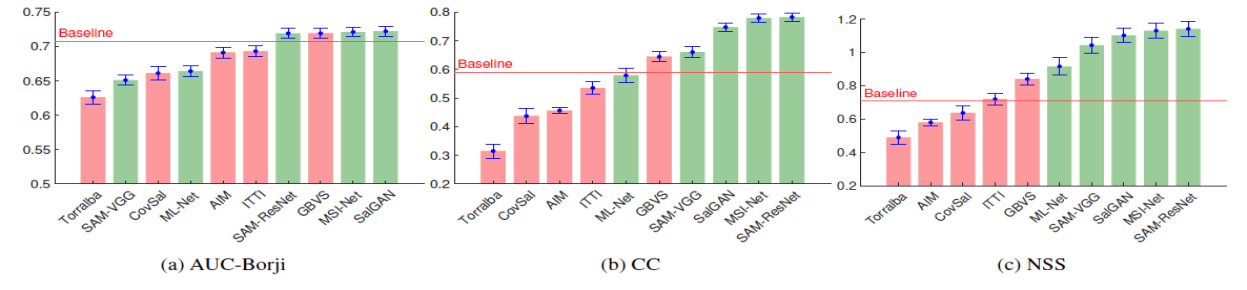



**Figure 2: A comparative analysis showing that deep learning (green bars) models outperform traditional (pink bars) models on distorted images: Deep Learning VS. Traditional Algorithms for Saliency Prediction of Distorted Images (Zhao et al., 2021).**

This project investigates how computational saliency models with varying input modalities and architectural designs perform across images with differing levels of scene complexity. The focus is on evaluating the extent to which these models can accurately replicate human gaze behaviour under diverse visual conditions. By examining their responses to complex and varied image content, the study aims to

uncover how well current models generalise beyond simple or uniform stimuli. Three deep learning-based models are assessed in this research: **SALICON**, which is trained on mouse-tracking data; Saliency Unification through Mamba (**SUM)**, which incorporates both eye-tracking and mouse-tracking data; and **DeepGaze II E**, trained exclusively on eye-tracking data. Each model is evaluated using standard saliency prediction metrics, including Area Under the Curve (AUC-Judd), Normalized Scanpath Saliency (NSS), and Pearson Correlation Coefficient (CC), to assess their predictive accuracy. Through this analysis, the study seeks to identify the merits and limitations of different input modalities and architectural approaches. Ultimately, the findings are intended to inform future research directions aimed at improving the accuracy and robustness of saliency models in replicating human visual attention across complex visual scenes.

## 1.1. Aims and Objectives

### 1.1.1. Aim

The primary aim of this project is to evaluate the performance of state-of-the-art deep learning-based saliency models with different input modalities and architectural designs in predicting human visual attention across images with varying scene complexities.

### 1.1.2. Objectives

To achieve this aim, I outline the following key objectives:

i. Develop an understanding of image complexity and investigate how variations in visual scene complexity influence saliency detection.
ii. Conduct a comparative evaluation of three state-of-the-art saliency models SALICON, SUM (Saliency Unification through Mamba), and DeepGaze IIE on image datasets with varying levels of visual complexity, using established evaluation metrics: AUC-Judd, NSS, and CC.
iii. Explore the impact of different input modalities including eye-tracking, mouse-tracking, and hybrid datasets on the performance and accuracy of saliency prediction models
iv. Examine the strengths and limitations of computational saliency models trained on varying input modalities such as eye-tracking, mouse-tracking, and hybrid datasets in predicting accurate saliency maps.
v. Derive actionable insights and propose recommendations to guide the development of next-generation saliency prediction models, informed by the results of the comparative evaluation.

## 1.2. Motivation for the project

In an increasingly visual digital world, understanding where people look and why is critical for designing effective, user-centred visual systems. From mobile applications and web interfaces to autonomous vehicles and medical imaging, the ability to accurately predict human gaze behaviour has a wide range of impactful applications. Visual saliency modelling seeks to replicate this attentional behaviour computationally, enabling machines to prioritise visual information in a way that aligns with human perception.

While traditional saliency models have laid the foundation for this field, their reliance on low-level features limits their ability to generalise across real-world visual environments, which often involve semantic understanding and contextual reasoning. Recent advances in deep learning have introduced more powerful models capable of incorporating both bottom-up and top-down cues. However, despite their success, key questions remain about how different input modalities (e.g., eye-tracking vs. mouse-tracking vs hybrid) and architectural choices affect performance especially when models are exposed to images of varying complexity.

This project is motivated by the need to better understand these underlying factors and evaluate how current state-of-the-art models perform under diverse conditions. By comparing models trained on different data types and built with distinct architectural approaches, this study aims to uncover their relative strengths and weaknesses. The goal is not only to improve the accuracy of saliency prediction but also to provide deeper insights that can guide the development of future models, ultimately bringing computational attention systems closer to human-like perception.

## 2. Background

Visual saliency modelling has progressed significantly over the past few decades, evolving from biologically inspired models to complex deep learning-based approaches. Early models, such as the one proposed by Itti, Koch, and Niebur (1998), were rooted in neurobiological theories of visual attention and primarily relied on bottom-up mechanisms. These models generated saliency maps by combining basic visual features like intensity, colour, and orientation.

In contrast, more recent models integrate both bottom-up and top-down cues, leveraging task relevance, contextual information, and learned representations through deep neural networks (Cornia et al., 2018; Kümmerer et al., 2016). These models benefit from large, annotated datasets and deep architectures capable of learning hierarchical feature representations. Alongside these architectural developments, the input modalities used for training saliency models have also evolved. While traditional models relied on precise eye-tracking data, recent models have increasingly adopted mouse-tracking and hybrid approaches to collect broader and more scalable datasets (Huang et al., 2015).

Existing research underscores the importance of both architectural design and input modality in determining a model's ability to accurately predict human gaze. Convolutional neural networks (CNNs) have been particularly effective at capturing complex visual patterns that correlate with human attention, especially when trained on high-resolution eye-tracking data (Bylinskii et al., 2019). However, studies also show that models trained on mouse-tracking data may not fully capture subtle aspects of visual attention, such as peripheral vision or rapid saccades (Chen et al., 2021).

The effectiveness of saliency models is thus tightly linked to the alignment between model architecture and the fidelity of the input modality. Hybrid datasets that combine eye and mouse-tracking data have emerged as a compromise, aiming to balance data availability with predictive accuracy.

Saliency models differ not only in their training data but also in how their architectures are structured to process visual information. Models trained on **eye-tracking data**, such as DeepGaze II E, are considered to offer high precision in predicting human fixation points (Kümmerer et al., 2016). In contrast, **mouse-tracking-based models** like SALICON leverage large-scale datasets collected through crowdsourcing, offering scalability at the expense of some spatial accuracy (Huang et al., 2015). Meanwhile, **hybrid models** like SUM incorporate both modalities to fuse the strengths of each (Islam et al., 2020).

Architecturally, some models employ attention mechanisms to refine spatial predictions, while others rely on feature extraction backbones such as VGG or ResNet. These design choices significantly impact how well a model handles complex scenes, including cluttered environments and multiple salient objects.

This project builds on the insights from previous research by examining the interplay between model architecture and input modality in deep learning-based saliency models. It focuses on how these factors influence performance across images with varying scene complexities. Three state-of-the-art models **SALICON**, **SUM**, and **DeepGaze II E** were selected for analysis. These models were trained on different modalities: mouse-tracking, a hybrid of mouse and eye-tracking, and eye-tracking data respectively.

The models are evaluated using established metrics, including **Area Under the Curve (AUC)**, **Normalized Scanpath Saliency (NSS)**, and **Correlation Coefficient (CC)**. Through comparative analysis, the study aims to uncover how training data and architectural differences affect saliency prediction and to provide recommendations for future improvements in the field.

The rest of this report is structured as follows: Section 3 details the methodology, including descriptions of the dataset, the saliency models evaluated, and the process for assessing scene complexity. Section 4 presents the results and evaluation, highlighting overall model performance, the impact of scene complexity, and comparative visual analyses. Section 5 discusses the conclusions drawn from the findings and suggests directions for future work. Section 6 reflects on the learning outcomes from undertaking this project. Supporting materials, such as figures, charts, and extended results, are included in the appendices, followed by the references used throughout the report.

# 3. Methodology

## 3.1. Data Description

The **MIT1003** dataset is a publicly available benchmark widely used in visual attention research (Judd et al., 2009). It consists of 1,003 natural images accompanied by eye-tracking data collected from 15 participants who viewed the images under free-viewing conditions. This dataset provides valuable insights into human visual fixation patterns and supports saliency modelling research.

For this project, I used a sample of images from the dataset featuring diverse stimuli, including indoor and outdoor environments, natural scenes, urban landscapes, objects, and people. This variety allows for a better evaluation of how selected computational models perform across different types of scenes, thereby helping to understand how scene complexity influences visual saliency.

## 3.2. Saliency Models Evaluated

Three deep learning-based models were evaluated in this project including DeepgazeIIE, SUM, and SALICON.

### 3.2.1. DeepgazeIIE

In this project I evaluated DeepgazeIIE, a state-of-the-art deep learning-based saliency model designed to predict human visual attention by generating probabilistic fixation density maps. DeepGaze IIE extends the original DeepGaze II framework by integrating multiple fixed CNN backbones pretrained on ImageNet including DenseNet201, ResNext50, ShapeNet-C, and EfficientNet-B5 to extract rich visual features. These features are passed through a readout network comprising 1×1 convolutions, layer normalization, and softplus activation, which transforms the features into spatial saliency representations. A centre-bias prior is then added before applying a softmax function to produce the final fixation probability map (Linardos et al., 2021). The model is pretrained on the SALICON dataset and fine-tuned on the MIT1003 dataset using maximum likelihood estimation. Evaluated via 10-fold cross-validation and various saliency metrics such as Information Gain (IG), AUC, NSS, and KL divergence, DeepGaze IIE provides a robust benchmark for predicting human fixations (Linardos et al., 2021). In this project, DeepGaze IIE was used to generate saliency maps for diverse visual stimuli, which were compared against human eye-tracking data. Its ability to generalize across different image types, combined with high interpretability and performance, makes it highly suitable for this project's objective. Figure 3 below depicts the DeepGaze II and DeepGaze IIE models. Part (a) outlines the process of testing various CNN backbones, while part (b) shows how DeepGaze IIE combines multiple backbones.

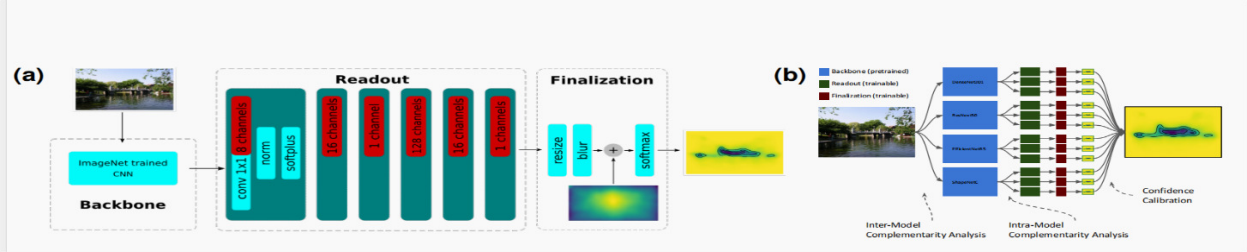



**Figure 3: DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modelling (Linardos et al., 2021, P. 4).**

### 3.2.1. 1. Saliency Map Generation Pipeline Using DeepGaze IIE

To apply the DeepGaze IIE model for saliency prediction, a dedicated inference pipeline was implemented using the open source **deepgaze_pytorch** package, publicly available via GitHub. The goal was to generate saliency maps for a custom dataset of natural images and evaluate their alignment with human eye-tracking data.

#### i. Software Dependencies

The pipeline utilized the following Python libraries for model inference, data manipulation, and visualization:

- **torch**: for loading the pretrained model and executing forward passes.
- **scipy**: for image interpolation and log-domain normalization.
- **numpy**: for managing image arrays and tensor transformations.
- **skimage.io**: for reading and handling input image files.
- **matplotlib. pyplot**: for generating and saving heatmap visualizations.
- **tqdm**: for displaying real-time progress during batch processing.

#### ii. Data Preprocessing

All input images were stored in a directory named **Stimuli**. Each image was loaded using skimage.io. imread () and standardized as follows:

- Grayscale images were converted to RGB by duplicating the single channel across the three-color channels.
- Images with an alpha (transparency) channel were reduced from RGBA to RGB by discarding the alpha layer.
- A predefined center-bias prior, originally defined for a resolution of 1024×1024 pixels, was incorporated to simulate natural gaze centrality. This prior was resized using **scipy. ndimage. zoom** to match the resolution of each input image and normalized in the log domain using **scipy. special. logsumexp**.

### iii. Inference and Saliency Map Generation

Each processed image and its corresponding center-bias map were passed into the model to obtain a log-probability saliency map:

**log_density** = model (image_tensor, centerbias_tensor)

The output log_density represents the unnormalized log-probabilities of gaze distribution over image pixels. This output was subsequently:

- Converted into a NumPy array for downstream quantitative analysis.
- Saved in. npy format for archival and numerical evaluation.
- Visualized as a heatmap using a 'hot' colormap and saved as a .png image for qualitative inspection.

### iv. Automation and Scalability

To streamline saliency generation across the dataset, the entire process was automated using a Python script. This script iterated over each file in the **Stimuli** directory, processing and visualizing saliency maps in a reproducible manner. A simplified structure of the loop is shown below:

for filename in tqdm (os. listdir (INPUT_FOLDER)):

**image, image_tensor** = load_image_as_tensor(img_path)

**centerbias** = zoom (centerbias_template, (height / 1024, width / 1024), order=0, mode='nearest')

centerbias -= logsumexp(centerbias)

**centerbias_tensor** = torch. from NumPy(centerbias). unsqueeze (0)

with torch.no_grad ():

log_density = model (image_tensor, centerbias_tensor)

np. save (output_path, log_density. Numpy ())

plt. imshow (np.exp(log_density), cmap='hot')

plt. axis ('off')

```
plt. savefig (visual_path)
```

This automated pipeline allowed for the efficient and consistent generation of saliency maps across all test images. The resulting visual and numerical outputs were then used for comparative analysis against ground truth eye-tracking data, enabling a systematic evaluation of the DeepGaze IIE model's predictive alignment with human visual behaviour.

#### 3.2.2. Saliency Unification through Mamba (SUM)

Saliency Unification through Mamba (SUM) is a recent state-of-the-art model designed for universal saliency prediction across diverse image types. It combines the efficient long-range dependency modelling capabilities of the Mamba architecture with the spatial precision of the U-Net framework, further enhanced by a novel Conditional Visual State Space (C-VSS) block. This block enables the model to dynamically adapt to various visual contexts such as natural scenes, e-commerce images, and user interfaces by modulating internal feature representations based on the input type (Zhou et al., 2024). Additionally, SUM incorporates a class-conditioned Multi-Layer Perceptron (MLP) module that applies adaptive scaling and shifting to fine-tune predictions according to domain-specific characteristics.

Trained and validated on six benchmark datasets encompassing a wide range of content types and acquisition modalities (including both eye-tracking and mouse-tracking), SUM achieves state-of-the-art performance across both location-based metrics (e.g., NSS, AUC) and distribution-based metrics (e.g., CC, SIM, KL divergence) (Zhou et al., 2024).

In this project, I employed SUM to generate saliency maps for a diverse set of visual stimuli and evaluated its predictions against ground-truth human fixation data. Its strong ability to generalize across modalities and image domains, combined with its computational efficiency and close alignment with human attention patterns, made it a highly suitable and impactful choice for my research objectives.

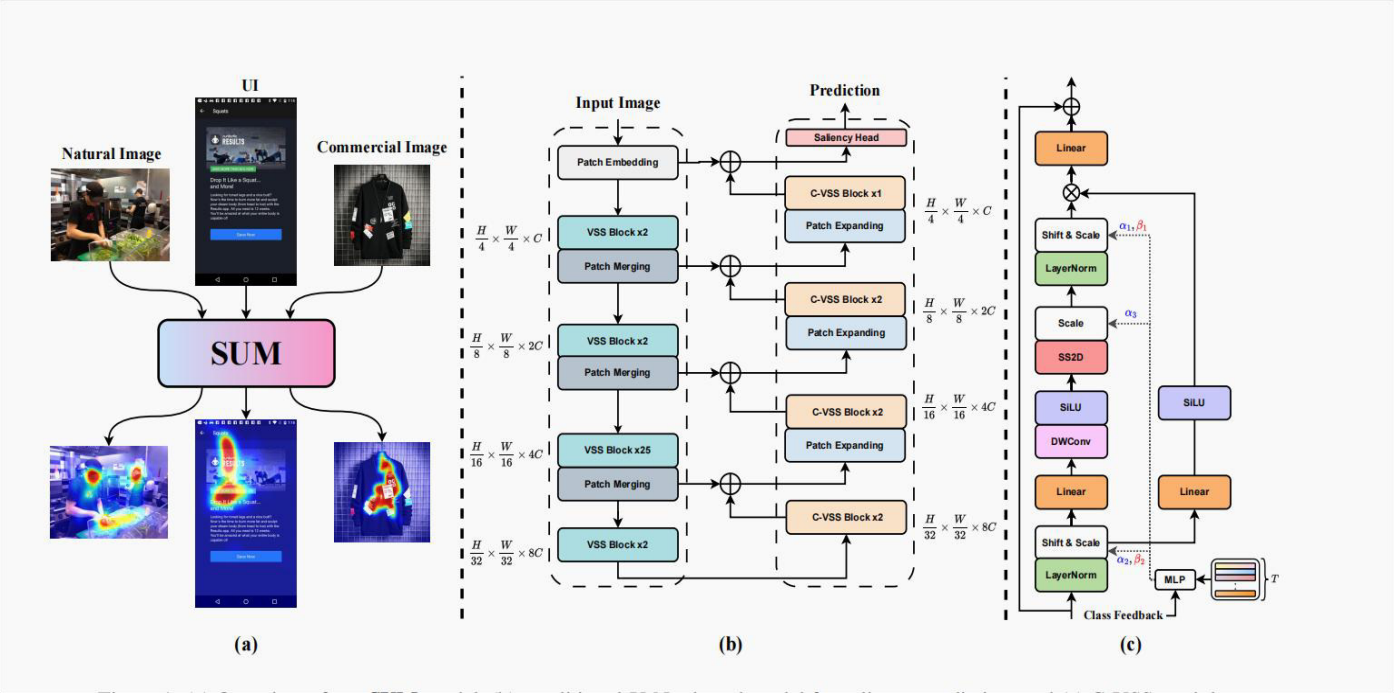



**Figure 4: (a) Overview of our SUM model, (b) conditional-U-Net-based model for saliency prediction, and (c) C-VSS module (Hosseini et al., 2024, P.4).**

#### 3.2.2.1. Saliency Map Generation Using SUM

To apply the SUM model to a custom dataset, a structured inference pipeline was established using the official PyTorch implementation available via the project's GitHub repository. The process involved acquiring the model, configuring the runtime environment, and executing a standardized inference script.

**Model Acquisition and Environment Setup**

The following steps were followed to integrate SUM into the experimental workflow:

1. **Download the Pre-trained Weights**
   The pretrained model weights (sum_model.pth) were retrieved from a Google Drive link provided in the official GitHub repository.

2. **Directory Placement**
   The downloaded weights were placed in the appropriate directory for model loading:

net/pre_trained_weights/sum_model.pth

3. **Runtime Environment**
   Inference was conducted using a GPU-enabled environment within **Google Collaboratory**, which facilitated efficient processing of high-resolution imagery and expedited computation.

**Inference Parameters and Execution**

The SUM model supports multi-domain saliency estimation through a script named **inference.py**, which accepts configurable parameters to tailor predictions based on content type and output format. The basic usage structure is as follows:

```
python inference.py \
    --img_path /path/to/image.jpg \
    --condition [0, 1, 2, 3] \
    --output_path /path/to/output \
    --heat_map_type [HOT, Overlay]
```

Key command-line arguments include:

- **--img_path**: Specifies the path to the input image.
- **--condition**: An integer flag to indicate the input domain:
    - 0: Natural scenes (mouse-tracking data).
    - 1: Natural scenes (eye-tracking data).
    - 2: E-commerce product images.
    - 3: User interface (UI) design layouts.
- **--output_path**: Target directory for saving saliency outputs.
- **--heat_map_type**: Determines the format of the output:
    - HOT: Standalone saliency heatmap.
    - Overlay: Heatmap superimposed on the original image.

**Saliency Map Generation Examples**

Several sample commands were executed to evaluate the model's response to different image categories. For example:

- **Standalone heatmap for natural scenes (eye-tracking):**

```
python inference.py \
    --img_path input_image.jpg \
    --condition 1 \
    --output_path output_results \
    --heat_map_type HOT
```

- **Overlay heatmap for e-commerce imagery:**

```
python inference.py \
    --img_path input_image.jpg \
    --condition 2 \
    --output_path output_results \
    --heat_map_type Overlay
```

In both cases, the model loads the input image, processes it under the selected domain condition, and produces a saliency map either as a standalone heatmap or an overlay for contextual interpretation. These outputs were subsequently stored for both visual analysis and metric-based comparison against human fixation ground truth data.

#### 3.2.3. Salicon

Another saliency model evaluated in this study is SALICON (Saliency in Context), which serves both as a large-scale dataset and a computational framework for predicting human visual attention. Unlike traditional eye-tracking-based methods, SALICON employs a novel mouse-contingent multi-resolutional paradigm inspired by neurophysiological studies of peripheral vision. This setup approximates gaze behaviour using standard computer mice, enabling scalable and cost-effective collection of attentional data during free-viewing tasks (Jiang et al., 2015). The paradigm simulates foveal vision by presenting high-resolution image patches centred around the mouse cursor, with peripheral regions rendered at lower resolutions.

SALICON uses aggregated mouse movements from multiple observers to generate saliency maps, which strongly correlate with eye-tracking data in both qualitative patterns and quantitative metrics like shuffled AUC (sAUC). In this project, I used the SALICON model to predict saliency maps across a variety of stimuli and compared them to ground-truth human fixation data. Its scalability, ecological validity, and ability to approximate human visual exploration make it a valuable for my research.

##### 3.2.3.1. Saliency Map Generation Using Salicon

Given that OpenSALICON relies on the legacy **Caffe** deep learning framework and its Python interface (pycaffe), the model required a specialized computational environment. All experiments were conducted using a **GPU-enabled Google Collaboratory** runtime to ensure efficient execution.

The following configuration steps were taken:

1. **Caffe Installation**
   A custom build of Caffe was compiled with Python layer support, which is required for executing custom inference routines:

```
WITH_PYTHON_LAYER: = 1
```

2. **Python Package Dependencies**
   Core Python libraries (numpy, PIL, matplotlib, scipy) were installed to support image loading, preprocessing, and visualization.

3. **Path Integration**
   The Caffe Python path was manually added to the system path in the model script:

sys. path. Insert (0, 'caffe/install/python') Adjusted to actual installation path

4. **GPU Configuration**
   The model was configured for GPU inference to accelerate computation:

caffe.set_mode_gpu ()

caffe.set_device (0)

5. **Model Files**
   pre-trained model weights (. caffemodel) and corresponding architecture files (. prototxt) were obtained from the official repository and placed in the working directory. No modifications to file paths were required when using default settings.

#### 3.2.5.3 Saliency Map Generation Process

Once the environment was configured, saliency maps were generated using the Salicon class provided in the OpenSALICON codebase. This class encapsulates both image preprocessing and model inference. The general process involved the following steps:

1. Convert input images to **RGB format** and resize to **256×256 pixels**, matching model input requirements.
2. Use the compute_saliency () function to generate a saliency map from the processed image:

from Salicon import Salicon

sal = Salicon ()

map = sal. Compute_saliency('face.jpg')

The function returns a **floating-point saliency map**, which can be visualized using standard Python libraries. The model does not include built-in preprocessing; therefore, input images were manually prepared to ensure compatibility.

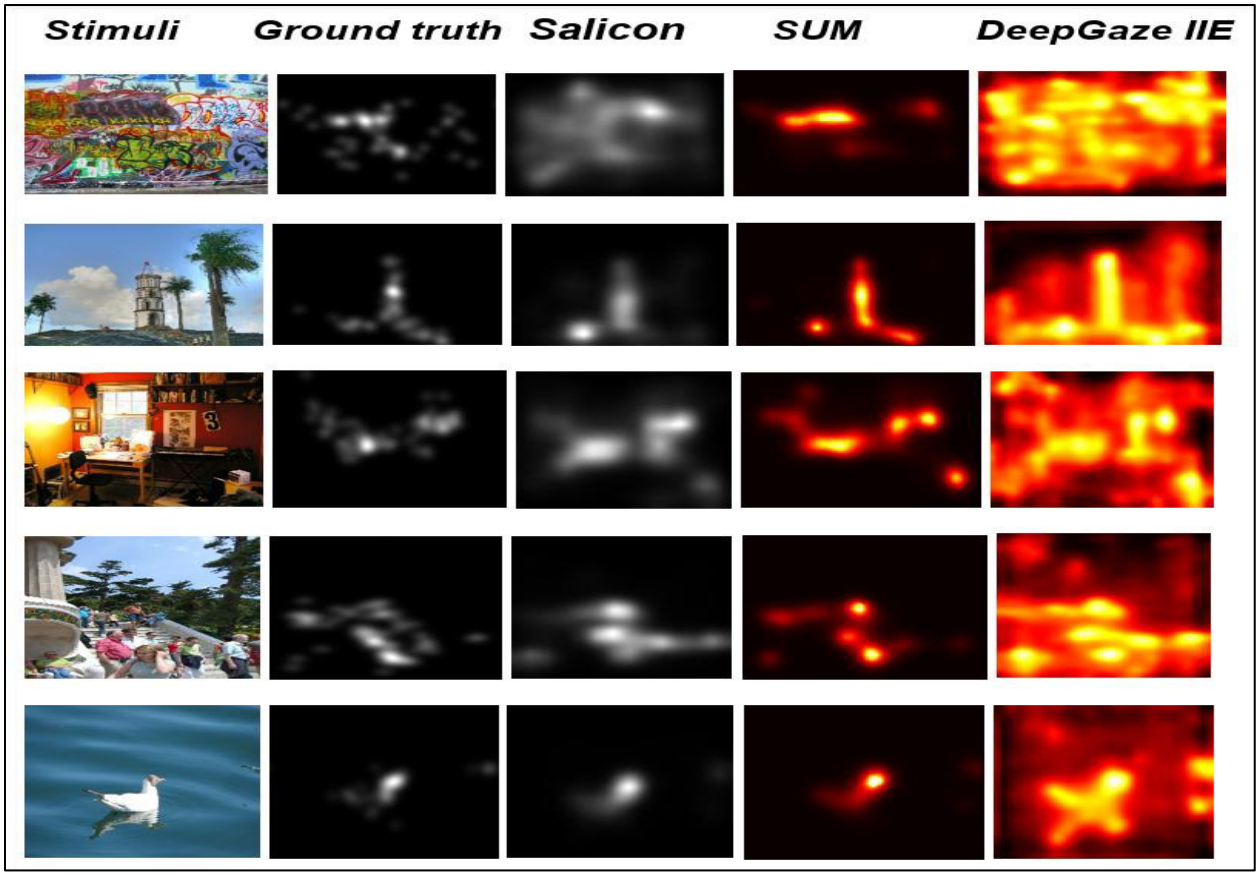



**Figure 5: A sample of image(s) with their corresponding ground truth saliency maps and predicted saliency maps generated by the computational models.**

### 3.3. Image Complexity

#### 3.3.1. Understanding Scene Complexity

Scene complexity refers to the degree of visual intricacy present within an image, encompassing elements such as the number of distinct objects, variations in colour, texture, and structural arrangement. In the context of visual saliency research, scene complexity plays a critical role in influencing human gaze behaviour, as more visually dense and heterogeneous scenes often present greater competition for attentional resources (Borji & Itti, 2013). Scenes with high complexity typically contain numerous edges, varied textures, and rich colour information, whereas low-complexity scenes are characterized by uniform regions and minimal visual variation.

Quantifying scene complexity is essential for evaluating the generalizability of saliency models beyond simple or synthetic stimuli toward more naturalistic, cluttered environments. High-complexity images pose a greater challenge for computational models, as the presence of numerous salient cues can dilute or obscure the primary focus of attention (Zhao et al., 2021). Therefore, incorporating a reliable measure of scene complexity is necessary to systematically assess model robustness across diverse visual conditions.

### 3.3.2. Computing Scene Complexity Scores

To objectively quantify the complexity of images used in this study, a two-pronged approach was implemented, leveraging **edge density** and **colour variation** as core indicators of visual richness.

#### 3.3.2.1. Image Preprocessing

Each image was loaded using OpenCV, a widely adopted computer vision library. Images that failed to load due to format inconsistencies or file errors were excluded to maintain data integrity and ensure consistent computational analysis.

#### 3.3.2.2. Edge Density Calculation

The grayscale version of each image was generated to simplify intensity-based operations. Subsequently, the Canny edge detector (Canny, 1986) was applied to identify regions of rapid intensity change, corresponding to edges and object boundaries.

Edge Density was defined as the ratio of the number of detected edge pixels to the total number of pixels in the image:

$$\textbf{Edge Density} = \frac{\textbf{Number of Edge Pixels}}{\textbf{Total Number of Pixels}}$$

A higher edge density indicates a structurally more complex scene, suggesting the presence of numerous objects, contours, and fine details.

#### 3.3.2.3. Colour Variation Measurement

In parallel, the colour diversity within each image was assessed by calculating the standard deviation of pixel intensities across all three-color channels (RGB). Before computation, pixel values were normalized to the [0,1] range to ensure consistency across images of different resolutions and bit depths.

The colour standard deviation serves as a proxy for chromatic richness: images with diverse and highly varied colours yield higher standard deviation values, whereas scenes dominated by uniform colours exhibit lower variability.

#### 3.3.2.4. Scene Complexity Score Computation

The final Scene Complexity Score was derived by averaging the normalized edge density and colour variation values:

$$\textbf{Scene Complexity score} = \frac{\textbf{Edge Density + Colour Standard Deviation}}{\textbf{2}}$$

This balanced formulation ensures that both structural and chromatic features contribute equally to the complexity estimation, aligning with previous research emphasizing the multifaceted nature of visual scene understanding (Bruce & Tsotsos, 2005).

#### 3.3.2.5. Visualization and Analysis

To facilitate interpretation, the computed complexity scores were tabulated, and images were visually presented in descending order of complexity. Each image was displayed alongside its corresponding complexity score, enabling qualitative cross-validation of the computed metrics.

For clarity and manageability, a maximum of 9 images were selected for visualization. However, the computational framework developed is scalable to accommodate larger datasets without modification.

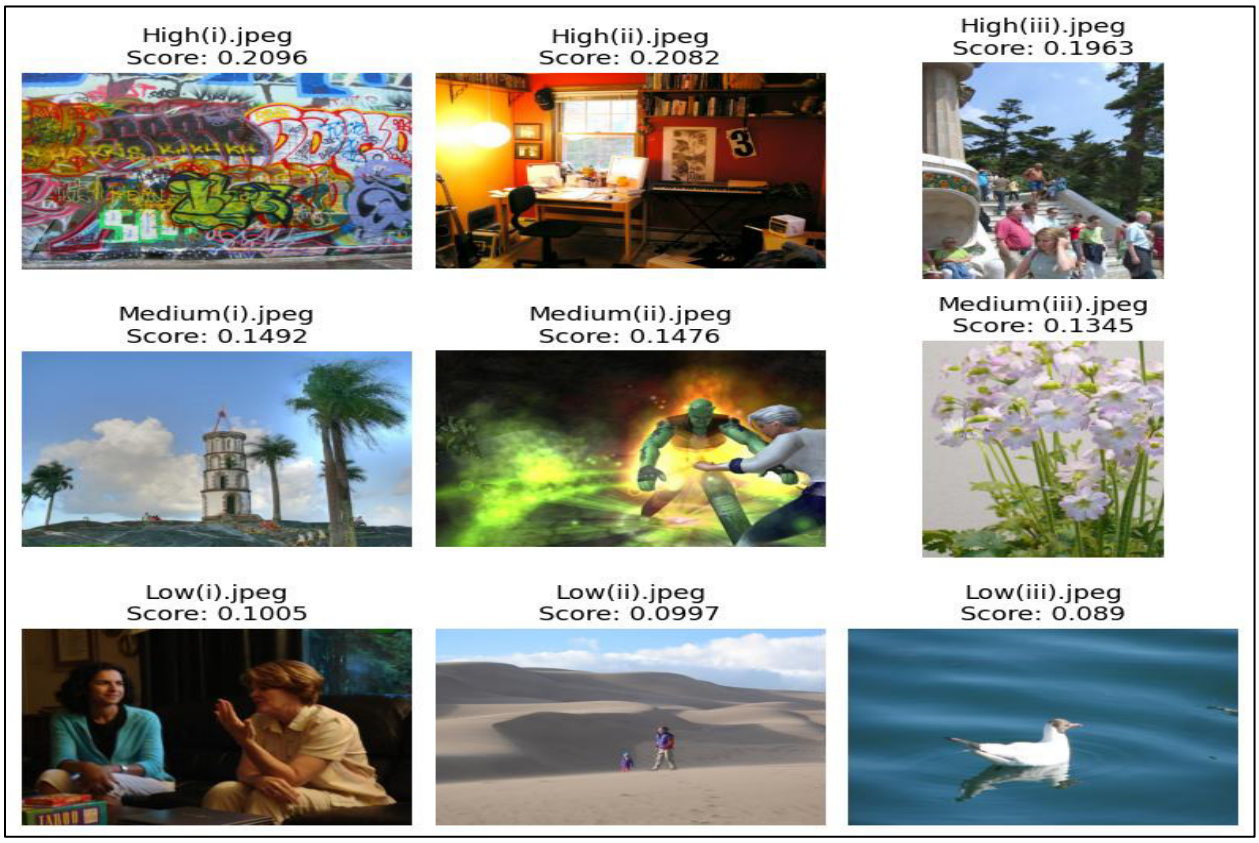



**Figure 6: Images with their complexity score arranged in the descending order of their complexity.**

#### 3.3.2.6. Rationale for Combining Edge Density and Colour Variation

The decision to integrate both edge density and colour variation into the scene complexity score stems from the understanding that visual complexity is inherently multidimensional. Edge density captures the structural intricacy of a scene by quantifying the number of object boundaries, textures, and fine spatial details, which are critical for bottom-up saliency mechanisms (Itti & Koch, 2001). However, relying solely on structural cues may overlook important variations in chromatic information that also contribute significantly to attentional deployment (Bruce & Tsotsos, 2005). Conversely,

colour variation alone may fail to account for scenes that are structurally rich but chromatically uniform, such as black-and-white photographs or low-saturation images.

By averaging these two complementary measures, the resulting complexity score offers a more holistic and balanced representation of visual richness, ensuring that both structural and chromatic dimensions of the scene are equally considered. This combined approach aligns with the broader understanding of human visual perception, where both form and colour interact to influence attentional prioritization.

Eventually, with the scene complexity scores computed for the selected image set, the next phase of the study involves evaluating the performance of three state-of-the-art deep learning-based saliency models SALICON, SUM, and DeepGaze II E across varying levels of visual complexity. By systematically comparing predicted saliency maps against human gaze data using established evaluation metrics, this analysis aims to assess the extent to which each model accurately replicates human attentional patterns under both simple and complex visual conditions.

The following sections present the results of this evaluation and discuss the observed trends in model performance relative to scene complexity.

# 4. Results and Evaluation

## 4.1. Evaluation Metrics used

To quantitatively assess the performance of the saliency prediction models **SALICON**, **SUM**, and **DeepGaze II E**, this study employed three widely used evaluation metrics: **AUC-Judd**, **Pearson Correlation Coefficient (CC)**, and **Normalized Scanpath Saliency (NSS)**. These metrics are standard in saliency research and offer complementary insights into how well the predicted saliency maps align with ground truth human eye fixation data. All metrics were implemented in Python using the scikit-learn, NumPy, and SciPy libraries.

Prior to evaluation, predicted saliency maps were resized and aligned with the resolution of the ground truth fixation maps. Binary fixation maps and continuous fixation density maps were derived from the eye-tracking dataset used in this study. Each metric was computed per image and then averaged across the test set for model comparison.

### 4.1.1. AUC – Judd

The **Area Under the Receiver Operating Characteristic Curve – Judd variant (AUC-Judd)** evaluates the discriminative power of a saliency map in distinguishing between fixated and non-fixated pixels (Judd et al., 2009). This metric treats the predicted saliency map as a probabilistic classifier: human fixations from the ground truth binary map are considered positive samples, while all other locations are negatives.

AUC-Judd is computed by sweeping a threshold across the predicted saliency map and calculating the true positive rate (TPR) and false positive rate (FPR) at each threshold,

thereby forming the ROC curve. The area under this curve represents the final score. A perfect saliency map yields an AUC of 1.0, while a score of 0.5 corresponds to random guessing. Notably, AUC-Judd is robust to false positives and does not require prior normalization or thresholding of the saliency map.

In this study, AUC-Judd was calculated using sklearn.metrics.roc_curve to generate the ROC curve and sklearn.metrics.auc to compute the area under the curve.

#### 4.1.2. Pearson Correlation Coefficient (CC)

The **Pearson Correlation Coefficient (CC)** measures the linear relationship between the predicted saliency map and the ground truth continuous fixation density map (Bylinskii et al., 2019). It is defined as:

$$\mathrm{CC}(S, G) = \frac{\mathrm{cov}(S, G)}{\sigma_S \sigma_G}$$

where S is the predicted saliency map, G is the ground truth fixation density map, cov (S, G) is the covariance between the two, and σS, σG are their standard deviations. CC values range from -1 to 1, with 1 indicating perfect positive linear correlation and 0 indicating no correlation.

For this evaluation, the saliency maps were flattened and standardized before calculating CC using numpy. corrcoef. This metric captures how well the spatial distributions of predicted and actual fixations agree, making it suitable for assessing spatial coherence in saliency predictions.

#### 4.1.3. Normalized Scanpath Saliency (NSS)

The **Normalized Scanpath Saliency (NSS)** metric measures how well the predicted saliency values correspond to actual fixation locations (Peters et al., 2005). Unlike CC, which evaluates global similarity, NSS focuses on local accuracy at human fixation points.

The predicted saliency map is first normalized to have zero mean and unit variance. NSS is then calculated as the average saliency value at ground truth fixation coordinates:

$$\mathrm{NSS}(S, F) = \frac{1}{N} \sum_{i=1}^{N} \hat{S}(F_i)$$

Where S^ is the normalized saliency map, Fi are the fixation locations, and N is the total number of fixations. An NSS score of 1 indicates that fixated points are, on average, one standard deviation above the mean saliency value.

In this project, the predicted saliency maps were normalized using scipy. stats. zscore, and NSS was computed by sampling the normalized map at fixation locations. This metric is particularly interpretable and penalizes both false positives and negatives, making it effective for identifying precise attention alignment.

## 4.2. Results

### 4.2.1. Overall Models Performance

To assess overall model performance across the test dataset, I calculated the average scores of the three models using the evaluation metrics explained in the previous sections **AUC-Judd (Area Under the ROC Curve – Judd**, **NSS (Normalized Scanpath Saliency)**, and **CC (Pearson's Correlation Coefficient).** For each of the three models: SALICON, SUM, and DeepGaze II E. This multi-metric evaluation approach provides a more comprehensive assessment of saliency prediction quality. AUC-Judd evaluates how well the model ranks fixated locations, NSS measures the spatial precision of predicted saliency at fixation points, and CC quantifies the structural similarity between predicted and ground-truth saliency maps. By averaging the scores across all test images, I ensured a fair and balanced comparison of model performance. Among the three models, SUM achieved the highest overall performance, with average scores of 0.91 (AUC-Judd), 4.05 (NSS), and 0.88 (CC). DeepGaze IIE with 0.88 (AUC-Judd), 2.14 (NSS), and 0.60 (CC), while SALICON recorded 0.92, 2.59, and 0.71 respectively.

The table 1 below summarises how each saliency model performed across the full dataset without yet splitting by scene complexity.

| Model | AUC-Judd | NSS | Pearson's CC |
|---|---|---|---|
| **DeepGaze IIE** | 0.88 | 2.14 | 0.60 |
| **Salicon** | 0.92 | 2.59 | 0.71 |
| **SUM** | 0.91 | 4.05 | 0.88 |

**Table 1: Overall performance of all models in all images.**

### 4.2.2. Impact of Scene Complexity on Model Performance

#### 4.2.2.1. Performance on Low-Complexity Images

In visual saliency research, low complexity images are those that feature simple visual content, such as a limited number of objects, low clutter, homogeneous textures, and a clear foreground-background distinction. These images reduce the cognitive load required to identify salient regions and are often used to evaluate how effectively a model can detect obvious attention-attracting elements (Kümmerer, Wallis & Bethge, 2016). Since the salient features are typically more prominent and isolated, models tend to perform better on such images compared to high complexity ones, where multiple stimuli compete for attention.

| Model | AUC-Judd | NSS | Pearson's CC |
|---|---|---|---|
| **DeepGaze IIE** | 0.90 | 2.78 | 0.62 |
| **Salicon** | 0.92 | 3.17 | 0.78 |
| **SUM** | 0.93 | 5.31 | 0.92 |

**Table 2: Performance of the models in low complexity images.**

From table 2 above, the SUM model outperforms both DeepGaze IIE and Salicon across all metrics, indicating superior ability to predict human attention in low complexity scenes. The high NSS (Normalized Scanpath Saliency) and Pearson's CC (Correlation Coefficient) values suggest strong agreement with actual human gaze patterns, while a high AUC-Judd indicates robust classification of salient versus non-salient regions.

#### 4.2.2.2. Performance on High-Complexity Images

In visual saliency, high complexity images refer to scenes that contain dense, diverse, and often overlapping visual information such as many objects, textured backgrounds, intricate patterns, or dynamic elements. These images pose greater challenges for computational models because multiple competing stimuli can make it difficult to isolate the truly salient regions. In such cases, saliency models must rely more heavily on higher-level cognitive cues such as context, semantics, and task relevance to accurately predict where human attention is likely to be directed (Zhang et al., 2018). As a result, performance often decreases on high complexity images compared to low complexity ones.

| Model | AUC-Judd | NSS | Pearson's CC |
|---|---|---|---|
| **DeepGaze IIE** | 0.85 | 1.34 | 0.46 |
| **Salicon** | 0.78 | 1.84 | 0.63 |
| **SUM** | 0.80 | 2.47 | 0.84 |

**Table 3: Performance of all the models on high complexity images.**

From table 3 above, we see that model performance drops on high complexity images across all metrics when compared to low complexity scenes. The SUM model still outperforms the others in terms of NSS and Pearson's CC, suggesting it handles visual clutter and complex semantic content more effectively. However, the overall decrease in AUC-Judd and NSS indicates that predicting human attention becomes significantly more difficult in such settings, particularly for models like Salicon, which relies more on coarse feature maps.

Generally, DeepGaze II E demonstrates consistent performance across both low- and high-complexity images, highlighting its robustness in modelling human visual attention. While all the other models experience a performance drop in high-complexity scenes, DeepGaze II E shows the least degradation, with relatively stable AUC-Judd (from 0.90 to 0.85), NSS (from 2.78 to 1.34), and Pearson's CC (from 0.62 to 0.46). This contrasts with larger performance declines seen in Salicon and SUM. Despite not achieving the highest absolute scores, DeepGaze II E's stable metrics across different image

complexities suggest strong generalization ability and reliability in varied visual environments.

#### 4.2.3. Comparative analysis and visualisations.

##### 4.2.3.1. Models Performance under AUC-Judd.

The **AUC-Judd** (Area Under the Curve - Judd) metric evaluates how well a saliency map can distinguish between fixated and non-fixated locations. Higher values (closer to 1) indicate better performance.

**DeepeGaze IIE.**

DeepGaze II E, although slightly behind at an average score of **0.88**, still delivered solid performance. Trained purely on high-fidelity eye-tracking data, it may be better suited for precise fixation prediction (as seen in other metrics like its stability across scene complexity) rather than broad fixated-region ranking.

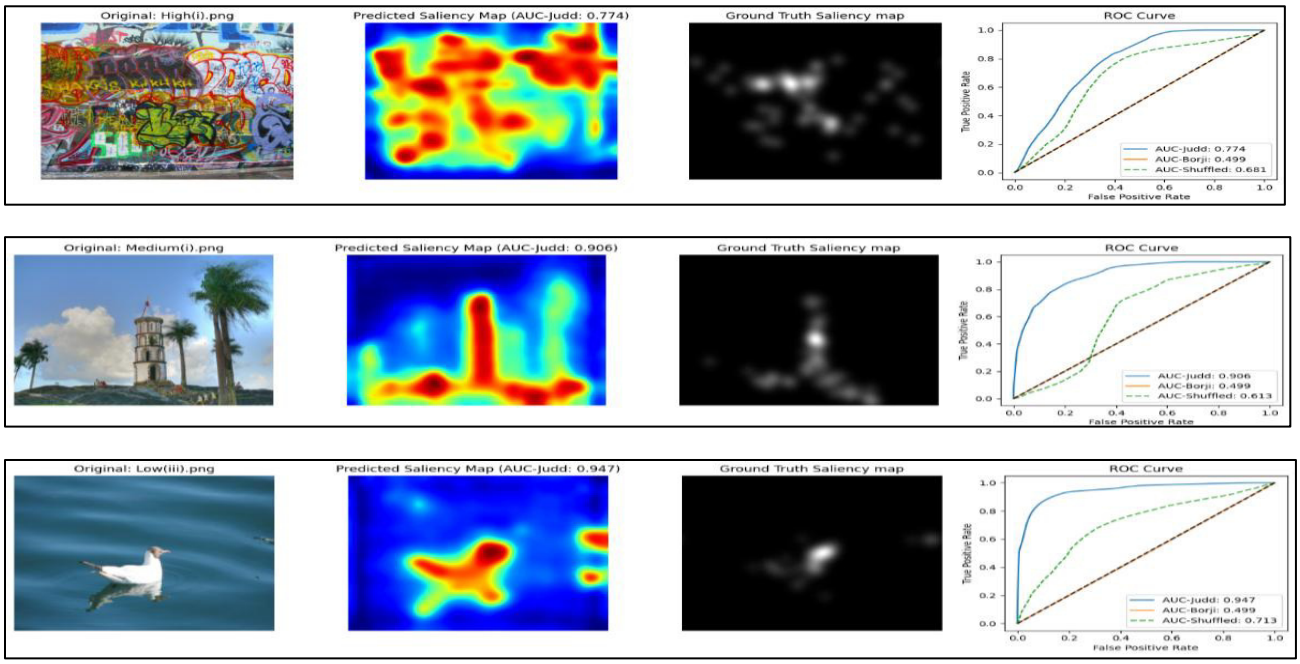



**Figure 7: Saliency prediction results of DeepGaze IIE evaluated using AUC-Judd on example images with high and low complexity. The figure shows the original image, predicted saliency map with AUC-Judd score, ground truth saliency map, and their corresponding ROC curve.**

**Salicon**

SALICON performed with an average **AUC-Judd score of 0.92**, performing strongly despite relying on mouse-tracking data alone. Its performance demonstrates that large-scale, crowd-sourced mouse-tracking datasets can still support highly competitive saliency prediction.

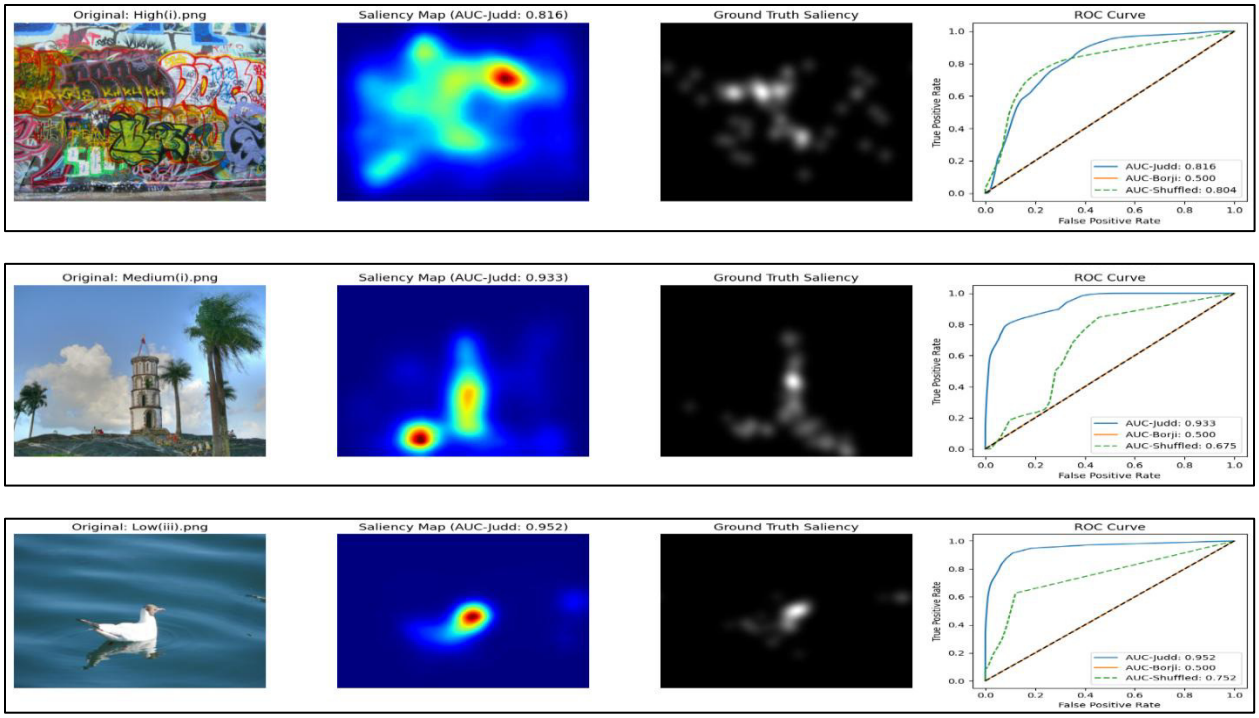
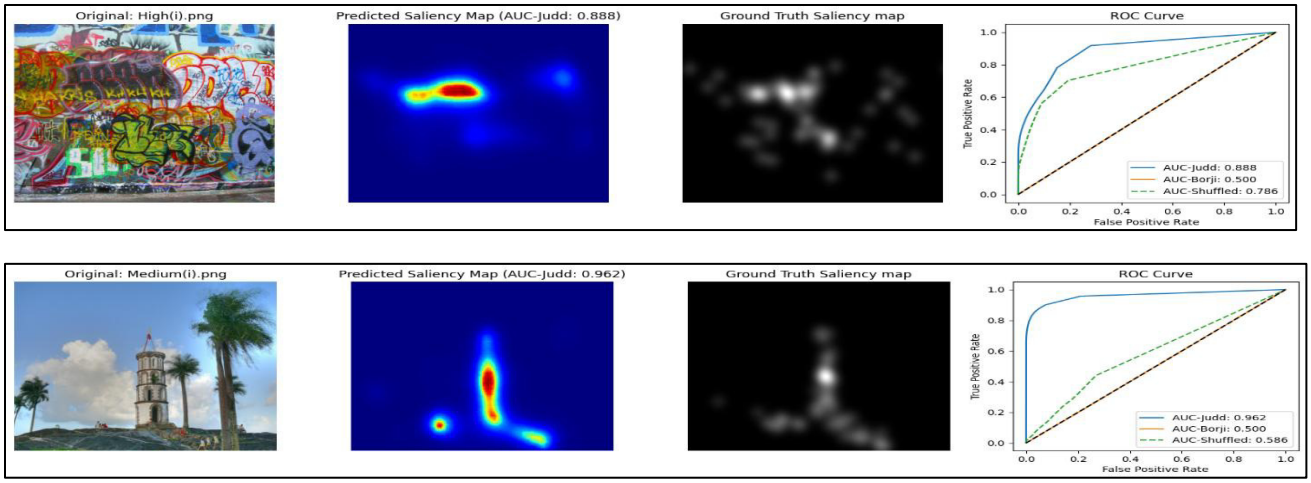



**Figure 8: Saliency prediction results of Salicon evaluated using AUC-Judd on example images with high and low complexity. The figure shows the original image, predicted saliency map with AUC-Judd score, ground truth saliency map, and their corresponding ROC curve.**

### Saliency Unification through Mamba (SUM)

**SUM** achieved the average **AUC-Judd score (0.91)**, suggesting it is effective in ranking true human fixation points over non-fixated regions. This indicates robust generalisation across varying scenes, likely aided by its hybrid training on both eye- and mouse-tracking data.



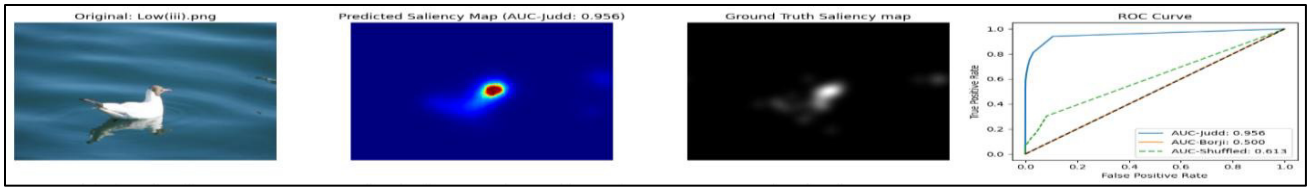



**Figure 9: Saliency prediction results of SUM evaluated using AUC-Judd on example images with high and low complexity. The figure shows the original image, predicted saliency map with AUC-Judd score, ground truth saliency map, and their corresponding ROC curve.**

#### 4.2.3.2. Models Performance Under Normalised Scanpath Saliency (NSS).

The **NSS** metric evaluates how well a predicted saliency map aligns with actual human fixation points, normalized by the map's standard deviation. Higher values reflect better correspondence to true human gaze.

**DeepGaze IIE**

**DeepGaze II E**, despite being trained solely on eye-tracking data, had the **lowest average NSS score of 2.14**. While this model is known for capturing general attention distribution and semantic saliency, it may struggle with pinpointing exact fixation locations, especially in images with high spatial variability.

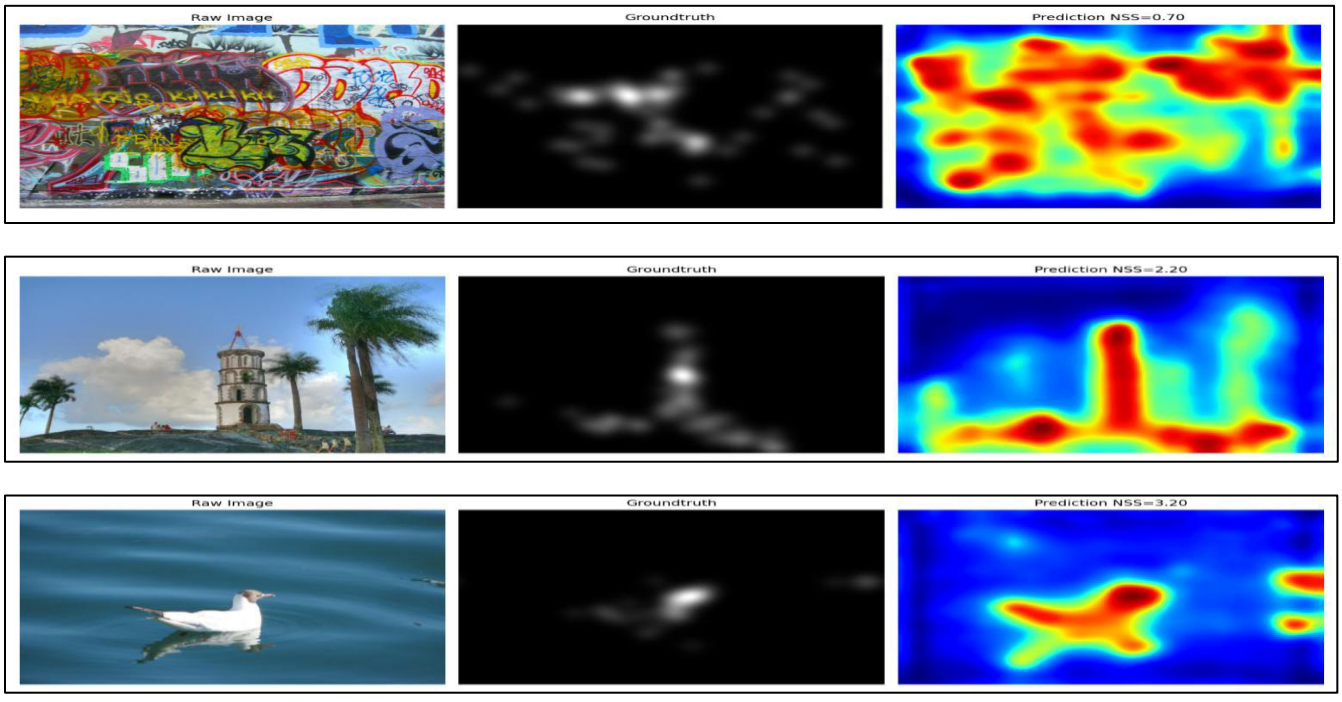



**Figure 10: Example saliency predictions and their corresponding Normalized Scanpath Saliency (NSS) score results of DeepGaze IIE. Each row shows the raw image, the ground truth saliency map, and the model's prediction with its NSS value.**

### Salicon

**SALICON** followed with a moderate **NSS score of 2.59**, showing that while its mouse-tracking-based training enables decent fixation prediction, it lacks the spatial precision seen in SUM. This may be due to the inherent noise and variability in mouse-tracking data compared to eye-tracking.

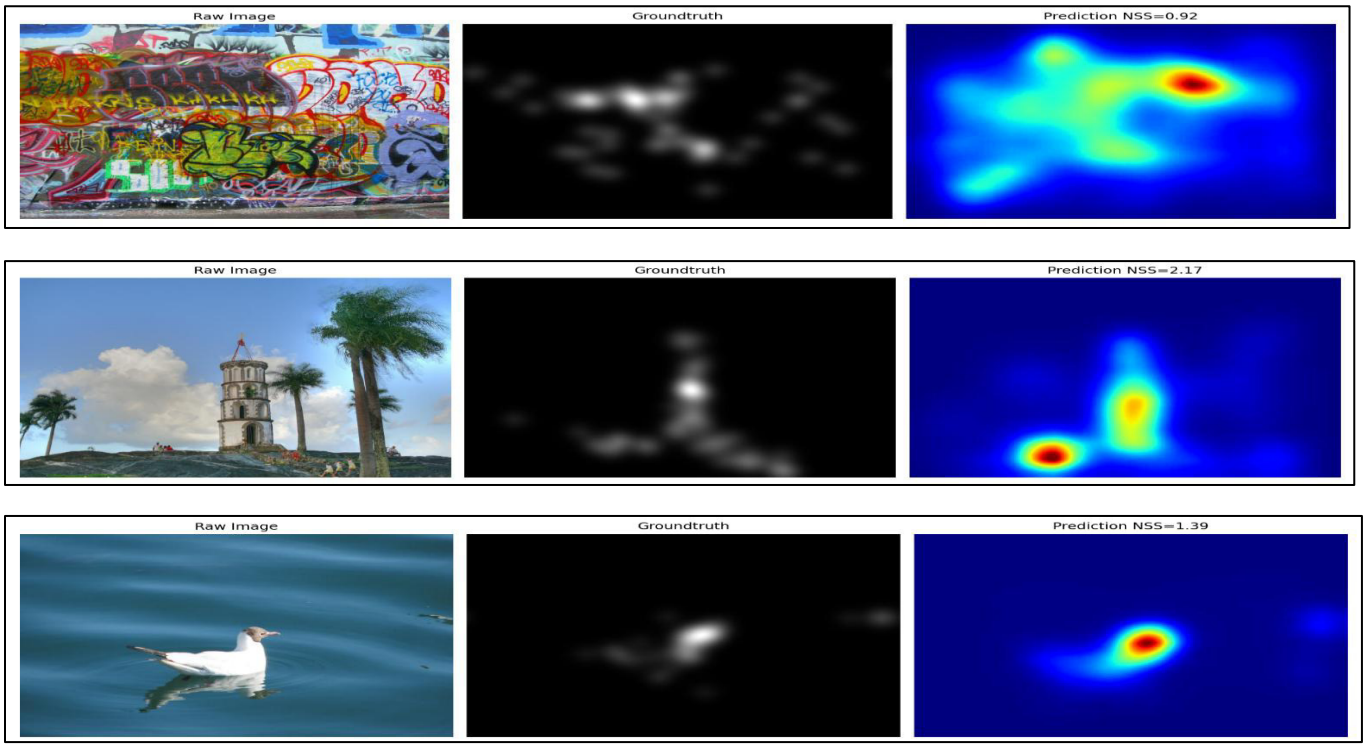
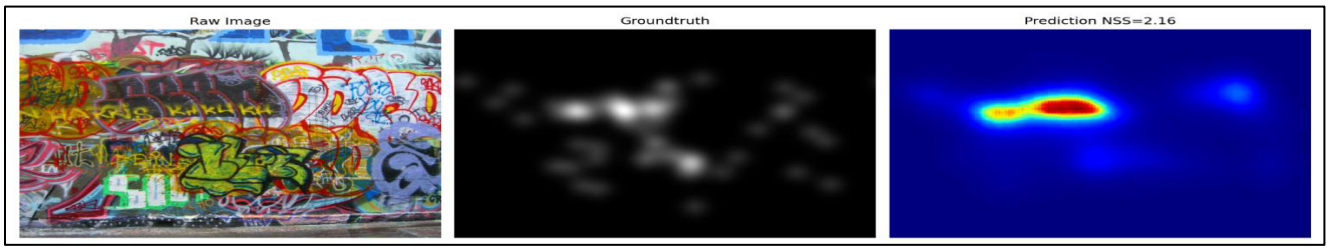



**Figure 11: Example saliency predictions and their corresponding Normalized Scanpath Saliency (NSS) score results of Salicon. Each row shows the raw image, the ground truth saliency map, and the model's prediction with its NSS value.**

### Saliency Unification through Mamba.

**SUM** again outperformed the other models with an **average NSS of 4.05**, indicating its strong capability to produce saliency maps that closely match human fixations. This high score suggests the model generates sharp, well-localised saliency peaks aligned with human gaze, particularly due to its hybrid training and advanced architecture.



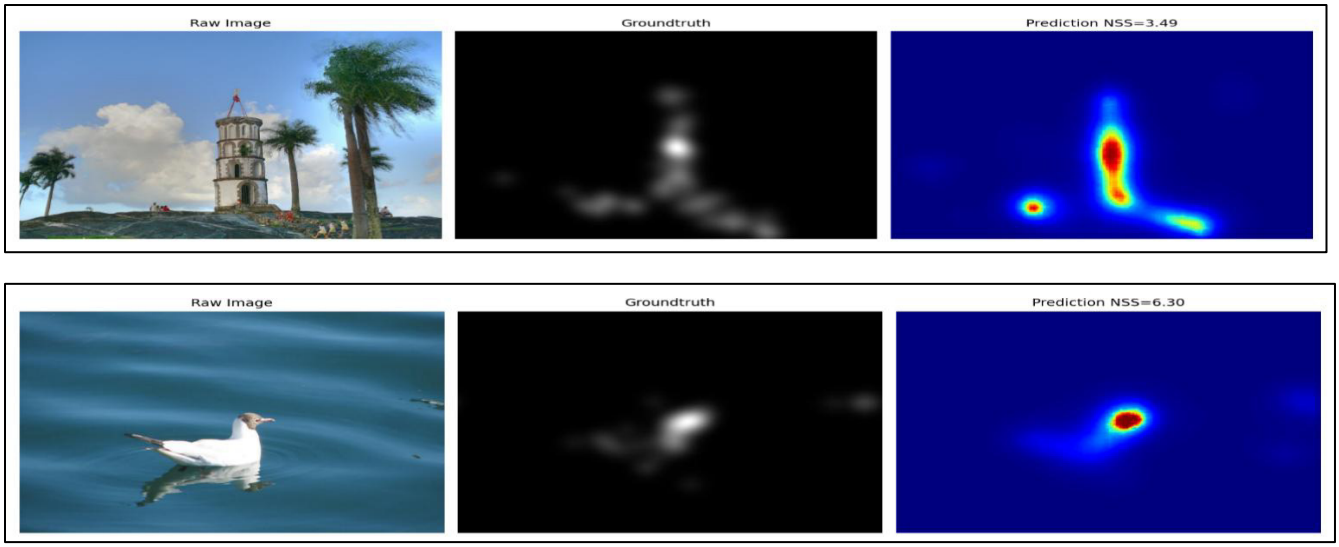


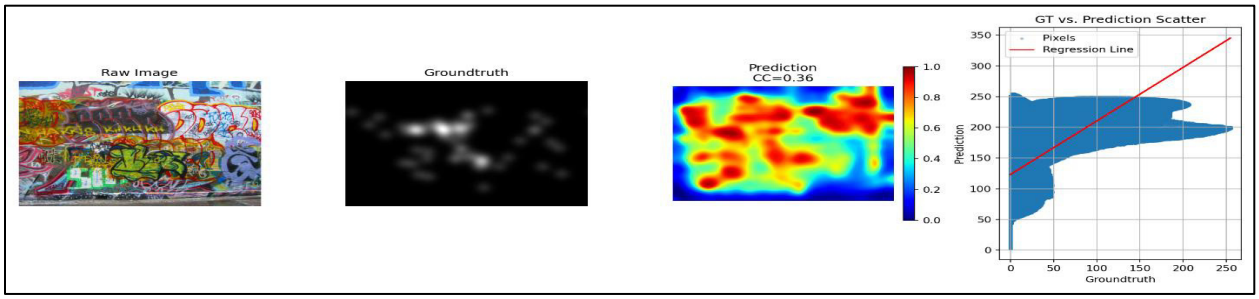

**Figure 12: Example saliency predictions and their corresponding Normalized Scanpath Saliency (NSS) score results of SUM. Each row shows the raw image, the ground truth saliency map, and the model's prediction with its NSS value.**

#### 4.2.3.3. Models Performance Under *Pearson's Correlation* Coefficient (CC).

Pearson's Correlation Coefficient (CC) quantifies the linear relationship between the predicted saliency maps and the ground-truth human fixation maps. A higher CC value indicates that the model's predictions align more closely with human attention patterns in terms of spatial structure and distribution.

**DeepGaze IIE**

**DeepGaze II E** scored an average CC score of **0.60**, indicating a strong correlation between its saliency maps and human fixations. This superior performance is attributed to its use of high-fidelity eye-tracking data and a more sophisticated architecture that includes probabilistic output and a fixed center bias, which better captures the spatial structure of human gaze.



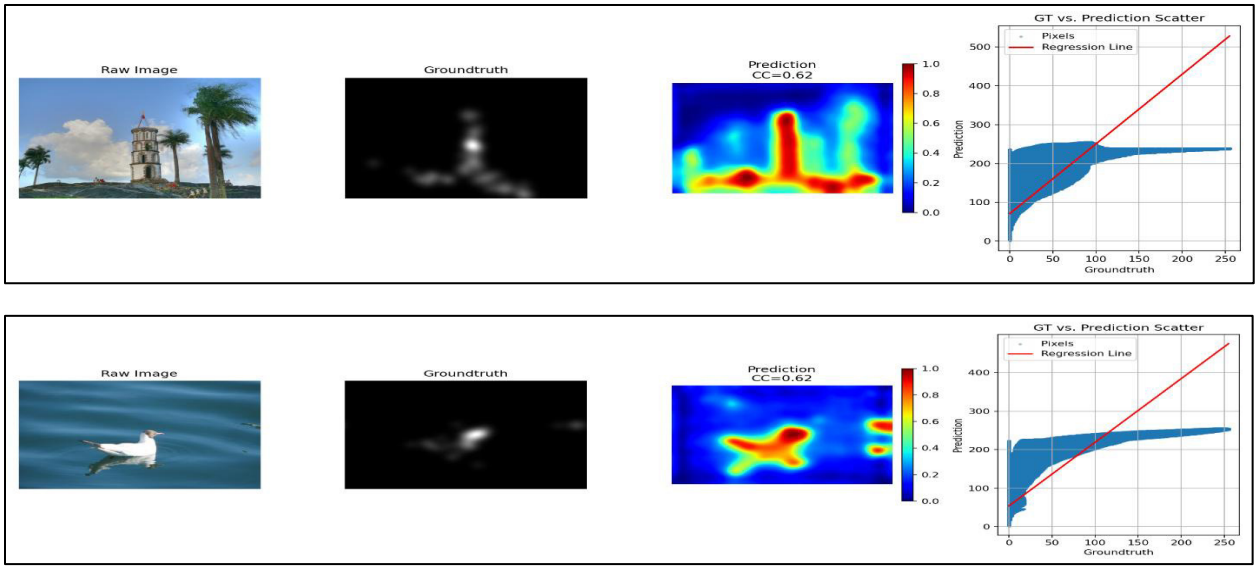
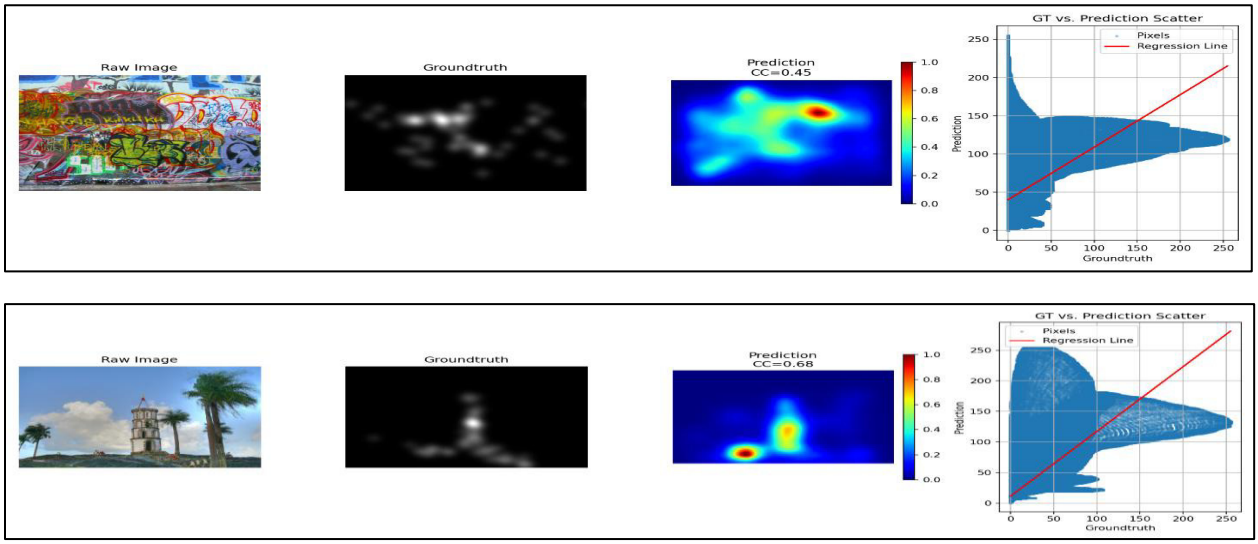



**Figure 13: Comparison of ground truth saliency maps and those predicted by DeepGaze IIE using the correlation coefficient (CC). Each row displays the raw image, the ground truth saliency map, the model's prediction, and a scatter plot comparing ground truth and predicted values, along with the corresponding CC score.**

### Salicon

**SALICON** achieved an average CC of **0.71**, indicating strong alignment between its predictions and ground-truth maps. While respectable, the performance suggests that SALICON may struggle to capture finer spatial patterns in attention, likely due to its reliance on mouse-tracking data, which lacks the precision of eye-tracking.



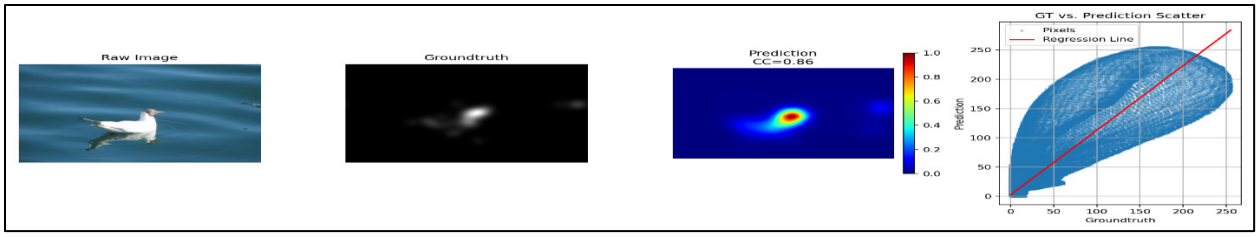



**Figure 14: Comparison of ground truth saliency maps and those predicted by SALICON using the correlation coefficient (CC). Each row displays the raw image, the ground truth saliency map, the model's prediction, and a scatter plot comparing ground truth and predicted values, along with the corresponding CC score.**

**Saliency Unification through Mamba (SUM).**

**SUM (Saliency Unification through Mamba)** performed better, with a CC score of **0.88**. This reflects an improvement in structural similarity, likely due to its hybrid training on both mouse and eye-tracking data. The fusion of modalities helps SUM better approximate human gaze distribution.

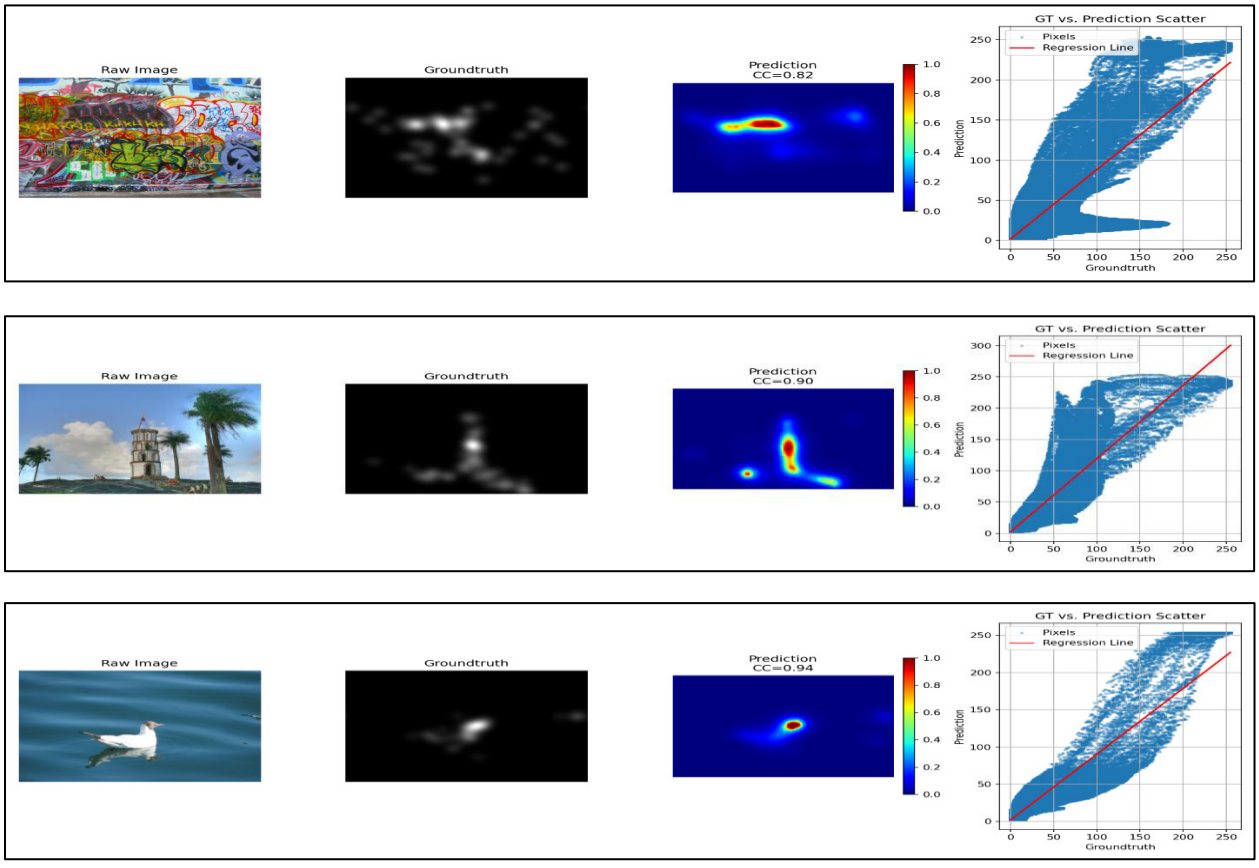



**Figure 15: Comparison of ground truth saliency maps and those predicted by SUM using the correlation coefficient (CC). Each row displays the raw image, the ground truth saliency map, the model's prediction, and a scatter plot**

**comparing ground truth and predicted values, along with the corresponding CC score.**

#### 4.2.4. Summary of Results and Findings

This study conducted a comprehensive evaluation of three state-of-the-art saliency models **SALICON**, **SUM (Saliency Unification through Mamba)**, and **DeepGaze II E** to assess their performance in predicting human visual attention across images with varying scene complexities. The results reveal key insights into how model architecture and input modality influence prediction accuracy and robustness.

1. **DeepGaze II E had consistent performance in different complexities** across all evaluation metrics (AUC-Judd, NSS, and Pearson's CC). This suggests a higher capacity to capture semantically meaningful and spatially precise visual attention cues.
2. **All three models performed significantly better on low-complexity images**, confirming that visual clutter and increased object density introduce challenges for saliency prediction models. However, the performance degradation in DeepGaze II E was less pronounced compared to SALICON and SUM.
3. **Scene complexity negatively impacts saliency prediction**, but the extent varies across models. DeepGaze II E showed strong generalization and stability, indicating its robustness to cluttered, semantically rich, or multi-object scenes. SUM maintained moderate consistency, while SALICON showed greater performance degradation under high-complexity conditions.

These trends were confirmed both quantitatively via aggregated metric scores and qualitatively through visual inspection of saliency maps. DeepGaze II E generated sharper and more accurate predictions, even in scenes with multiple competing salient regions. SALICON, while scalable due to its reliance on mouse-tracking data, struggled with spatial precision. SUM performed better than SALICON due to its hybrid training approach but lacked the high-level semantic understanding present in DeepGaze II E.

In the **Discussion** section that follows, these findings are further contextualized by analysing architectural and methodological distinctions that likely account for DeepGaze II E's superior performance. In particular, the integration of pretrained semantic features, probabilistic fixation modelling, and a learned center bias appear to be critical design choices supporting its accuracy and robustness.

## Discussion

The quantitative and qualitative results presented in the previous section clearly indicate that **DeepGaze II E** had consistent performance than SALICON and **SUM**, even on images of high visual complexity. This section explores the architectural and

methodological design choices that likely contribute to DeepGaze II E's superior accuracy and robustness in saliency prediction.

### Architectural Advantages of DeepGaze II E

One of the key strengths of DeepGaze II E lies in its pretrained feature extractor specifically, a VGG-19 network trained on the ImageNet dataset. This enables the model to extract high-level semantic features, such as faces, animals, and text, which are known to strongly influence human visual attention (Simonyan & Zisserman, 2015). In contrast, models trained from scratch or using less powerful backbones may lack the representational richness to detect such cues reliably.

Another distinctive component is its probabilistic readout model, which generates a log-probability distribution over the image pixels. This mechanism enables DeepGaze II E to directly model the probability distribution of human fixations, unlike SALICON and SUM, which rely on empirical heatmap matching. Furthermore, the integration of a learned center bias reflects the natural human tendency to fixate near the image center (Tatler, 2007), enhancing the model's ecological validity.

### Training Strategy Differences

From a training perspective, DeepGaze II E adopts a **data-efficient strategy** by freezing the VGG-19 backbone and only fine-tuning the readout layer. This approach minimizes the risk of overfitting, particularly when using relatively small eye-tracking datasets, while preserving the generalization capability learned from large-scale pretraining.

The model also employs a **log-likelihood loss function**, directly optimizing for fixation prediction accuracy in a probabilistic sense, rather than using mean squared error or L2 losses common in SALICON and SUM. This probabilistic framework aligns more closely with the statistical nature of human gaze patterns.

Moreover, unlike conventional models that generate static saliency maps, DeepGaze II E models the **true fixation distribution**, offering a more theoretically grounded representation of visual attention (Kümmerer et al., 2016).

### Limitations of SALICON and SUM

In contrast, **SALICON** utilizes a dual-column CNN trained on mouse-tracking data, which while scalable can introduce **noisier and less precise** spatial saliency signals compared to eye-tracking data (Huang et al., 2015). Mouse clicks tend to lag behind fixations and are often biased by task demands, which affects model accuracy.

**SUM**, designed with computational efficiency in mind, features a lightweight architecture optimized for speed. While beneficial for real-time applications, this simplicity limits its

ability to capture **complex semantic relationships** in cluttered scenes, impacting its overall predictive performance.

### Robustness to Scene Complexity

DeepGaze II E demonstrates **robust generalization across different levels of scene complexity**, a result of its rich feature extraction, probabilistic modelling, and data-efficient training. Even in highly cluttered or semantically dense scenes, the model maintains stable performance by identifying salient objects amidst noise. This contrasts with SALICON and SUM, whose performance tends to degrade more significantly under the same conditions.

These findings underscore the importance of considering **both architectural design and training modality** when developing saliency models. DeepGaze II E's approach anchored in high-level feature extraction, probabilistic readout, and efficient fine-tuning sets a benchmark for robustness and accuracy in the field.

## 5. Conclusions and Future work

### 5.1. Conclusions

This project evaluated the performance of three state-of-the-art deep learning-based saliency models SALICON (Huang et al., 2015), SUM (Islam et al., 2020), and DeepGaze II E (Kümmerer et al., 2016) across images of varying scene complexity. Using established metrics such as AUC-Judd, NSS, and Pearson's Correlation Coefficient (CC), the analysis revealed important trends regarding the influence of model architecture and input modality on saliency prediction performance.

DeepGaze II E emerged as the most robust and consistent model across both low- and high-complexity scenes. Its superior performance is largely attributable to its training on high-fidelity eye-tracking data and its architecture, which leverages deep convolutional feature maps extracted from a pretrained VGG network and applies a probabilistic readout mechanism optimized for fixation prediction (Kümmerer et al., 2016). In contrast, SALICON, trained solely on mouse-tracking data, showed weaker generalization in complex scenes, confirming previous findings that mouse-tracking may lack the spatial precision needed to model human gaze accurately (Chen et al., 2021). The SUM model, trained on hybrid datasets combining both modalities, showed modest improvements over SALICON, but did not match the performance of DeepGaze II E, suggesting that hybrid models still face challenges in effectively fusing data sources with differing noise and resolution levels (Islam et al., 2020).

These findings reinforce the notion that input modality and model architecture are central to saliency modelling success. While mouse-tracking data offers scalability, eye-tracking remains the most reliable modality for capturing human visual attention (Bylinskii et al., 2019). Architecturally, models that integrate context-aware mechanisms,

deep feature hierarchies, and probabilistic saliency readouts are better equipped to handle visual scenes of varying complexity.

### 5.2. Future Work

Building on the findings and limitations of this study, several avenues for future research are proposed to further advance the field of visual saliency modelling:

- **Complexity-Aware Architectures:** Future research should explore the design of saliency models that dynamically adapt to varying levels of scene complexity. Incorporating visual clutter classifiers, scene-type detectors, or attention-modulated routing mechanisms within model architectures may enhance performance in visually diverse and cluttered environments. These adaptive mechanisms could improve saliency prediction by allocating computational resources according to the perceptual demands of the input image.
- **Advanced Hybrid Training Strategies:** Although hybrid datasets combining mouse- and eye-tracking data offer a scalable alternative to traditional eye-tracking, the performance of models trained on such data is often affected by the domain gap between modalities. Future work should investigate advanced learning strategies such as domain adaptation, uncertainty modelling, and multi-task learning to effectively bridge this gap and enhance model robustness. These methods would allow for better exploitation of large-scale mouse-tracking datasets while maintaining the spatial precision characteristic of eye-tracking data (Huang et al., 2015; Tavakoli et al., 2017).
- **Scalable Eye-Tracking Techniques:** The high cost and limited availability of eye-tracking hardware pose barriers to data collection at scale. Emerging webcam-based systems, such as GazeCapture (Krafka et al., 2016), present promising alternatives for large-scale gaze data collection in uncontrolled environments. Future studies could investigate the integration of such scalable techniques to build more diverse and representative datasets, thereby improving the generalizability of saliency models across populations and use cases.
- **Task-Dependent and Personalized Saliency Models:** Human attention is often guided by goals, intentions, and contextual relevance. As such, future models should incorporate task-specific or user-personalised attention mechanisms. These approaches hold potential in applications such as adaptive interfaces, targeted advertising, educational technologies, and assistive systems. Modelling attention based on user behaviour and contextual cues would represent a significant step toward achieving human-like and functionally useful saliency prediction (Judd et al., 2009).
- **Statistical Analysis of Scene Complexity Effects:** Due to time constraints, this study was limited in its statistical treatment of the relationship between scene complexity and model performance. Future work should incorporate robust statistical methods such as Pearson or Spearman correlation analyses to

quantify the sensitivity of saliency models to variations in scene complexity. Additionally, statistical significance testing (e.g., ANOVA or t-tests) could be used to rigorously assess differences in performance across models and complexity levels.

- **Expansion of Model and Dataset Scope:** This project was constrained to three deep learning-based saliency models and a single dataset. With more time and resources, future studies should consider evaluating a broader range of models, including transformer-based or diffusion-based saliency predictors, and using multiple publicly available datasets with varied content and annotation modalities (e.g., MIT1003, CAT2000, or SALICON). This would provide a more comprehensive understanding of model generalizability across data domains.
- **Inclusion of Additional Evaluation Metrics:** While this study employed AUC-Judd, NSS, and Pearson Correlation Coefficient (CC), future research could benefit from incorporating other complementary metrics such as Kullback–Leibler Divergence (KLDiv), Earth Mover's Distance (EMD), and Similarity (SIM). These metrics can offer additional insights into the spatial distribution and perceptual relevance of predicted saliency maps.

Pursuing these research directions will support the development of next-generation saliency models that are not only more accurate, but also more adaptive, scalable, and cognitively aligned with human perception. Such advancements are expected to enhance the impact of computational attention modelling in domains including human-computer interaction, autonomous navigation, medical imaging, and cognitive neuroscience.

## 6. Reflection on Learning

Undertaking this project has been a significant learning experience that has deepened my understanding of both the technical and theoretical aspects of visual saliency modelling. From the initial stages of conducting a literature review to the final phase of evaluating model performance, I have developed a comprehensive appreciation for the interdisciplinary nature of computational saliency research combining elements from computer vision, cognitive psychology, machine learning, and human-computer interaction.

One of the most valuable aspects of this project was gaining practical experience with deep learning-based models such as SALICON, SUM, and DeepGaze II E. This involved not only understanding their architectural differences and training modalities but also implementing and evaluating them using standard metrics such as AUC-Judd, Normalized Scanpath Saliency (NSS), and Pearson Correlation Coefficient (CC). Working hands-on with these models helped me develop stronger skills in model evaluation, result interpretation, and the challenges of working with saliency datasets.

I also learned the importance of data preprocessing and feature extraction, particularly in the context of computing scene complexity. Designing a framework to quantify complexity using edge density and colour variation taught me how low-level image statistics can affect high-level model behaviour. Additionally, visualizing these metrics and correlating them with model performance offered insights into the nuanced relationship between image content and attention prediction.

Another important takeaway from this project was recognising the practical limitations of computational saliency models especially those related to data quality and scalability. This led me to explore emerging trends such as hybrid data collection techniques and scalable eye-tracking alternatives, which I now see as vital for advancing real-world applications.

While I am pleased with the depth and breadth of this project, time constraints limited the scope of the evaluation. I would have liked to include more computational models, explore additional datasets, and conduct more rigorous statistical analysis to validate the findings. Evaluating performance with a wider range of metrics, such as KL Divergence and Earth Mover's Distance, could also have enriched the study. These are areas I hope to explore in future work or academic pursuits.

Overall, this project has significantly improved my research, programming, and critical thinking skills. It has reinforced my interest in human-centred AI systems and provided a strong foundation for future research in visual attention modelling and computer vision more broadly.

## References

Borji, A. and Itti, L., 2013. State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1), pp.185–207.

Bruce, N.D.B. and Tsotsos, J.K., 2005. Saliency based on information maximization. In: Advances in Neural Information Processing Systems (NIPS 2005). Vancouver, Canada, 5–8 December 2005. MIT Press, pp.155–162.

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A. and Durand, F., 2019. What do different evaluation metrics tell us about saliency models? IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(3), pp.740–757.

Canny, J., 1986. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8(6), pp.679–698.

Chen, T., Zhou, Y., Ma, K. and Ji, R., 2021. MouseView: Improving mouse tracking for attention prediction. IEEE Transactions on Multimedia, 23, pp.1942–1956.

Chen, Z., Zhao, Y., Shen, W., Yuan, Y., Hua, X.S. and Shao, L., 2021. Saliency prediction with mouse tracking data: A comparative study. Pattern Recognition, 116, p.107950.

Cornia, M., Baraldi, L., Serra, G. and Cucchiara, R., 2018. Predicting human eye fixations via an LSTM-based saliency attentive model. IEEE Transactions on Image Processing, 27(10), pp.5142–5154.

Dhara, G. and Kumar, K.R., 2023. Concepts and Techniques of Graph Neural Networks. [no publisher provided].

Hofheimer, J.A. and Lester, B.M., 2008. Neuropsychological Assessment. In: T.D. Wachs and M.H. Bornstein, eds. Encyclopedia of Infant and Early Childhood Development. Oxford: Academic Press, pp.425–438.

Hosseini, A., Kazerouni, A., Akhavan, S., Brudno, M. and Taati, B., 2024. SUM: Saliency Unification through Mamba for Visual Attention Modeling. arXiv preprint, arXiv:2406.17815.

Huang, X., Shen, C., Boix, X. and Zhao, Q., 2015. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Santiago, Chile, 7–13 December 2015. IEEE, pp.262–270.

Islam, M.T., Misu, T., Ma, K., Felsen, P. and Oliva, A., 2020. A hybrid model for predicting human visual attention on natural images. In: Proceedings of the European Conference on Computer Vision (ECCV). Glasgow, UK, 23–28 August 2020. Springer, pp.496–512.

Islam, M.T., Rashid, S., Siddiquie, B., Metaxas, D. and Roy-Chowdhury, A.K., 2020. Gaze-based video summarization using an attention-unified network. In: Proceedings of

the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp.2241–2250.

Itti, L. and Koch, C., 2001. Computational modelling of visual attention. Nature Reviews Neuroscience, 2(3), pp.194–203.

Itti, L., Koch, C. and Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11), pp.1254–1259.

Jiang, M., Huang, S., Duan, J. and Zhao, Q., 2015. SALICON: Saliency in context. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA, 7–12 June 2015. IEEE, pp.1072–1080.

Judd, T., Ehinger, K., Durand, F. and Torralba, A., 2009. Learning to predict where humans look. In: IEEE 12th International Conference on Computer Vision (ICCV). Kyoto, Japan, 27 September–4 October 2009. IEEE, pp.2106–2113.

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W. and Torralba, A., 2016. Eye tracking for everyone. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2176–2184.

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

Linardos, A., Dorr, M. and Bethge, M., 2021. DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modelling. arXiv preprint, arXiv:2111.12203.

Leiva, L.A., Huang, J. and White, R.W., 2020. Gaze prediction for mobile user interfaces. ACM Transactions on Interactive Intelligent Systems.

Peters, R.J., Iyer, A., Itti, L. and Koch, C., 2005. Components of bottom-up gaze allocation in natural images. Vision Research, 45(18), pp.2397–2416.

Ramanishka, V., 2016. Top-down visual saliency guided by captions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.7206–7215.

Simonyan, K. and Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations (ICLR).

Tatler, B.W., 2007. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. Journal of Vision, 7(14), pp.1–17.

Tavakoli, H.R., Borji, A., Laaksonen, J. and Suominen, H., 2017. Exploiting inter-image similarity and group information for personalized saliency prediction. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp.4781–4790.

Zhang, J., Sclaroff, S., Lin, Z., Shen, X., Price, B. and Mech, R., 2018. Minimum barrier salient object detection at 80 fps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1404–1412.

Zhao, Q., Jiang, M., Huang, S. and Wang, Y., 2021. Toward understanding scene complexity from saliency perspective. IEEE Transactions on Image Processing, 30, pp.1570–1583.

Zhao, R., Wang, X., Guo, Y., Xu, M., Li, Z., Ding, E. and Tao, D., 2021. Deep learning vs. traditional algorithms for saliency prediction of distorted images. Pattern Recognition, 110, p.107658.

**Additional Predicted Saliency Maps**

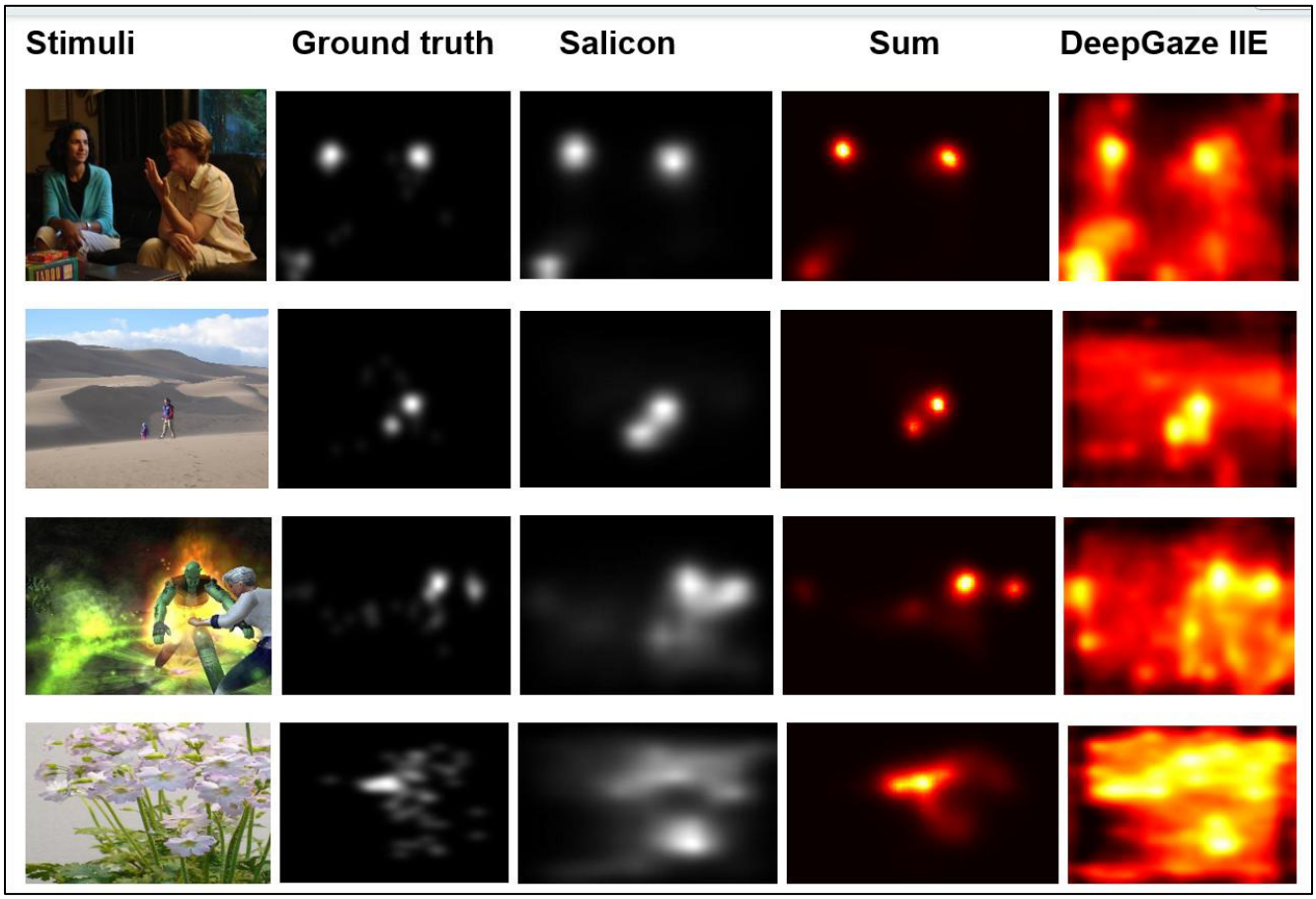
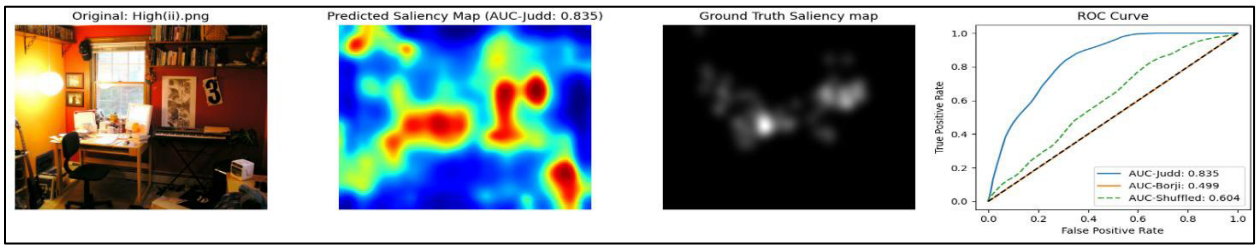



**Figure 16: Additional Image(s) with their corresponding ground truth saliency maps and predicted saliency maps generated by the computational models.**

**Additional results of models' Performance under AUC-Judd.**

**DeepGaze IIE**



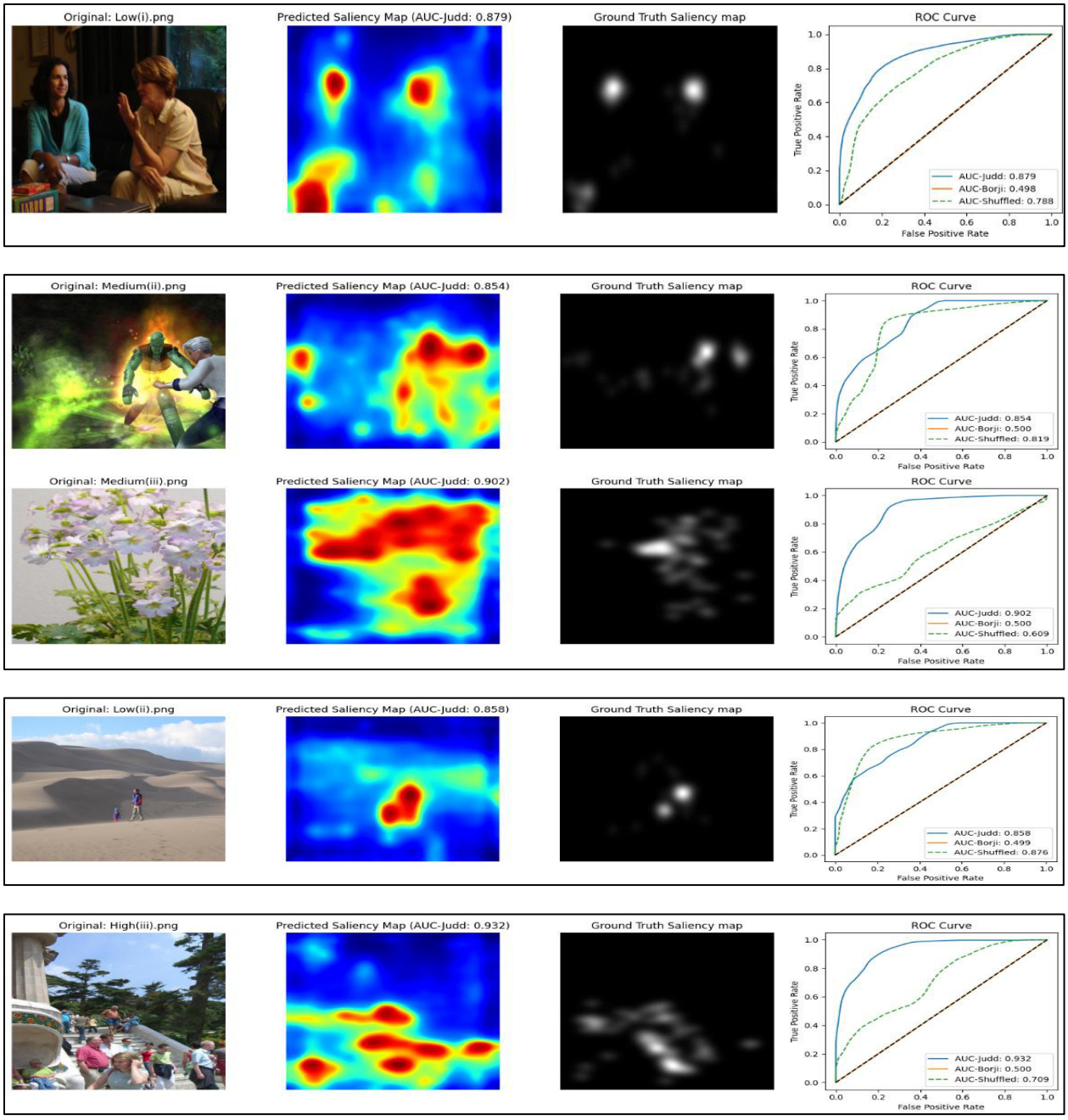



**Figure 17: Additional Saliency prediction results of DeepGaze IIE evaluated using AUC-Judd on example images with high and low complexity. The figure shows the original image, predicted saliency map with AUC-Judd score, ground truth saliency map, and their corresponding ROC curve.**

## Salicon

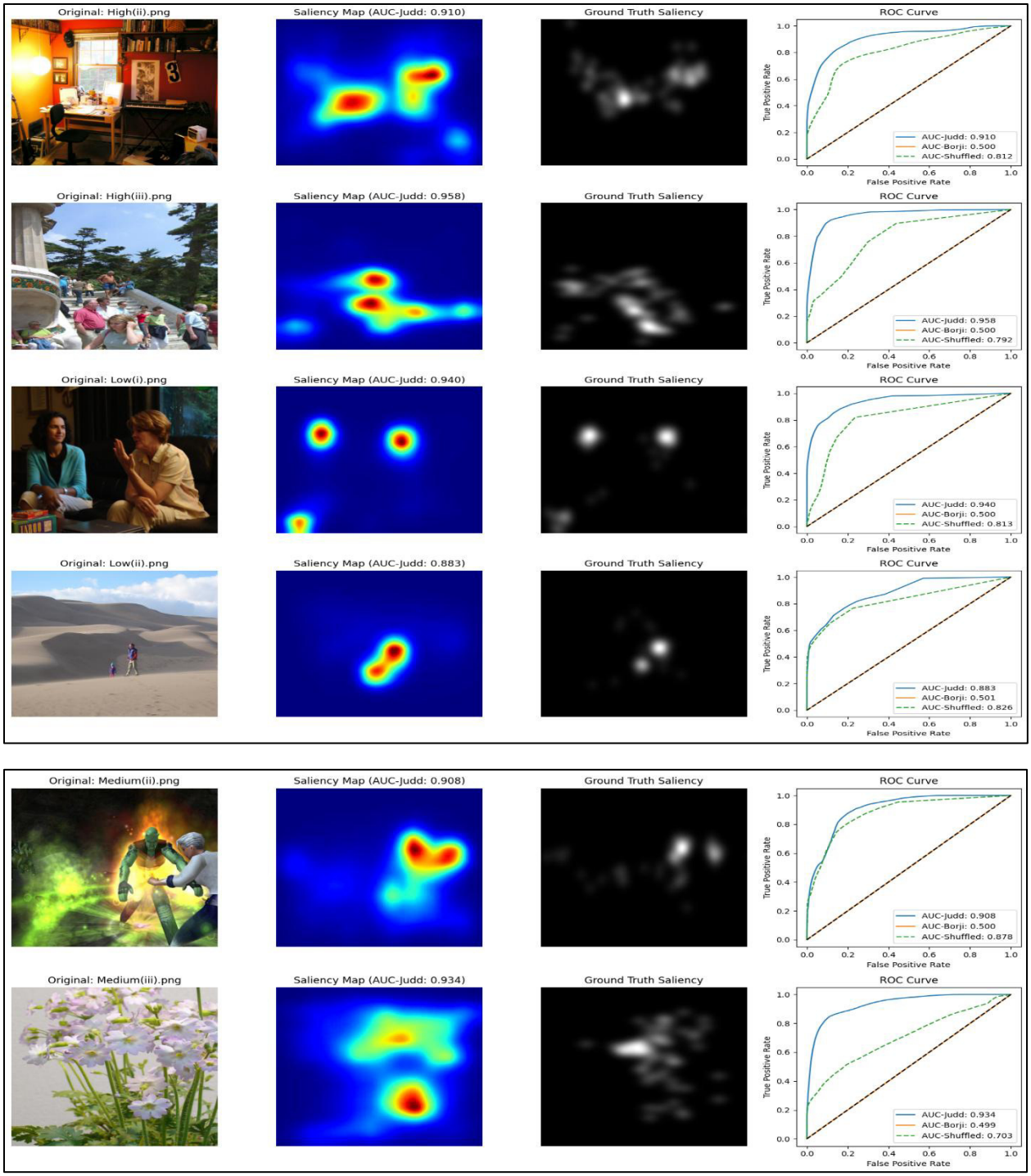



**Figure 18: Additional Saliency prediction results of DeepGaze IIE evaluated using AUC-Judd on example images with high and low complexity. The figure shows the original image, predicted saliency map with AUC-Judd score, ground truth saliency map, and their corresponding ROC Curve.**

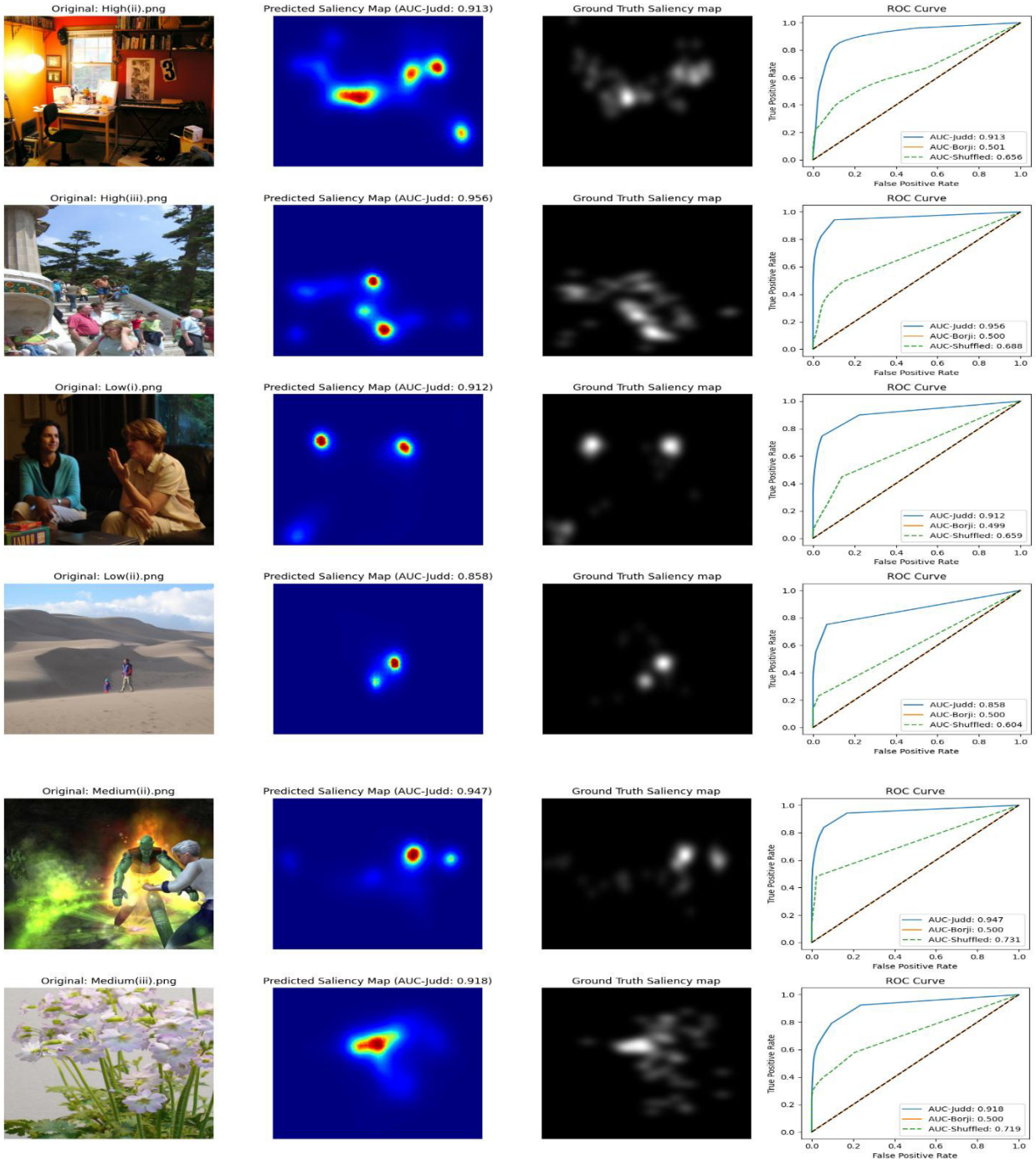
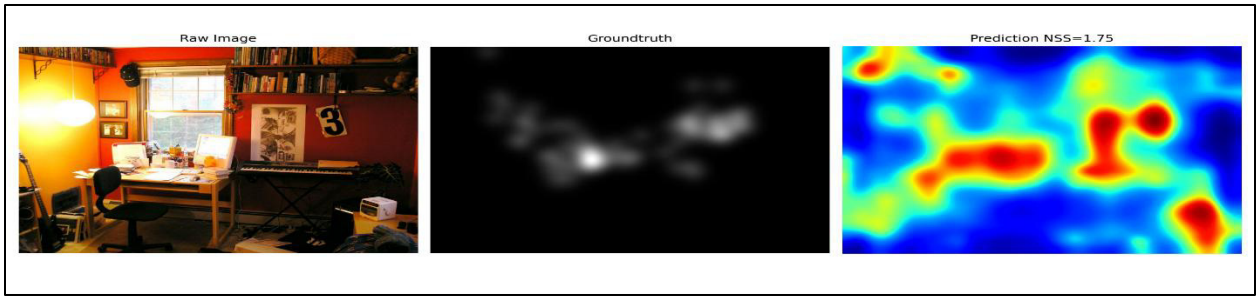
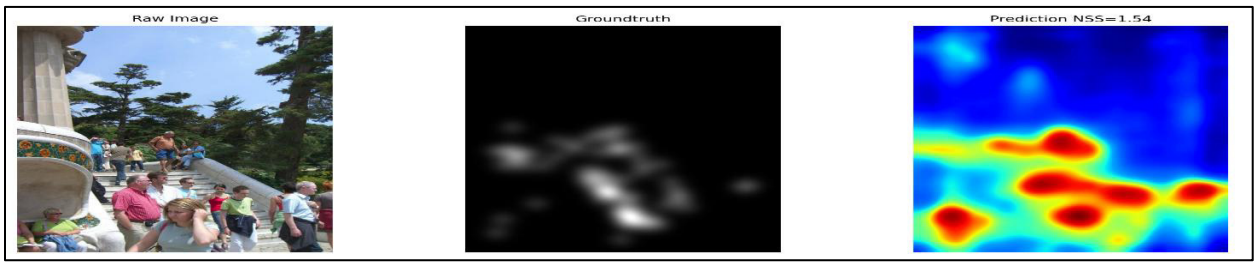


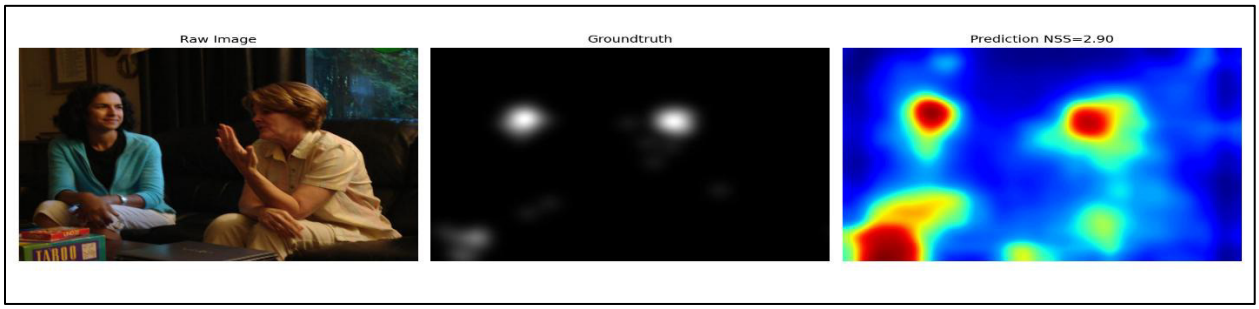

**Figure 19: Additional Saliency prediction results of DeepGaze IIE evaluated using AUC-Judd on example images with high and low complexity. The figure shows the original image, predicted saliency map with AUC-Judd score, ground truth saliency map, and their corresponding ROC Curve.**

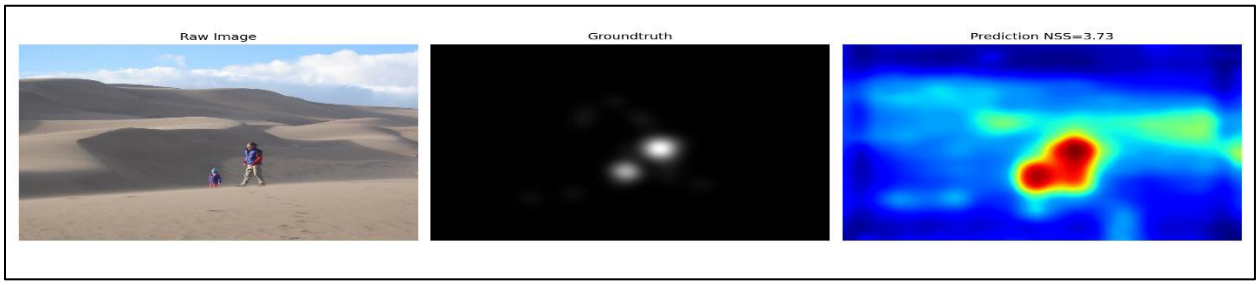
**Additional results of models' Performance under Normalised Scanpath Saliency (NSS).**

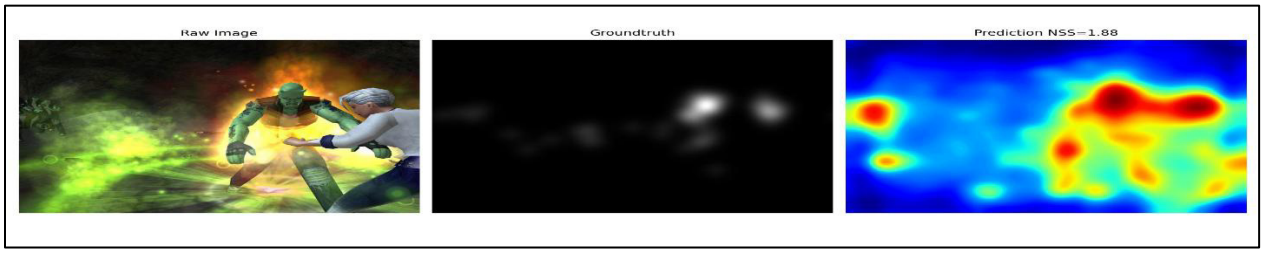
**DeepGaze IIE**


Raw Image
Groundtruth
Prediction NSS=1.75


Raw Image
Groundtruth
Prediction NSS=1.54


Raw Image
Groundtruth
Prediction NSS=2.90


Raw Image
Groundtruth
Prediction NSS=3.73


Raw Image
Groundtruth
Prediction NSS=1.88

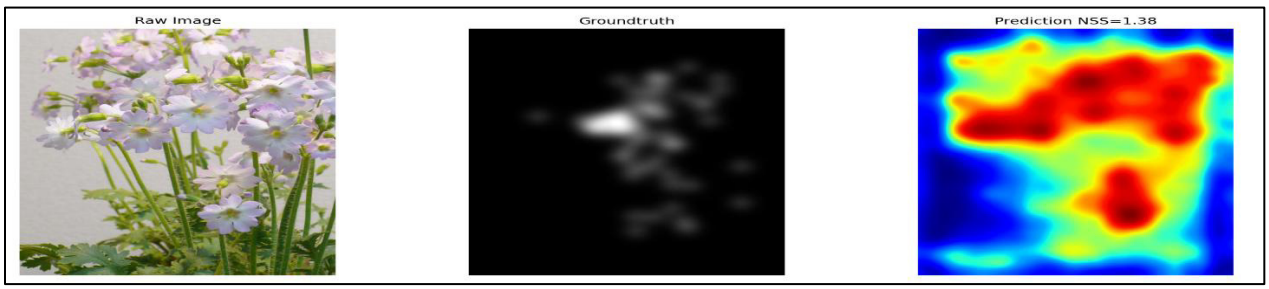
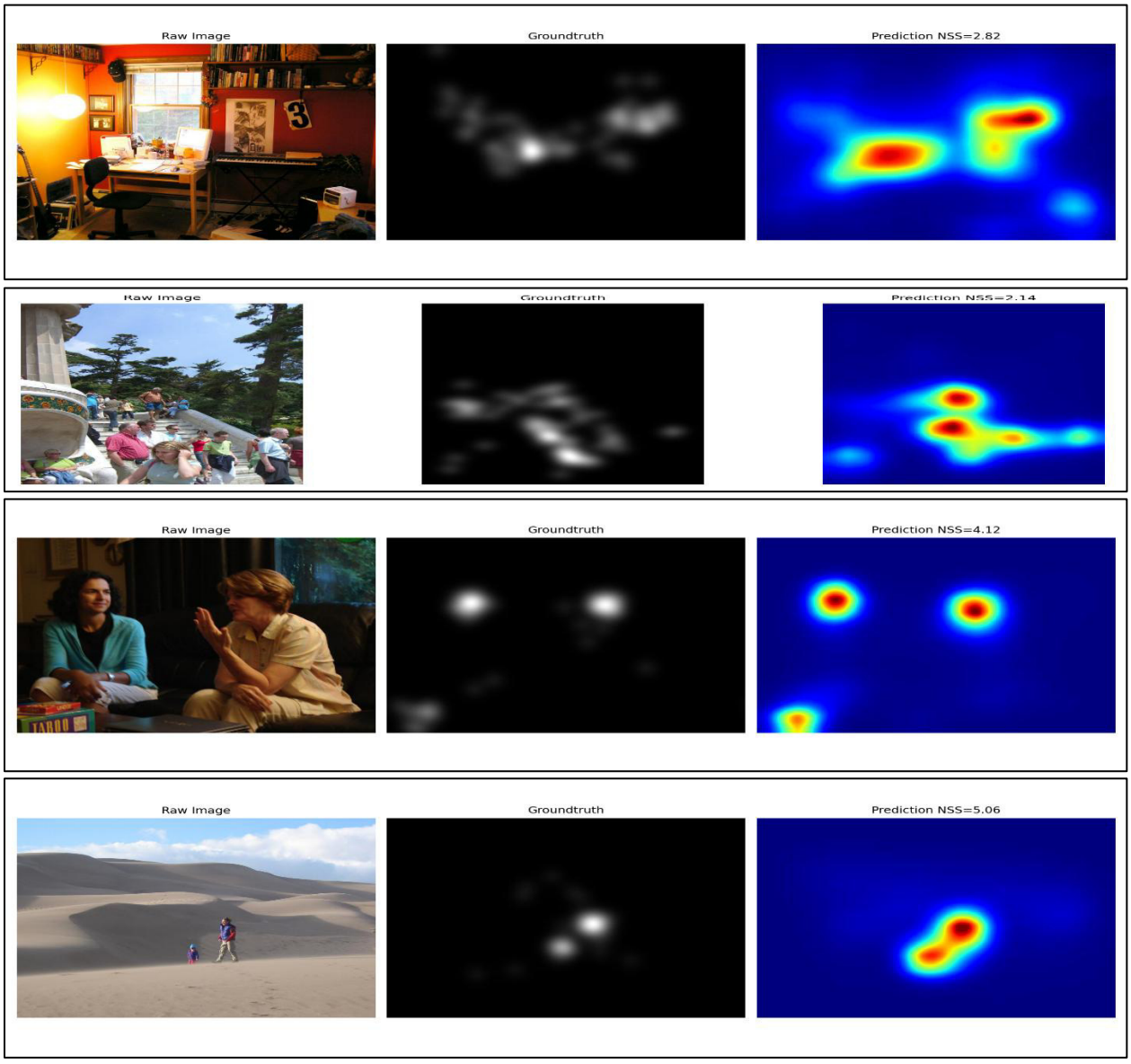

Raw Image
Groundtruth
Prediction NSS=1.38

**Figure 20: Additional Example saliency predictions and their corresponding Normalized Scanpath Saliency (NSS) scores using DeepGaze IIE. Each row shows the raw image, the ground truth saliency map, and the model's prediction with its NSS value.**

## Salicon



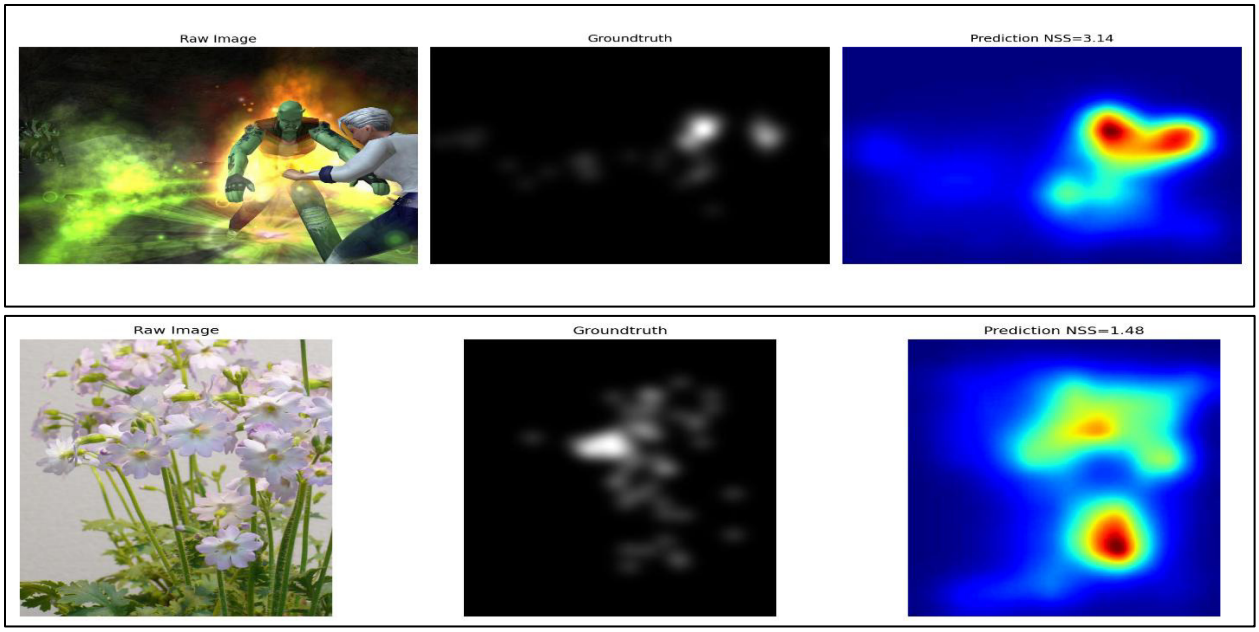
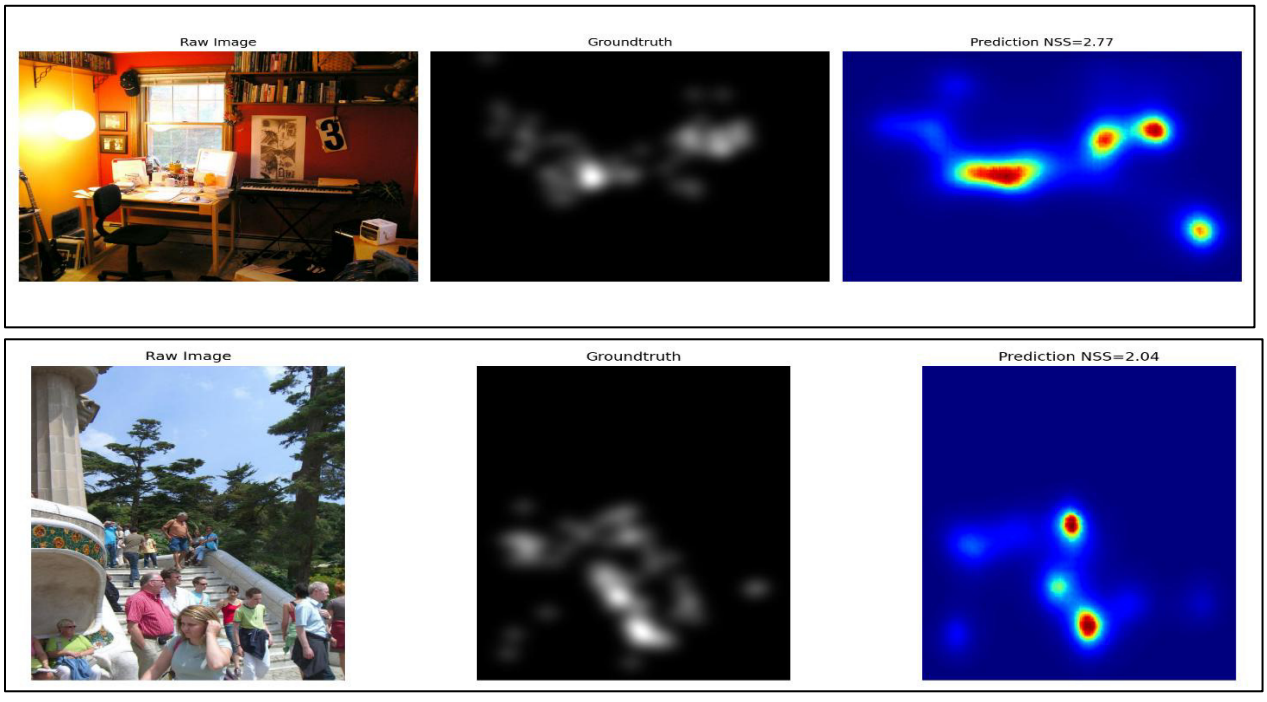



**Figure 21: Additional Example saliency predictions and their corresponding Normalized Scanpath Saliency (NSS) scores using Salicon. Each row shows the raw image, the ground truth saliency map, and the model's prediction with its NSS value.**

### Saliency Unification through Mamba (SUM)



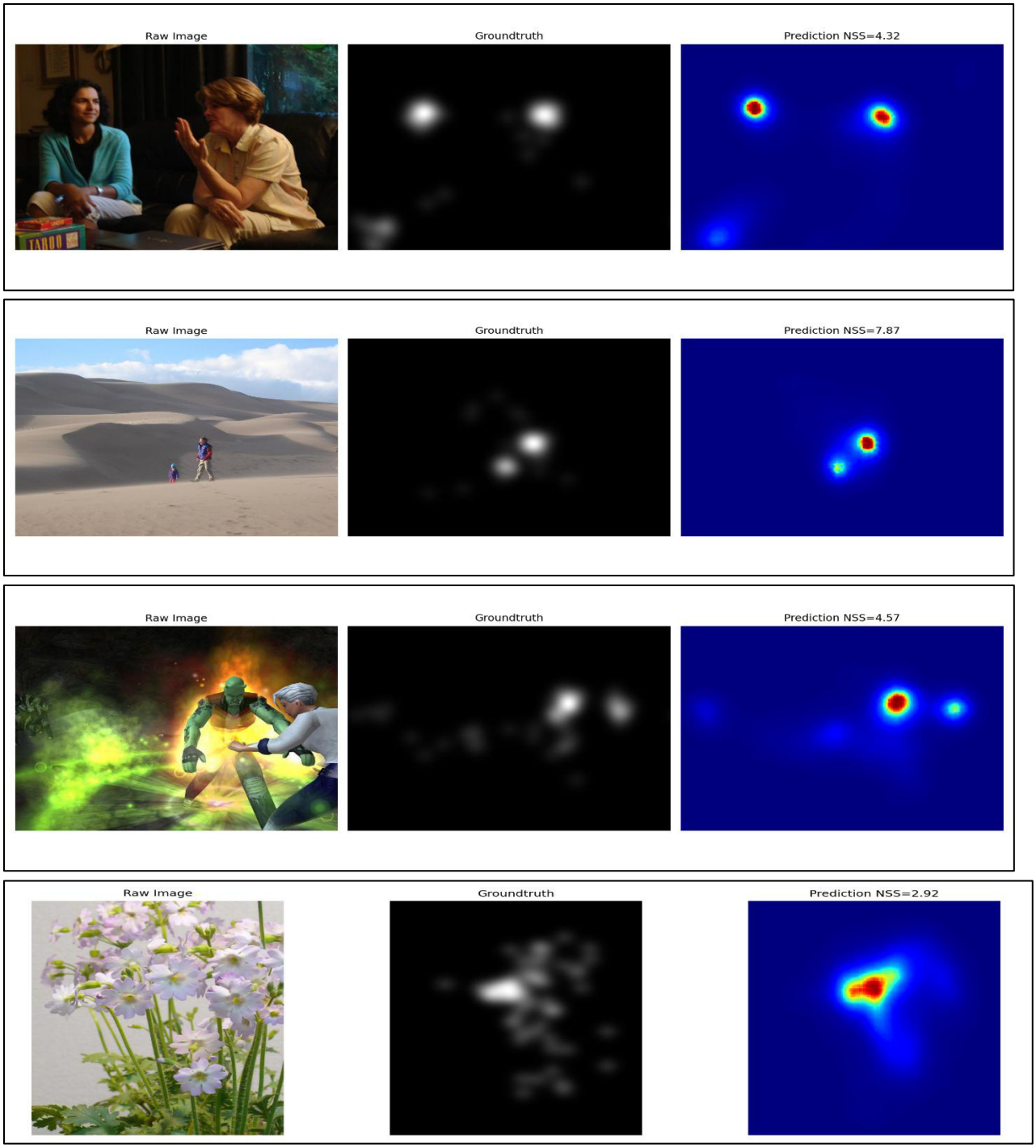
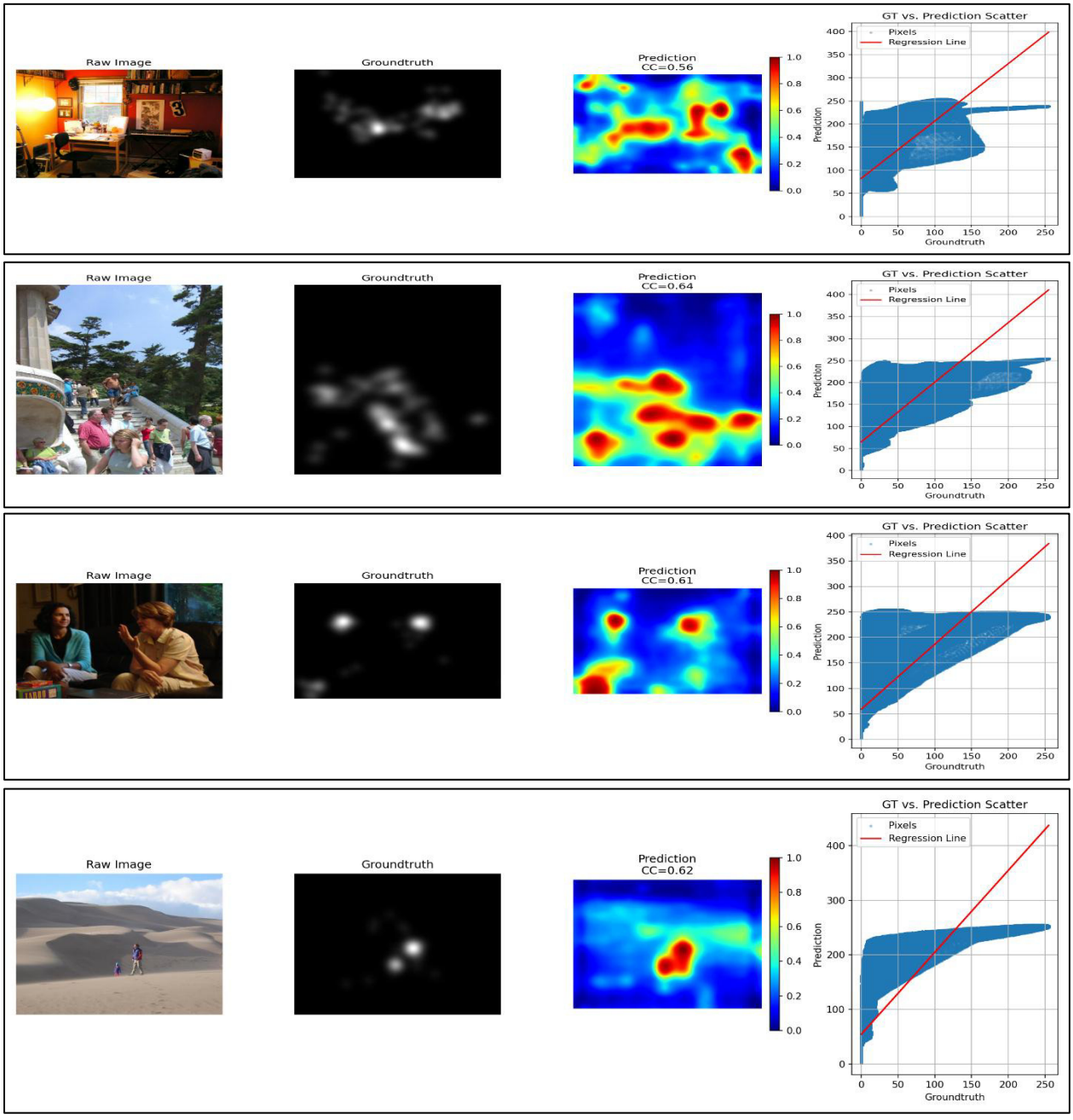



**Figure 22: Additional Example saliency predictions and their corresponding Normalized Scanpath Saliency (NSS) scores using Salicon. Each row shows the raw image, the ground truth saliency map, and the model's prediction with its NSS value.**

**Additional results of models' Performance under Pearson Correlation Coefficient (CC).**

## DeepGaze IIE



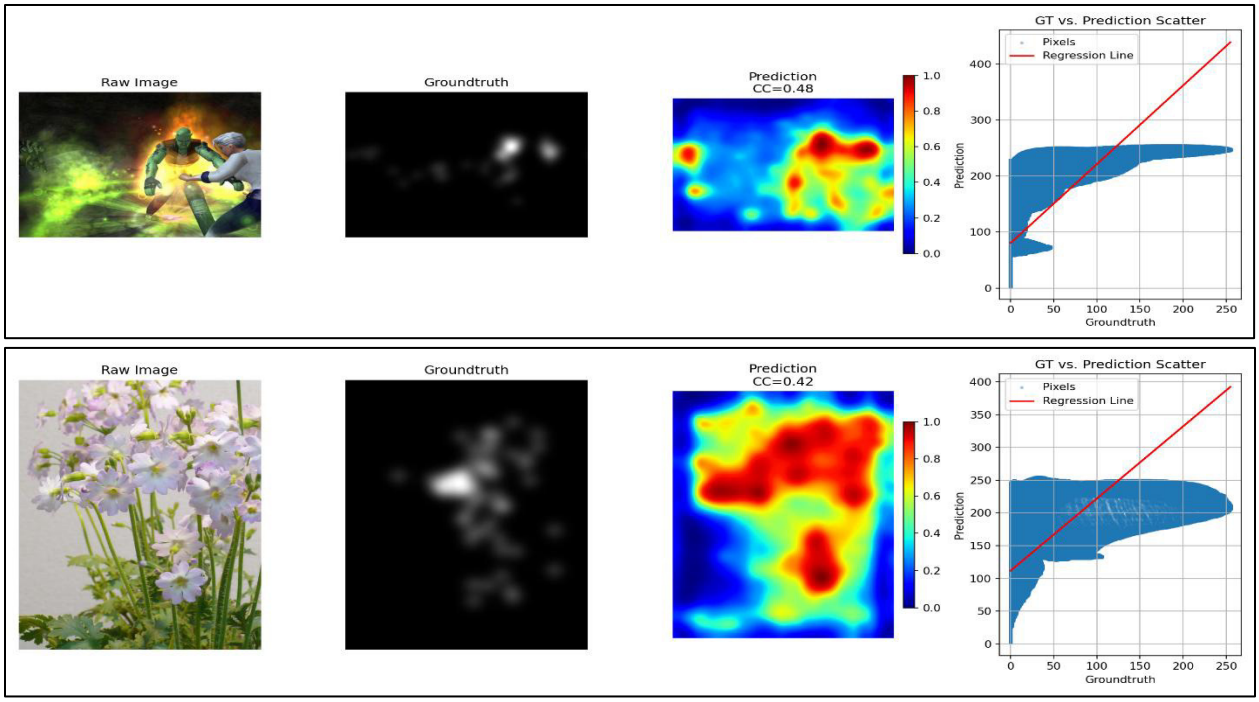
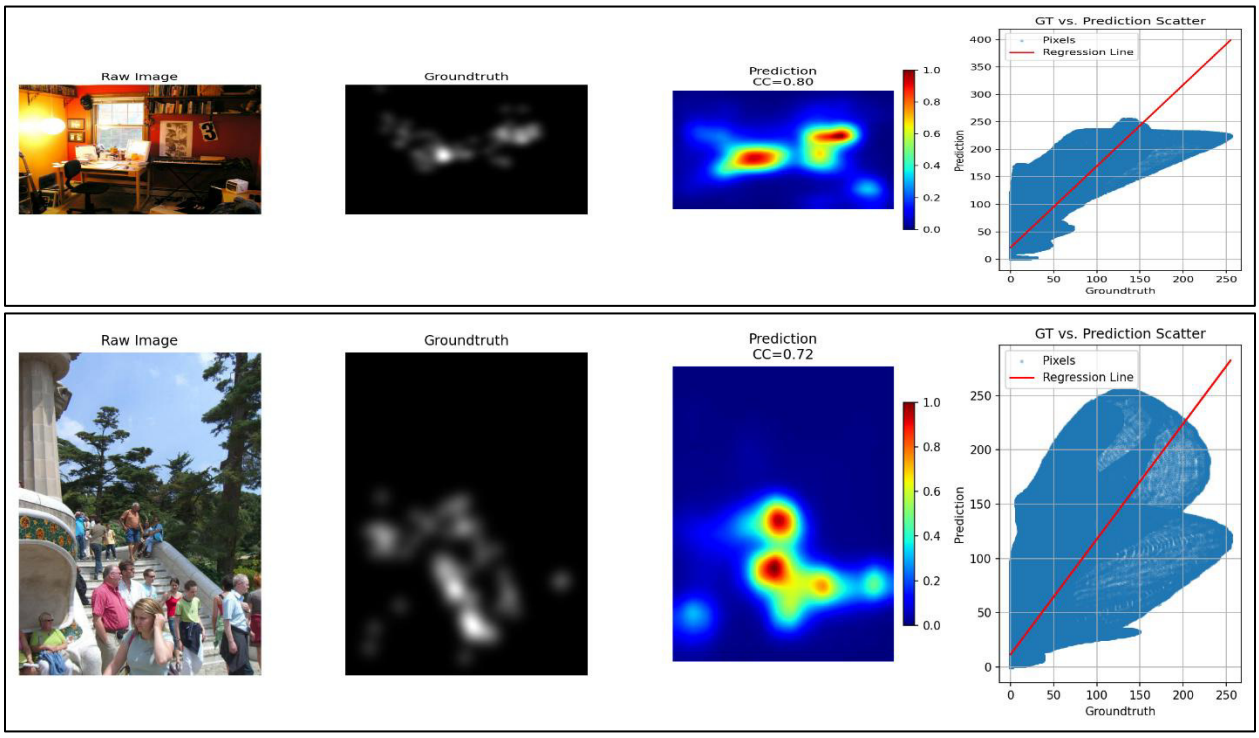



**Figure 23: Comparison of ground truth saliency maps and those predicted by DeepGaze IIE using the correlation coefficient (CC). Each row displays the raw image, the ground truth saliency map, the model's prediction, and a scatter plot comparing ground truth and predicted values, along with the corresponding CC score.**

## Salicon



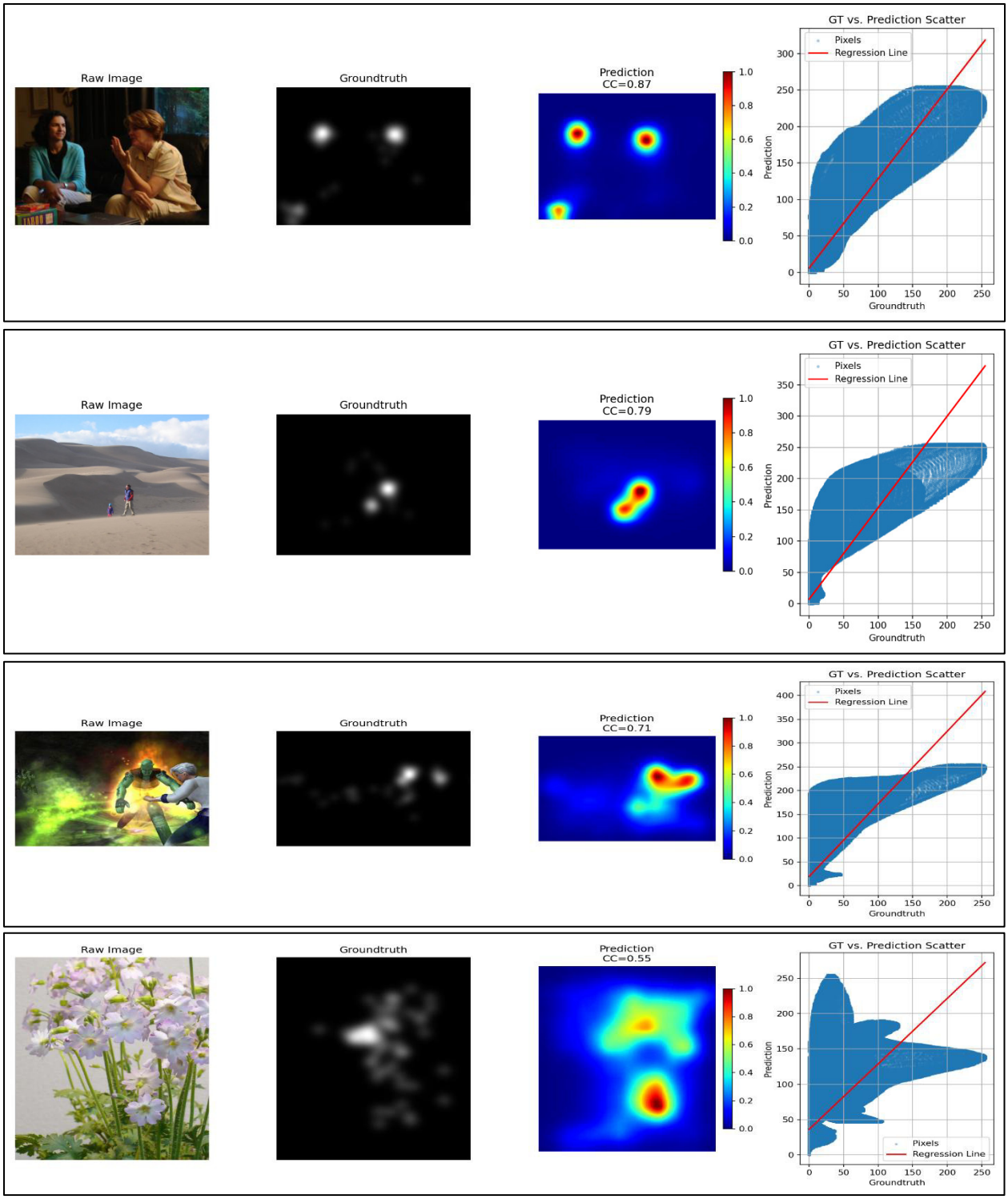
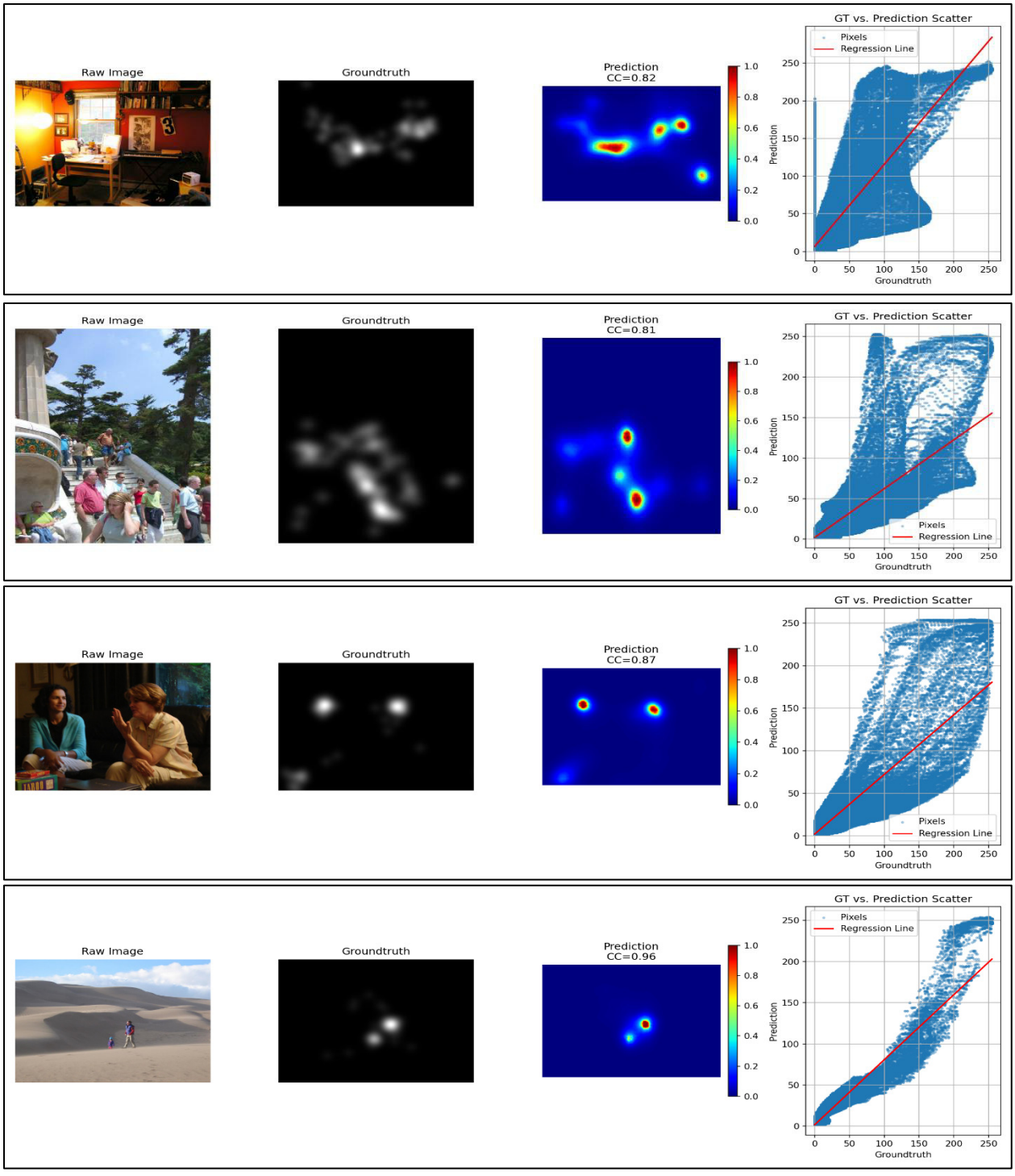



**Figure 24: Comparison of ground truth saliency maps and those predicted by SALICON using the correlation coefficient (CC). Each row displays the raw image, the ground truth saliency map, the model's prediction, and a scatter plot comparing ground truth and predicted values, along with the corresponding CC score.**

## Saliency Unification through Mamba (SUM)



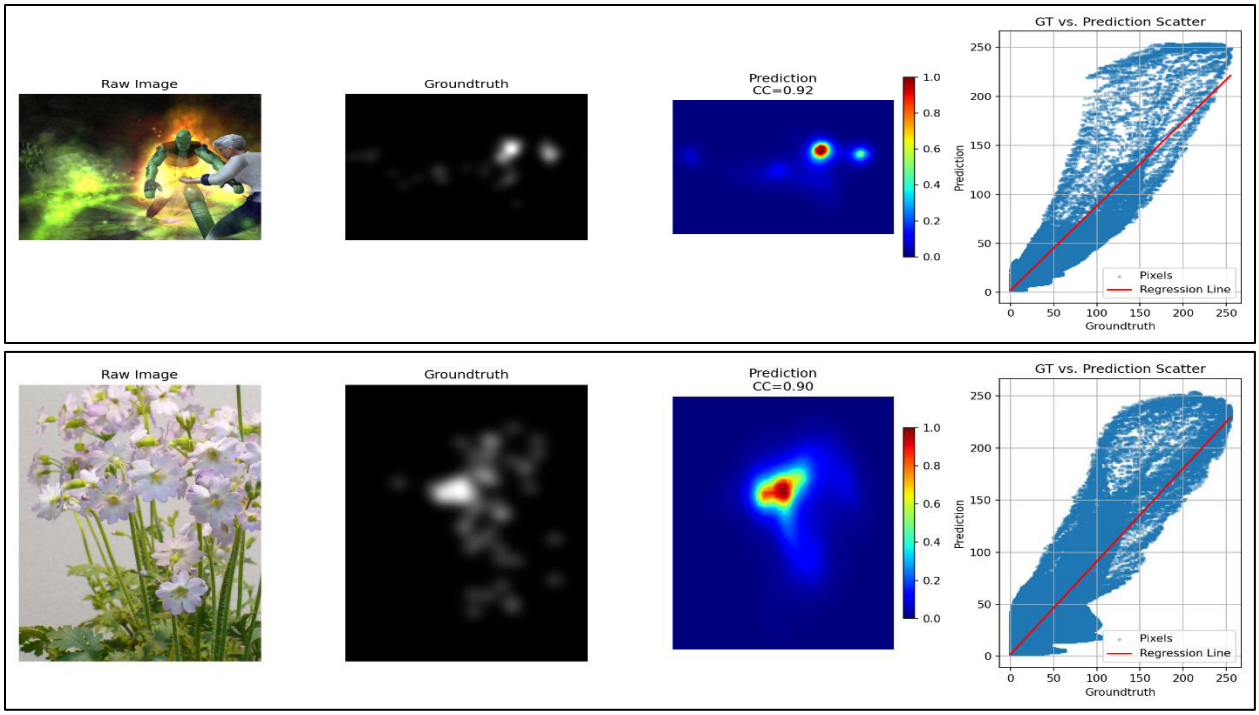



**Figure 25: Comparison of ground truth saliency maps and those predicted by SUM using the correlation coefficient (CC). Each row displays the raw image, the ground truth saliency map, the model's prediction, and a scatter plot comparing ground truth and predicted values, along with the corresponding CC score.**