“Where we look”

Approach to Predicting Visual Attention in Images using Deep Learning

Advanced AI for Data Science II

Adrián Galván Díaz
School of Engineering and Science
Instituto Tecnológico y de Estudios Superiores de Monterrey
Querétaro, México
A01704076@tec.mx

***Abstract* — Visual saliency prediction, particularly the ability to predict where humans will fixate in an image, remains a key challenge in computer vision and human-computer interaction. In this project, we present a deep learning approach that builds upon the state-of-the-art DeepGaze IIE architecture [11] by introducing a novel category-specific weighting mechanism. This model integrates four backbones—DenseNet201, ResNet50, EfficientNetB5, and ShapeNetC—each initialized with fixed seed values to ensure robust and diverse feature extraction. The model was trained and evaluated exclusively on the CAT2000 dataset, containing eye-tracking data from 120 observers across 4000 images spanning 20 categories [1].**

We propose a modified inter-model (combination of individual pretrained backbones) ensemble architecture that dynamically applies category-specific weights to saliency maps after Gaussian smoothing, normalization, and a learned center bias weight are applied. This approach optimizes the contribution of each backbone based on the image’s category, addressing challenges in saliency prediction while preserving spatial resolution and ensuring computational efficiency through a higher downsampling factor.

Experimental results demonstrate that this model effectively predicts eye fixation patterns across diverse image categories, achieving decent performance on evaluation metrics such as Information Gain (IG), and Area Under the Curve (AUC). Notably, the dynamic weighting mechanism enhances interpretability by identifying backbones that excel in specific contexts.

This work contributes to the field by leveraging category-aware weighting to improve both the accuracy and interpretability of saliency predictions. These findings provide new insights into the relationship between backbone architectures, image categories, and human visual attention, offering a robust and computationally efficient framework for advancing saliency prediction systems.

Keywords* — *Visual Saliency Prediction, Deep Learning, Probabilistic Modeling, Category-Specific Weighting, DenseNet201, ResNet50, EfficientNetB5, ShapeNetC, Saliency Maps, Gaussian Smoothing, Human-Computer Interaction, Image Attention Modeling, Log-Likelihood, Information Gain, Normalized Scanpath Saliency, CAT2000 Dataset, Downsampling, Multimodel Ensemble, Fixation Mask, Transfer Learning, Gaze Behavior Modeling, Visual Stimuli Analysis.

I. Introduction

Understanding where people naturally direct their attention within an image offers profound insights into human perception and can reshape the way industries approach design, media, and user engagement. By predicting fixation points, deep learning models enable new ways to interpret human attention and interest, creating valuable applications across a range of industries. Knowing where a person will focus their gaze can provide critical insights for marketing, where advertisements can be optimized to capture attention; in film and photography, where scenes and compositions can be crafted to guide the viewer’s experience; and in healthcare, where understanding visual attention can aid in diagnosing conditions related to cognition, eye health, and neurological disorders.

Eye-tracking devices are pivotal in this field. These devices capture eye movements, allowing researchers to record where and how long a person looks at specific elements within a visual scene. Typically, eye trackers employ infrared light and high-speed cameras to precisely

map gaze positions on a screen or environment. By studying these gaze patterns, it is possible to reveal subconscious attention behaviors, which can be highly informative for personality studies, education (to analyze focus in learning environments), and even safety research, such as tracking driver attention in autonomous vehicles. For instance, atypical gaze behaviors can indicate conditions such as autism or ADHD, while specific gaze patterns help in understanding the progression of neurodegenerative diseases like Alzheimer's [19]. In the context of deep learning, these datasets of eye-tracking data serve as ground truth for training models to predict visual saliency maps — heat maps indicating likely areas of focus.

By integrating data from eye-tracking devices into deep learning models, we can create predictive systems that simulate human gaze, empowering diverse fields to tailor visual content more effectively. This project explores these possibilities through the development of a deep learning model aimed at predicting visual saliency, showcasing the transformative potential of gaze prediction.

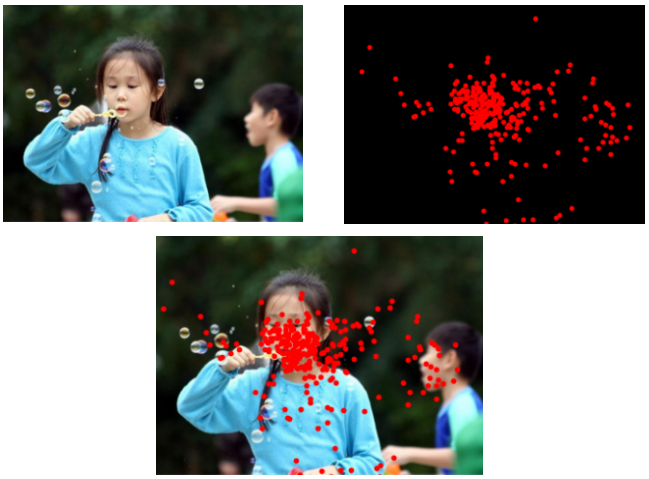


Fig. 1. This figure showcases an example image from the CAT2000 dataset [1] alongside its corresponding fixation data. The top-left image represents the original stimulus presented to observers, while the top-right visualization displays the fixation points (in red) recorded during the free-viewing task. The bottom image overlays these fixation points onto the original stimulus, illustrating where human attention was predominantly directed during the 5-second viewing window.

II. Related Work

Research in visual saliency prediction has evolved significantly [20], driven by its utility in fields like computer vision and human-computer interaction. Early models, such as Itti and Koch's [2], employed handcrafted features based on color, contrast, and orientation to approximate human attention. However, these approaches were limited in adaptability and struggled with complex image types.

The introduction of deep learning led to a paradigm shift, enabling models to automatically learn features directly from data. SalGAN [3], one of the pioneering models in this field, employs an encoder-decoder architecture based on VGG networks, with skip connections that allow the encoder's feature maps to bypass certain layers and connect directly with the decoder. This design preserves finer details from the encoding layers, which improves the model's pixel-wise saliency predictions. SalGAN introduced an innovative use of adversarial training, similar to GANs, where a discriminator is trained alongside the generator to distinguish between real and generated saliency maps. This adversarial setup, beyond conventional binary cross-entropy loss, encourages the generator to produce maps that not only align pixel-by-pixel with the ground truth but also mimic the broader distribution of human gaze, making the predictions more realistic and naturally aligned with human attention patterns [4].

DeepGaze I [5] took a simpler transfer learning approach by leveraging a pre-trained AlexNet model, which had been trained on the large-scale ImageNet dataset. Rather than retraining the model end-to-end, DeepGaze I extracts features from each layer of AlexNet and resizes these feature maps for consistency. By combining these layers into a single set of features, DeepGaze I applies logistic regression to output a final saliency map. This straightforward design, complemented by a center bias assumption (where human attention is often drawn towards the image center), proved effective and achieved competitive results on the MIT saliency benchmark [6]. This model demonstrated the feasibility of using pre-trained networks in saliency prediction without extensive additional training

Building on these insights, multi-resolution approaches like MLNet were developed to better capture attention patterns at multiple spatial levels [7]. Unlike DeepGaze I, which uses a linear combination of pre-trained features, multi-resolution models train end-to-end and use convolutional layers for richer feature learning. These models typically integrate a learned center bias directly from training data, reflecting more accurate natural gaze tendencies without imposing a hard-coded bias. Some architectures even employ multiplicative operations instead of addition to merge feature maps from different levels, capturing more complex feature interactions that enhance spatial detail and gaze accuracy. By adapting their center biases dynamically based on data distributions, these models are highly responsive to the actual patterns in human gaze data.

SALICON, another multi-scale model, incorporates this concept by employing the same network architecture at multiple resolutions [8]. SALICON combines feature maps via concatenation to refine saliency prediction across scales. It innovatively incorporates saliency evaluation metrics—such as Kullback-Leibler (KL) divergence, Normalized Scanpath Saliency (NSS), and correlation coefficient—as loss functions during training. By directly embedding these evaluation metrics, SALICON aligns its learning process with the criteria by which its performance will be judged. Although the effectiveness of this method in enhancing predictive accuracy remains an area of ongoing research, it represents a novel attempt to integrate evaluation criteria directly into the model's optimization process.

DeepFix, one of the more advanced models, modifies the VGG-16 architecture to improve spatial context handling [9]. By removing the max pooling layers typically found in VGG-16, DeepFix avoids excessive downsampling, retaining essential spatial information. In place of max pooling, it incorporates dilated convolutions, which expand the receptive field without sacrificing resolution, a key feature for saliency prediction tasks that demand both local detail and a broad field of view. Furthermore, DeepFix introduces "location-based convolutions," which provide explicit spatial cues by adding location images to the convolutional layers. This innovation enables the model to recognize important regions, ignoring the center, where attention is often focused. These adaptations allow DeepFix to maintain the integrity of learned VGG-16 features while enhancing the model's sensitivity to spatial layout and context.

More recent developments have explored domain-specific augmentations, such as frequency-domain information, to enhance model performance. For example, a model presented at AAAI 2019 named Sal-DCNN introduced Fourier-transformed inputs alongside the standard RGB channels [10]. By adding frequency-based features, this model captures patterns that might be less apparent in spatial data alone, potentially enhancing saliency prediction in scenes with complex textures or repetitive elements. In parallel, complex convolutions with non-linear activations have been utilized to exploit these Fourier features, yielding deeper representations that respond to high-frequency variations and patterns in the data.

The progression in visual saliency prediction has shown a clear shift from basic handcrafted feature models to complex deep learning architectures, each iteration aiming for greater precision and adaptability. As the field evolved, a trend emerged toward models capable of handling spatial detail while capturing broad contextual information. Techniques like multi-resolution architectures and location-based convolutions helped preserve fine details and adapt to natural gaze biases. However, these advancements often require high computational resources, extensive training data, and considerable processing power, which can be a barrier to practical deployment, especially in limited-resource environments.

DeepGaze IIE marks a significant advancement in visual saliency prediction by employing a probabilistic framework for both in-domain and out-of-domain predictions of human gaze patterns [11]. Building on earlier iterations like DeepGaze I and II, DeepGaze IIE introduces several novel techniques that set it apart as a state-of-the-art model for saliency prediction.

This model distinguishes itself with a probabilistic framework, prioritizing Information Gain (IG) as a principal evaluation metric. Unlike conventional saliency metrics, IG measures how much predictive utility DeepGaze IIE adds over a central baseline model, offering a more interpretable assessment of performance. To enhance predictive robustness, it uses an ensemble of multiple backbones, including DenseNet-201, EfficientNet-B5, ResNet-50, and ShapeNet-C, each configured with different random initializations (seed values). The authors conducted extensive experiments [11] to determine the most effective layers for feature extraction, selecting the final three to four layers within each backbone. These layers were found to focus on high-level, abstract features that are critical for predicting attention. Each backbone connects to a dedicated readout network that processes these selected features via 1x1 convolutions, normalization, and softplus activations. These readout layers produce intermediate saliency maps for each backbone instance, which are subsequently averaged to create a final, unified saliency map. For training, DeepGaze IIE was pretrained on the SALICON dataset to capture general saliency patterns and then fine-tuned on MIT1003 [12] using a 10-fold cross-validation scheme, reinforcing its adaptability across diverse types of visual stimuli. Last but not least, DeepGaze IIE also incorporates a trainable center bias weight that adjusts to reflect natural human tendencies to focus on central areas within an image [11].

III. Dataset

For this project, we employ exclusively the CAT2000 dataset for training and the evaluation set. CAT2000 is a large-scale dataset curated to foster innovation in saliency modeling by offering a diverse selection of image categories and high-quality eye-tracking data. Developed by Ali Borji and Laurent Itti and presented at the CVPR 2015 workshop "Future of Datasets," CAT2000 includes 4000 images across 20 categories (e.g., cartoons, art, low-resolution images, indoor and outdoor scenes, and line drawings), with each category containing 200 images [1]. In this case I'm only using 1280 images for training and evaluation since the test folder doesn't have fixation locs by image. This diversity challenges models to adapt to various visual contexts and aids in avoiding overfitting to specific image types.

CAT2000 images have a resolution of 1920x1080 pixels, with each pixel approximating 1 degree of visual angle (DVA) ≈ 38 pixels. The dataset includes eye fixation data from 120 observers, with each image viewed for 5 seconds. In contrast to SALICON, which uses mouse-traced approximations of human attention, CAT2000 provides precise fixation data recorded at 1000 Hz using the EyeLink1000 eye-tracker [21]. This focus on real fixations ensures a more accurate representation of human attention patterns, offering an advantage for training saliency models that aim for precise fixation prediction.

IV. Methodology

DeepGaze IIE currently stands as the top-performing model on the MIT/Tuebingen Saliency Benchmark [13], achieving these results by intelligently combining multiple backbones through transfer learning. Inspired by the success of DeepGaze IIE and its innovative ensemble-based architecture, this project introduces a refined approach to leveraging category-specific information for visual saliency prediction. The authors emphasize that different backbones excel at capturing distinct features of images—shapes, textures, or compositions.

Building on this foundation, we propose a modification that dynamically assigns category-specific weights to the saliency maps generated by each backbone before they are combined. This weighted combination is designed to amplify the contribution of the most effective backbones for each image category, enhancing predictive performance. Our approach aims to push gaze prediction boundaries further by testing whether assigning greater weight to image categories could yield even better results.

Moreover, recognizing that each backbone inherently specializes in extracting unique features from an image, my methodology seeks to enhance this strength by introducing the weight matrix specifically tailored for certain situations. It is intuitive to assume that human gaze behavior varies significantly depending on the type of stimulus encountered—for example, the way we look at a natural landscape differs from how we explore an abstract painting. By replicating this behavior, the proposed model adds a weight matrix trained specifically for distinct categories or scenarios. This matrix dynamically adjusts the importance of each backbone's saliency map in response to the visual context, aiming to mimic human-like variations in gaze patterns and further refine the model's predictions.

A. Model Architecture

The architecture of the proposed model builds upon the core principles of DeepGaze IIE [11], retaining its ensemble-based approach while introducing significant modifications to accommodate computational constraints and enhance category-specific predictions. The model consists of four non-trainable backbones—DenseNet201, ResNet50, EfficientNetB5, and ShapeNetC—each optimized to extract complementary features from input images (see Figure 2). These backbones are pre trained on ImageNet and process the downsampled input images to produce feature maps.

The proposed architecture begins with the processing of input images and a center bias model, which are passed through all backbones to extract a diverse range of features. Humans naturally tend to fixate near the center of an image (see Figure 1), a tendency that makes predicting the center a relatively easy task for any visual saliency model. However, in a framework focused exclusively on saliency prediction, the goal is not simply to identify the center but to replicate human behavior in visual attention. Predicting where a person is most likely to look is fundamentally different from predicting the precise points where they will fixate during a specific timeframe. The center bias model plays a crucial role here by guiding the model to extract features from regions where people are likely to look, without confining predictions solely to fixation points observed within 5 seconds.

The gradient flow adjusts the center bias weight dynamically. For example, if the weight is too high, predictions may overly depend on the center bias, causing a large loss when fixation points deviate from the center. The gradient will reduce the weight to emphasize image-specific features instead. Conversely, if the weight is too low, the model may ignore the natural human tendency to fixate near the center, increasing the loss when fixations align with the center. In such cases, the gradient increases the weight, leveraging the center bias more effectively.

One of the significant innovations by the authors in DeepGazeIIE was the initialization of three random seed values for each backbone (see Figure 3, readout and finalizer network). These seeds influence the initial weights of the backbones during training, ensuring diversity in the feature extraction process. By initializing each backbone with distinct seed values, we create multiple instances of the same backbone architecture, each with unique learned representations of the input data.

Each backbone generates an initial saliency map through the readout network and finalizer network all together (see Figure 3). The readout network transforms the high-dimensional feature maps generated by each backbone into intermediate saliency maps. The initial convolution reduces the feature space to 8 channels, applying normalization and softplus activation. Subsequent convolutions iteratively refine the representation into 16, 128, and eventually a 1-channel saliency map (see Figure 2, top panel). This structure not only reduces dimensionality but also ensures that critical spatial details are preserved, enabling each backbone to specialize in capturing unique visual patterns.

These maps are further refined in the finalizer. The finalizer applies post-processing steps to the saliency maps generated by the readout network (see Figure 3, readout network). It begins by taking the normalized output from the readout network and applying a Gaussian smoothing filter. This step removes small inconsistencies or noise in the map, enhancing its clarity and interpretability by emphasizing broader saliency regions. Following this, the trainable center bias weight is added, reflecting the natural human tendency to fixate near the center of an image. This bias is represented as a spatially distributed prior, learned during training, which adjusts the saliency map to align with typical gaze tendencies observed in human visual attention data.

After the smoothing and biasing steps, the saliency map undergoes an upscaling operation to restore its resolution to match the original input dimensions. This upscaling is performed using bicubic interpolation, which ensures that the map retains spatial consistency without introducing artifacts. The final step in the process is normalization, where the map is converted into a probabilistic distribution by dividing all pixel values by their sum. This normalization ensures that the saliency map can be interpreted as a probability density function, where each pixel's value represents the likelihood of human fixation at that location.

At a high level, each backbone (initialized with three random seed values and processed through 5 cross-validation folds) generates a total of 15 individual saliency maps. These maps are combined internally within each backbone complete network to produce a single combined saliency map per backbone.

After each backbone produces its combined saliency map, a category-specific weighting matrix is applied to dynamically adjust the contribution of each backbone based on the image category. The negative log-likelihood (NLL) loss [14], which serves as the loss function, is calculated directly on the final saliency map by comparing it to the ground truth fixation mask. This ensures that the loss reflects the accuracy of the entire model pipeline, including the combined contribution of all backbones. The NLL loss plays a crucial role in optimizing the model by simultaneously updating the center bias weight, category-specific weights (fine-tuning the contribution of each backbone for different image categories), adjusting the readout networks within each backbone instance to ensure that extracted features are optimally transformed into saliency maps, and refining the finalizers to improve the normalization and smoothing processes.

B. Pipeline

The flow of information in the model begins with an input image of size 1920×1080 which undergoes a downsampling process and a predefined center bias model. The downsampled image is then processed by the four backbones operating in parallel. Once the backbones generate their feature maps, these are passed to the readout network, where they are transformed into intermediate saliency maps. These maps are subsequently refined by the finalizer. The resulting maps, now resized back to the original image dimensions, are then subject to category-specific weighting. At this stage, the model dynamically adjusts the center bias weight and the contribution of each backbone's saliency map based on the image category. For example, weights are assigned differently for categories such as “Art” or “Outdoor” to prioritize backbones that excel in those contexts. The weighted saliency maps are finally combined into a single probabilistic map that represents the model’s optimized prediction of visual attention.

C. Adaptations

The implementation of these modifications required significant adaptations to the original DeepGaze IIE framework to align with computational constraints and the new weighting mechanism. The first major adaptation involved increasing the downsampling factor from 2 (used in DeepGaze IIE) to 8.5. Our input images had a 1920x1080 size, the downsampling factor and the readout network reduced the image to approximately 14x8 pixels (as can be further observed in the predictions) which dramatically decreased computational demands but risked spatial detail for saliency prediction. To account for the reduced resolution, the Gaussian smoothing step and upscaling process in the finalizer were fine-tuned to ensure that the saliency maps retained clarity and interpretability.

Unlike DeepGaze IIE, which uses SALICON for pretraining, MIT1003 for fine tuning and 10 cross-validation folds, our model was trained exclusively on CAT2000 with a 5-fold cross-validation to ensure exposure to all image categories, allowing the model to generalize across different visual contexts. This approach aims to maximize saliency map prediction metrics without relying on the pre-training step with SALICON, thus maintaining computational efficiency and focusing on high-quality eye-tracking data

With these changes comes the second critical adaptation which is the introduction of the category-specific weighting mechanism. This required extracting each image’s category in the dataset (e.g., “Action,” “Cartoon,” “Art”). During training, the learnable weight matrix was integrated into the model, dynamically adjusting the importance of each backbone's saliency map based on the image category.

Another adaptation involved addressing the center bias model. The original DeepGaze IIE implementation leverages a center bias model fine-tuned on the MIT1003 dataset, which is specifically tailored to their training data. However, due to dataset differences and computational constraints, we implemented a simpler Gaussian center bias model. Unfortunately, we were unable to replicate the original center bias model, and our hypothesis is that this limitation could potentially impact our results, as the Gaussian center bias model lacks specialization for our CAT2000 training data. To minimize variability and maintain consistency throughout the project, we decided to keep the Gaussian center bias model fixed during the entire development process. This approach allowed us to isolate and evaluate the performance improvements introduced by the modified DeepGaze IIE model without introducing additional sources of variation.

Just as DeepGazeIIE the final loss was computed as the negative log-likelihood of the combined saliency map relative to the fixation mask but it was modified to accommodate the weighted combination of saliency maps. This adjustment ensured that the model learned to optimize the contribution of each backbone to the overall prediction. Finally, the evaluation metrics, including Log-Likelihood (LL), Information Gain (IG), and NSS [14], were recalculated based on the weighted saliency maps during training to reflect the overall performance of the combined model rather than individual backbones.

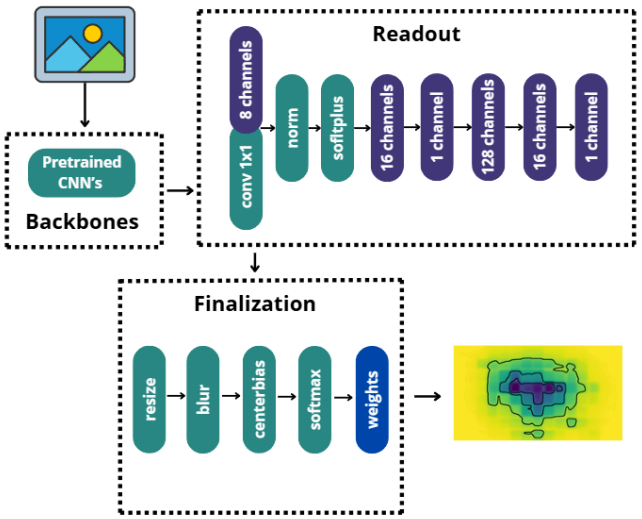


Fig. 2. Illustration of the internal processing steps of the readout network and the finalizer. The readout network processes high-dimensional feature maps generated by the backbones through a series of convolutional layers. The finalizer refines the saliency maps using Gaussian smoothing, applies a trainable center bias to incorporate human fixation tendencies, and normalizes the output into a probabilistic saliency map.

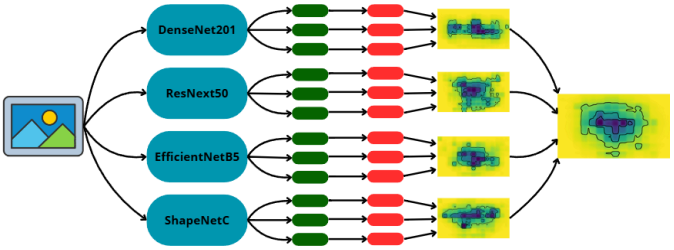


Fig. 3. Complete model pipeline with input parallel processing by multiple backbones (DenseNet201, ResNet50, EfficientNetB5, and ShapeNetC). Each backbone generates a saliency map, which is refined through the readout network (green) and finalizer (red). The resulting maps are dynamically weighted based on the input image category, leveraging a learned weight matrix. The weighted saliency maps are combined into a single probabilistic output.

D. Training

As mentioned before, the model was trained and evaluated using the CAT2000 dataset provided by the MIT Saliency Benchmark. To establish a robust training and evaluation pipeline, the dataset was partitioned as follows:

Evaluation Set: For each category (20 categories), 20 random images were selected from the available 100 images per category. This resulted in a total of 400 evaluation images. The evaluation metrics were calculated using an adaptation of the PySaliency library [17].

Training Set: From each category, 64 images were used for training, leading to a total of 1280 training images across all categories.

Validation Set: The remaining 16 images per category were allocated for validation, resulting in a total of 320 validation images.

No augmentation techniques were applied to the training images or fixation masks, as the problem does not necessitate such transformations. Augmenting fixation masks could introduce unintended distortions or biases, potentially degrading the quality of saliency predictions. An advantage of the CAT2000 dataset is its inclusion of diverse categories such as "BlackWhite," "Cartoon," "Fractal," "Inverted," "Jumbled," "Line drawing," "Low Resolution," "Noisy," and "Sketch." These inherently varied image types provide sufficient diversity in the training data, reducing the need for additional augmentations (see Figure 4).

Fig. 4. This figure illustrates sample images from four distinct categories within the CAT2000 dataset, highlighting the dataset's inherent diversity: Jumbled (Top Left), Inverted (Top Right), Low Resolution (Bottom Left) and Noisy (Bottom Right).

The training process employed the ReduceLROnPlateau learning rate scheduler [18], which dynamically adjusts the learning rate based on the validation loss. This scheduler helps prevent overtraining by reducing the learning rate when the validation loss plateaus, eventually stopping the training process when no significant improvements are observed. The scheduler was designed to handle two distinct parameter groups within the optimizer (group 0 and 1), each with different initial learning rates and reduction strategies. Group 0 includes the foundational parameters of the model, such as those in the readouts and finalizers. These parameters were initialized with a lower learning rate 1e-4 with a minimum threshold of 1e-6 to allow for fine adjustments without destabilizing the learned features. Group 1 encompasses the dynamically learned weights that adjust the contribution of each backbone based on image categories. To facilitate faster convergence for these weights, their initial learning rate was set to be 10 times higher than that of the base parameters. The training process stops when group 0 reaches the minimum threshold.

Due to computational constraints, a batch size of 2 was used for training. While this is lower than typical values in deep learning, it was sufficient given the model's architecture and the dataset size.

During training, the model's progress and state were periodically saved to ensure reproducibility and recovery in case of interruptions. The category-specific weight matrices were saved as .pyc files for each category. These matrices capture the dynamic weighting adjustments made during training. The final trained model, including all backbones, readouts, and the finalizer, was saved as a .pth file. This file contains the complete state of the model, ready for evaluation and deployment.

V. Iterations

Throughout the development of our saliency prediction model, we conducted three major iterations, each building on the previous one to address identified shortcomings and explore potential improvements. This section outlines the iterative process, detailing the modifications implemented, their impact on visualizations, and the corresponding metric evaluations.

A. First Iteration: Initial Baseline Model

The first model replicated the architecture of DeepGazeIIE with minimal modifications. The key changes included training on the CAT2000 dataset instead of SALICON or MIT1003 and using a fixed Gaussian center bias. As mentioned before, unlike the original DeepGazeIIE, we could not successfully implement their official center bias model from MIT1003. The aggressive downsampling factor of 8.5 compounded a major issue, reducing the image size to 226×127, with further reductions through the readout factor to approximately 14×8. This limitation, coupled with a lack of an upsampling method, resulted in acceptable metrics but visually poor saliency maps (see Figure 5 top row and Table 1).

B. Second Iteration: Weighted Category-Specific Classification

In the second iteration, we retained the baseline setup but incorporated a category-specific weighting matrix. This addition introduced clear improvements, particularly in the alignment of visual saliency maps with the underlying image categories. While the lack of optimized upsampling still hampered visual clarity, significant differences were observed between visualizations generated with and without category classifications (see Figure 5 first and second row).

We also evaluated visualizations with correctly and incorrectly assigned categories to analyze the model's response to classification errors. As expected, saliency predictions improved noticeably when the correct category was provided, suggesting that accurate classification is essential for robust predictions. With the wrong category, the dominant backbone cannot specialize in relevant areas of activation and instead shows a tendency toward central fixation, diluting the model's effectiveness (see Figure 5 second and third row).

C. Third Iteration: Enhanced Upsampling

The third and final iteration introduced several refinements to address the limitations of the previous versions and kept the weighted specific classification:

1. Progressive Upsampling and Smoothing: To mitigate the artifacts caused by aggressive downsampling, we implemented a progressive upsampling method. The saliency maps were scaled up in steps, using bicubic interpolation at each stage:

From 8x15 → 16x30 → 64x120 → 512x960 → 1080x1920 to match the original image size.

The center bias was upscaled separately using bilinear interpolation, as it required less computational precision and complexity. This process smoothed the interpolations, producing more natural saliency maps with reduced noise (see Figure 5 fourth row).

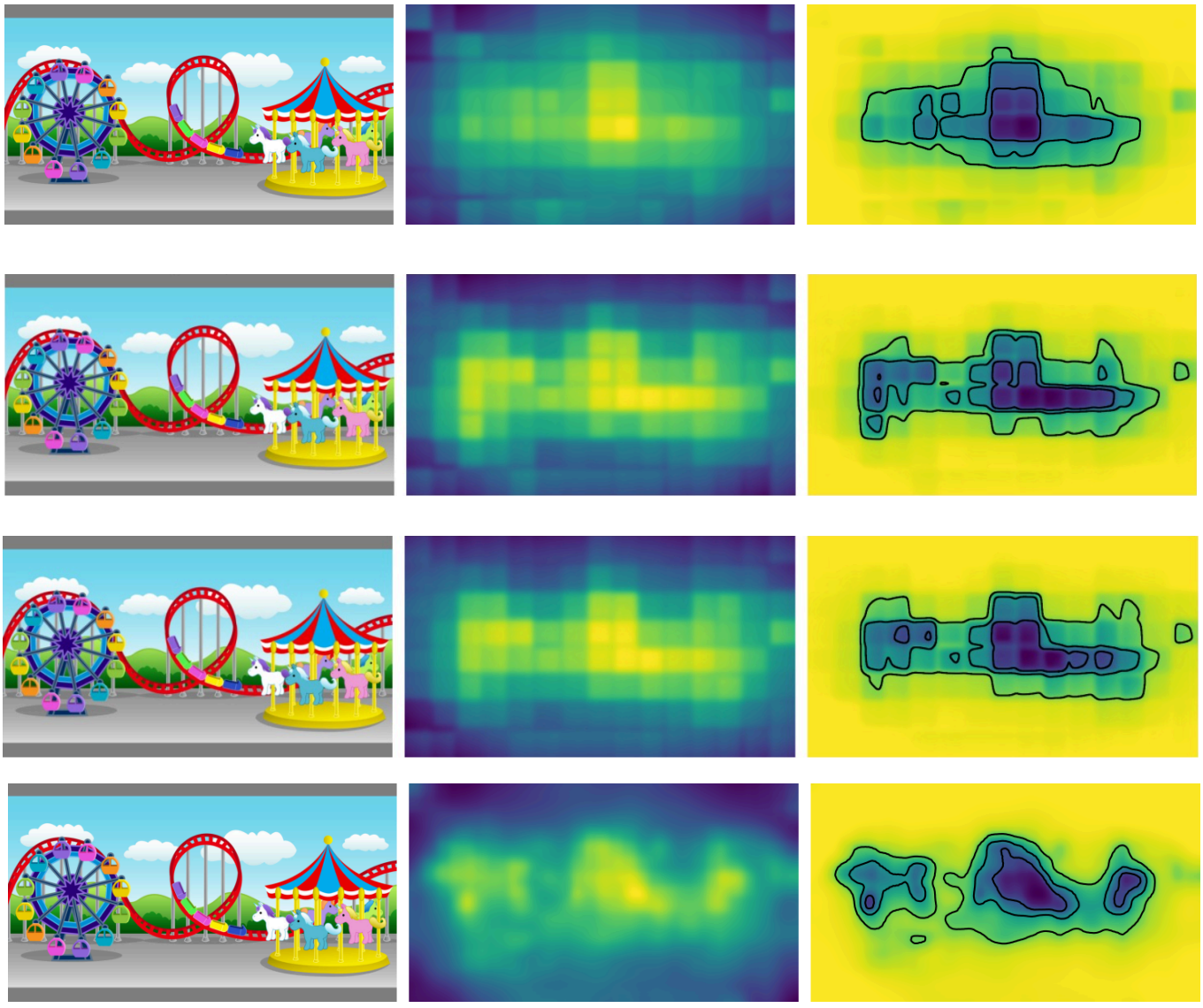
2. Backbone Initialization with Random Seeds: This iteration introduced three independent instances for each backbone, initialized with different random seeds. Unlike the previous versions where backbones shared the same initialization, this modification enabled diverse feature extraction and improved training stability. The result was a more efficient learning process and saliency maps that captured a broader range of visual cues across categories.

3. Category-Specific Weight Optimization: We retained the classification system but optimized it with the upsampled maps and Gaussian smoothing, leading to clearer predictions. This improvement is evident when comparing visualizations across the three models, particularly in category-specific maps, as seen in (see Figure 5 fourth and fifth row).

TABLE 1. ITERATIVE METRICS TABLE

Model	Loss	IG	LL	AUC	NSS	Epochs
1	13.5	3.07	-13.56	0.868	2.7	29
2	4428.47	-4478.56	-4495.20	0.865	2.02	37
3	4394.64	-4502.75	-4519.39	0.866	2.03	48

This table summarizes the performance metrics for the three iterations of the model, highlighting the progression in performance across different configurations. The metrics used for comparison include Loss, Information Gain (IG), Log Likelihood (LL), Area Under the Curve (AUC), and Normalized Scanpath Saliency (NSS). Additionally, the number of epochs each model trained is also listed. Although the metrics such as Loss, IG, and LL are not directly comparable to the first iteration due to the use of category-specific evaluation, this model showed promising improvements in guiding the saliency maps based on image categories as shown in Figure 5. AUC and NSS remained relatively stable, as these metrics are comparable across iterations. It is important to note that while AUC and NSS are comparable across iterations, they primarily served as indicators of validation performance during training. Definitive conclusions about model performance can only be drawn after obtaining benchmark metrics on saliency maps. These metrics were monitored for all iterations to ensure the models improved progressively throughout training. Furthermore, as model complexity increased, training time also extended, with the final iteration requiring significantly more epochs to converge due to the additional components and optimizations.



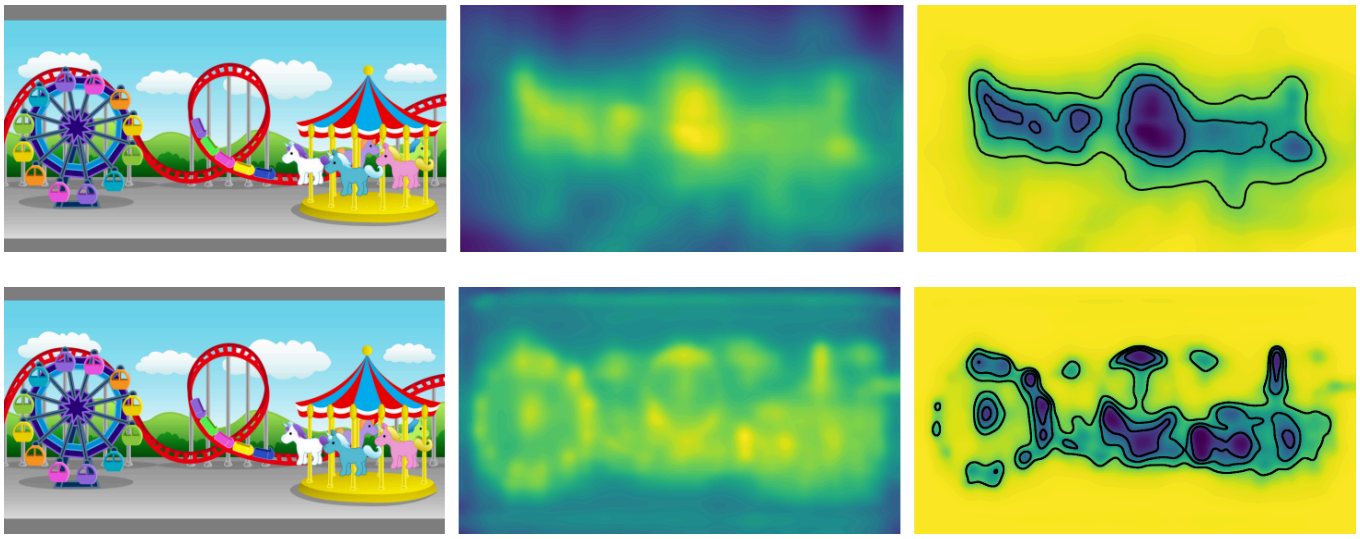


Fig. 5. For the purposes of this report and for illustrative clarity, we present the changes in saliency map predictions using a single image belonging to the Cartoon category, where the ResNext50 backbone was dominant in performance. Additional visualizations for other categories and scenarios can be found in the Github repository in appendices.

This figure illustrates the saliency map predictions across the different iterations of the model, comparing the visual outputs and the log density activations for each case. Each row corresponds to a specific configuration or iteration the model:

1. First Iteration (Top Row): This baseline model replicates the original DeepGazeIIE architecture, with a fixed Gaussian center bias. The lack of upsampling and the aggressive downsampling result in blurred and low-resolution saliency predictions, leading to limited interpretability in the final maps.
2. Second Iteration (Second Row): This iteration incorporates category-specific weights and uses the correct category classification for the input image. The log density activations and predictions show notable improvements, with clearer and more contextually accurate saliency maps compared to the baseline.
3. Second Iteration with Incorrect Category (Third Row): Using an incorrect category classification for the image demonstrates the sensitivity of the category-specific weights. The saliency maps show misaligned activations and a more center biased prediction, highlighting the importance of accurate categorization in this approach.
4. Final Iteration with Correct Category (Fourth Row): The final model includes progressive upsampling, Gaussian smoothing, category-specific weights and the initialization of random seeds. The saliency maps exhibit high resolution, reduced noise, and improved focus on the salient regions, closely aligning with human fixation patterns.
5. Final Iteration with Incorrect Category (Fifth Row): When provided with incorrect categories, even the refined final model shows reduced accuracy in saliency maps. Although improved compared to the second iteration's incorrect category case, the maps are less precise, indicating that the model still benefits significantly from accurate categorization; this can be perceived in the log density map.
6. DeepGazeIIE (Bottom Row): The original DeepGazeIIE predictions are included for comparison. These maps demonstrate strong performance in saliency prediction but lack the category-specific refinements introduced in our iterations.

Each column in the figure represents a different stage of the saliency prediction process: the leftmost column shows the input image, the middle column displays the log density activations, and the rightmost column illustrates the final saliency map predictions. This visualization highlights the progression in prediction quality and interpretability across the iterations, culminating in a model that offers superior saliency map clarity and accuracy than prior models.

VI. Adaptive Weighted Category Matrix Results

The results from the second & third iteration were promising, demonstrating the effectiveness of adaptive category-specific weighting, random seeds activation and upsampling enhanced process as seen above. Over the course of training, the model exhibited distinct patterns in weight adjustments for the backbones, reflecting their performance for specific categories. These patterns reveal that the model dynamically optimizes backbone contributions based on the category of the image. To illustrate these dynamics, we provide several visualizations that showcase the adaptability of the weights.

Fig. 6. This figure illustrates representative images from three specific categories used in the weight progression analysis:

1. Top Left: This image belongs to the Black and White category, characterized by its grayscale composition, which challenges the model to extract features without relying on color information.
2. Top Right: This image from the Low Resolution category demonstrates the challenges associated with blurry or pixelated visuals, where the model must rely on structural and coarse features to make predictions.
3. Bottom: This image belongs to the Affective category, designed to capture emotional or expressive visual stimuli, which require the model to interpret more abstract and compositional elements.

We present the progression of weights for three selected categories—Black and White, Low Resolution, and Affective (see Figure 6)—across the training epochs for each backbone. These graphs highlight how the model adjusts weights for specific categories in individual backbones, reflecting the strengths and weaknesses of each backbone for different visual contexts. For instance, certain backbones may exhibit higher weights for categories like Black and White while others dominate in categories such as Affective. These insights emphasize the nuanced decision-making of the model during optimization (see Figure 8).

In addition to observing weight progression across multiple categories, we also focus on a single category—Line Drawing—to examine how the model optimizes the contributions of different backbones for this specific context. This visualization reveals how the weights dynamically adapt for Line Drawing, highlighting, for example, that certain backbones are better suited for the sparse and structural features inherent in these images. Over training epochs, the model learns to allocate more weight to the most effective backbone for this category, improving both interpretability and predictive accuracy (see Figure 7).

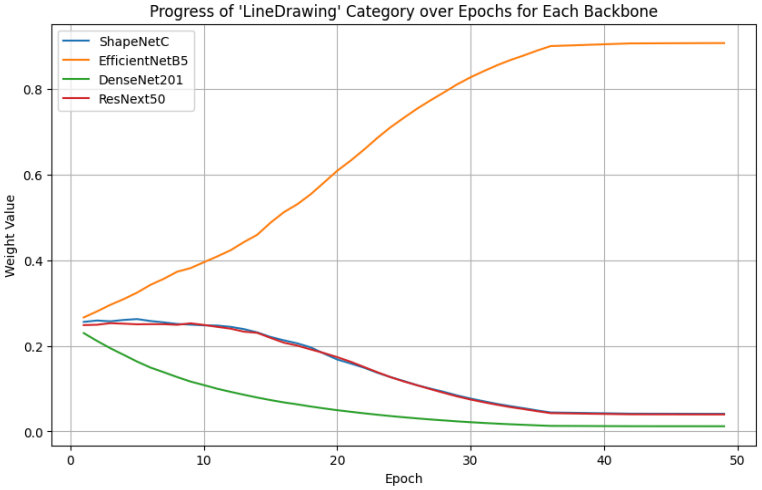
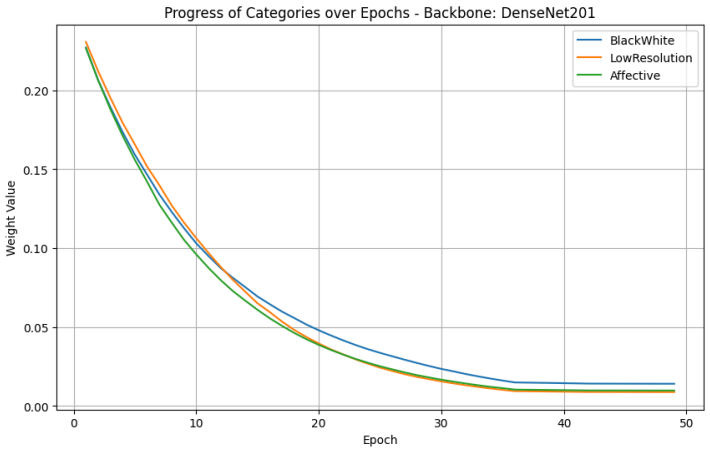
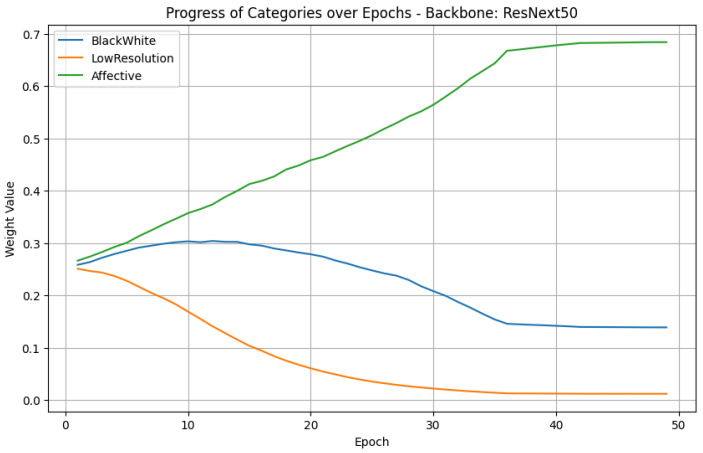
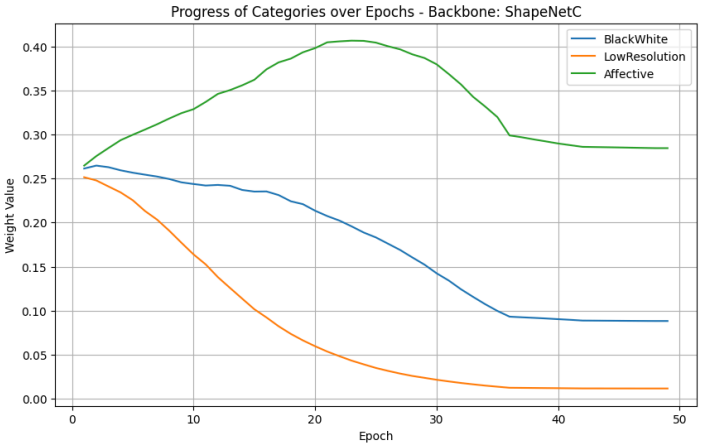


Fig. 7. This figure displays the evolution of weights for the Line Drawing category across epochs for each backbone (ShapeNetC, ResNext50, DenseNet201, and EfficientNetB5). The graph highlights how the ResNext50 backbone progressively becomes dominant as the model optimizes.







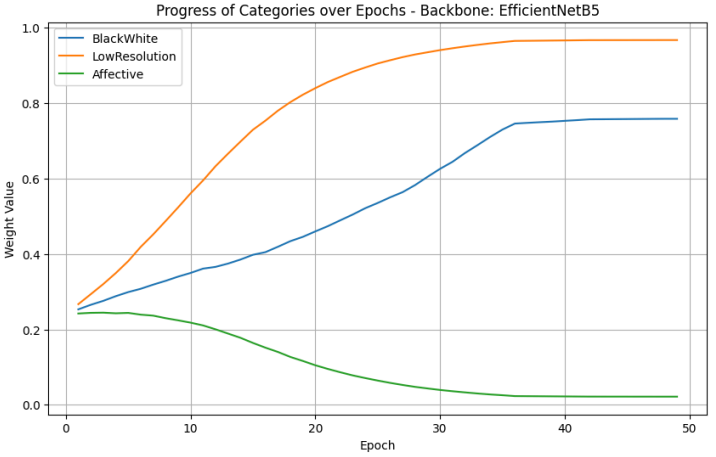


Fig. 8. This panel of graphs illustrates the weight progression for three categories—Black and White, Low Resolution, and Affective—across epochs for individual backbones (DenseNet201, ResNext50, ShapeNetC, and EfficientNetB5). Each graph highlights how the backbones adapt differently to these categories, with ResNext50 dominating in Affective, EfficientNetB5 excelling in Low Resolution, and varying performance across other categories.

To further validate the improvements brought by this adjustment iteration, we compare the weight progression of the same categories across the second and third iterations to see if results were random or changed but they stayed consistent. For this comparison, we use Line Drawing again as the focus category. In the second iteration, weights show less stability, reflecting the shorter convergence time and non random initialization seeds (see Figure 9). In contrast, the third and final iteration demonstrates smoother and more consistent weight adjustments across epochs, indicating that the longer training period allowed the model to converge more effectively and allocate weights optimally. This comparison highlights how increased training complexity and duration improved the model's ability to specialize in category-specific predictions.

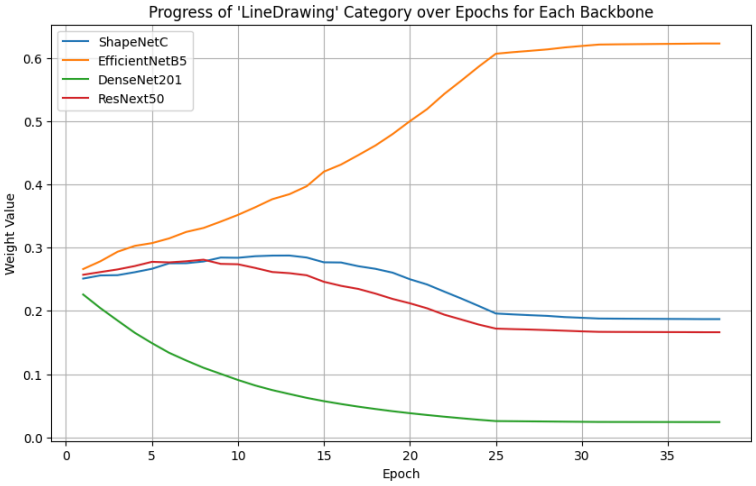


Fig. 9. This figure compares the weight progression of the Line Drawing category for all backbones during the second iteration. The third iteration (see Figure 7) demonstrates smoother and more consistent weight adjustments, reflecting improved convergence compared to the less stable changes observed in the second iteration. ResNext50 shows the most stable dominance in the final iteration.

The combination of these visualizations underscores the impact of adaptive weights and category-specific optimization in enhancing saliency prediction. By dynamically adjusting the contributions of backbones for specific categories, the model achieves improved performance and interpretability, adapting to the diverse visual challenges posed by the dataset.

VII. Final Results

In this section, we analyze the performance of our proposed model, DeepGazeIIE ACW (Adaptive Category Weighting), in comparison to state-of-the-art saliency prediction models using the CAT2000 dataset and other top performing models [7]. This evaluation is based on several established saliency metrics, as presented in Table 2. Additionally, we provide visual and quantitative justifications for these metrics, referencing the provided images, and the theoretical basis of each metric as explained in [16].

One of the key metrics, Information Gain (IG), measures how much better the model predicts fixation locations compared to a baseline, which in this case is the Center Bias Model (CBM). A positive IG value, such as the 0.1795 achieved by our model, confirms that it captures meaningful saliency information beyond the baseline. However, our IG is slightly lower than the original DeepGazeIIE (0.1893), suggesting that while our model excels at general saliency prediction, the reliance on the CBM limits its ability to adapt to more complex fixation patterns, particularly in categories like Pattern, where human attention is more distributed and less center-focused (Fig. 10 top 2 rows).

The Area Under the Curve (AUC) metric, which evaluates the model's ability to distinguish fixation from non-fixation points, further illustrates this point. Our model achieves an AUC of 0.8883, surpassing the original DeepGazeIIE (0.8692). This suggests a strong discriminative capability, but it's important to note that AUC does not penalize central bias, meaning that this high performance may stem from the model's tendency to predict central regions. This is evident in the top and last 2 rows of Fig. 10, where human fixations are naturally drawn to the center, resulting in higher AUC values. However, when we turn to Shuffled AUC (sAUC), which compensates for central bias, the story changes significantly. Our model's sAUC drops to 0.4946 compared to the original model's 0.6677, revealing its struggle to generalize beyond central tendencies.

This dependence on central bias also impacts the Normalized Scanpath Saliency (NSS) metric, which quantifies the alignment between predicted saliency and actual fixation points in units of standard deviation. Our NSS score of 1.2663 is substantially lower than the 2.1122 achieved by the original model. The lower NSS suggests that while our model identifies regions of interest, the predicted saliency maps lack the sharpness and confidence necessary to match fixation points closely. This can be

TABLE 2. MODELS SCORES ON THE CAT2000 BENCHMARK

Model	IG	AUC	sAUC	NSS	CC	KLDiv	SIM
DeepGazeIIE (SOTA)	0.1893	0.8692	0.6677	2.1122	0.8189	0.3448	0.706
DeepGazeII	0.0839	0.864	0.6498	1.9619	0.795	0.3815	0.6865
UNISAL	0.0321	0.8604	0.6684	1.9359	0.7399	0.4703	0.6633
ICF	-0.0229	0.8561	0.6187	1.9588	0.7791	0.4448	0.6697
SalFBNet		0.8549	0.633	1.8789	0.7027	1.1983	0.6425
DeepGaze I	-0.1546	0.8524	0.6184	1.843	0.7403	0.5094	0.6391
Ensembles of Deep Networks (eDN)		0.847	0.5782	1.2092	0.5003	0.9832	0.4508
DeepGazeIIE (ACW)	0.1795	0.8883	0.4946	1.2663	0.345	8.3372	0.3427

This table presents the performance of the top seven models [13], along with our proposed model (DeepGazeIIE ACW), evaluated on the CAT2000 dataset. As mentioned in the Dataset section, the evaluation utilized 20 images and their corresponding fixation data per category. I calculated each metric for each category and the mean of all of them is the one displayed. It is important to note that this evaluation does not follow the official CAT2000 benchmarking procedure, which involves submitting models to the centralized benchmark platform to assess their performance across various datasets, such as COCO Freeview, MIT300, and CAT2000. This omission reflects the adaptation of the evaluation process, as we used custom code derived from the evaluation process to calculate the metrics, rather than relying on the standardized benchmark submission process [23]. This could positively or negatively affect the model's metrics.

An important point to consider is that not all models listed in the table are probabilistic, meaning they do not produce outputs in the form of probability distributions. Consequently, these models are not directly comparable using Information Gain (IG), as IG specifically evaluates probabilistic models by comparing their predictions against a baseline like the Center Bias Model (CBM). For reference, probabilistic models generate a likelihood over possible fixation distributions, while non-probabilistic models output deterministic saliency maps that cannot be evaluated through IG [24].

observed in the smoother and more diffuse activation maps generated by our model. We can clearly observe this behavior in all our model predictions (Fig. 10).

The Correlation Coefficient (CC) provides additional insights into the spatial alignment between the predicted and ground truth saliency maps. With a CC of 0.345, our model falls significantly short of the 0.8189 achieved by DeepGazeIIE. This disparity underscores the diffuse nature of our predictions, which, while visually smooth, fail to capture the spatial precision required for higher correlation. As shown in Fig. 5 (third and fourth row), the original model's predictions exhibit tighter clusters, whereas our predictions are more spread out, likely due to the aggressive downsampling and smoothing in our architecture.

Similarly, the Kullback-Leibler Divergence (KLD), which measures the difference between the predicted and ground truth saliency distributions, highlights this issue. Our KLD score of 8.3372 is significantly higher than the original model's 0.3448, indicating a greater divergence. This is likely a direct consequence of our model's interpolation techniques, which, while effective for generating smoother predictions, result in less precise probability distributions.

Finally, the Similarity (SIM) metric, which evaluates the overlap between predicted and ground truth distributions, paints a consistent picture. Our SIM score of 0.3427, compared to the original model's 0.706, reflects the reduced overlap caused by the diffuseness of our predictions. This metric underscores the trade-off between smoothness and precision in our model's architecture, where the aggressive downsampling and reliance on the fine tuned upsampling result in saliency maps that are less aligned with ground truth fixations.

In summary, our model demonstrates strengths in metrics like AUC, where it outperforms the state-of-the-art by leveraging its discriminative power. However, its reliance on central bias and the inherent diffuseness of its predictions limit its performance in metrics like NSS, CC, KLD and SIM, which require precise spatial alignment. These findings suggest that while our model is effective at capturing general saliency patterns, further refinement is needed to improve its spatial precision and adaptability to more complex fixation behaviors. Future work could focus on reducing the reliance on central bias, optimizing interpolation methods, and refining the weight adjustment process to enhance performance across all metrics.

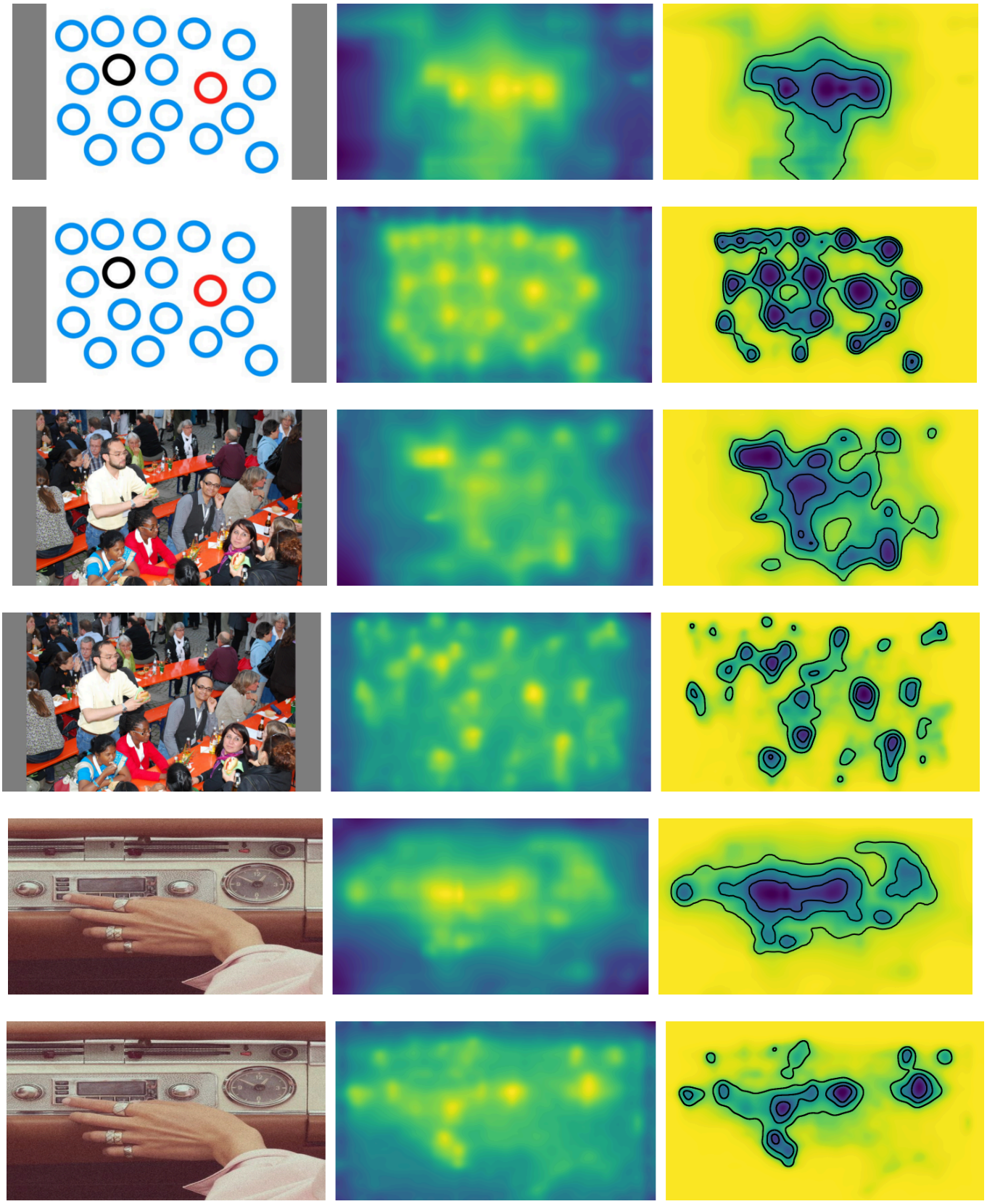


Fig. 10. This figure compares the visual predictions of our proposed model (DeepGazeIIE ACW) with the state-of-the-art (DeepGazeIIE) across multiple images. Each row pair represents the same input image: the first row showcases the predictions from our model, while the second row presents the corresponding predictions from the state-of-the-art model.

For each row, three panels are displayed:

- Original Image (left): The input image used for saliency prediction.
- Log Density (center): The activation maps showing the areas of interest predicted by the respective model.
- Predicted Saliency Map (right): The final saliency map highlighting areas of predicted visual attention, represented with contours over a heatmap.

VIII. Conclusion & Discussion

Through this project, we aimed to enhance the performance of the state-of-the-art DeepGazeIIE model by introducing category-specific weighting mechanisms and optimizing the architecture for better generalization. However, as reflected in the evaluation metrics, our model ultimately fell short of surpassing the state-of-the-art performance. Despite this, the journey of iterative development provided valuable insights into the intricacies of saliency modeling and opened promising directions for future work.

At the outset of this project, our model faced significant limitations, producing decent visualizations and lacking spatial coherence. The first iteration served as a baseline and demonstrated the challenges introduced by aggressive downscaling and the lack of an effective upsampling strategy. However, as we refined the model through subsequent iterations, we observed consistent improvements in visual outputs. By the final iteration, the model achieved metrics that, while still falling short of state-of-the-art benchmarks, were nonetheless comparable to top-performing models on the CAT2000 dataset. This progression underscores the importance of iterative experimentation and evaluation in deep learning research.

One of the key discoveries in this project was the implementation of category-specific weighting matrices across the backbones. This innovative approach demonstrated how leveraging the inherent differences in visual stimuli across categories can improve model generalization. This mirrors the natural variability in human visual attention, where our gaze is influenced by the type of visual stimulus we encounter. This finding suggests a broader potential for saliency models to specialize in specific domains, opening new avenues for applications in health (e.g., diagnosing visual impairments or designing visual aids), marketing (e.g., optimizing advertisements for target demographics), and autonomous systems (e.g., enhancing robotic navigation in diverse environments).

Despite these advancements, several challenges limited the model's performance. The heavy reliance on a fixed Gaussian center bias (which isn't tailored for our training data) and the aggressive downsampling steps contributed to less precise predictions, as reflected in metrics such as sAUC, NSS, CC, SIM and KLDiv. While the introduction of progressive upsampling and smoothing in later iterations helped visual results, there remains significant room for improvement in preserving spatial detail and reducing over-reliance on central predictions. Future work could explore alternative center bias implementations or more sophisticated upsampling techniques to address these limitations.

In conclusion, while our model did not achieve state-of-the-art performance, the iterative development process revealed promising directions for future research. The ability to incorporate category-specific adaptations highlights the potential for more context-aware saliency models that better align with human visual attention patterns. This opens the door to domain-specific saliency prediction. Moving forward, addressing the challenges identified in this project and building on the insights gained could pave the way for more robust and adaptive saliency prediction models.

Appendices

1. Github Repository:
https://github.com/AdrianGalvanDiaz/DeepGazeIIE_ACW

References

[1] Borji, A., & Itti, L. (2015). Cat2000: A large scale fixation dataset for boosting saliency research. arXiv preprint arXiv:1505.03581.

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

[3] Pan, J., Ferrer, C. C., McGuinness, K., O'Connor, N. E., Torres, J., Sayrol, E., & Giro-i-Nieto, X. (2017). Salgan: Visual saliency prediction with generative adversarial networks. arXiv preprint arXiv:1701.01081.

[4] Image Processing Group - UPC/BarcelonaTECH. (2020, 30 January). Visual Saliency Prediction with Deep Learning - Kevin McGuinness - UPC TelecomBCN Barcelona 2019 [Video]. YouTube. https://www.youtube.com/watch?v=L0-Sm8ERvVU

[5] Kümmerer, M., Theis, L., & Bethge, M. (2014). Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. arXiv preprint arXiv:1411.1045.

[6] MIT Saliency Benchmark. (s. f.). http://saliency.mit.edu/results_mit300.html

[7] Cornia, Marcella & Baraldi, Lorenzo & Serra, Giuseppe & Cucchiara, Rita. (2016). A Deep Multi-Level Network for Saliency Prediction. 10.1109/ICPR.2016.7900174.

[8] Huang, X., Shen, C., Boix, X., & Zhao, Q. (2015). SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks. 2015 IEEE International Conference on Computer Vision (ICCV), 262-270.

[9] Kruthiventi, S. S., Ayush, K., & Babu, R. V. (2017). Deepfix: A fully convolutional neural network for predicting human eye fixations. IEEE Transactions on Image Processing, 26(9), 4446-4456.

[10] Jiang, L., Wang, Z., Xu, M., & Wang, Z. (2019). Image Saliency Prediction in Transformed Domain: A Deep Complex Neural Network Method. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 8521-8528. https://doi.org/10.1609/aaai.v33i01.33018521

[11] Linardos, A., Kümmerer, M., Press, O., & Bethge, M. (2021). DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 12919-12928).

[12] Judd, T., Durand, F., & Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations.

[13] MIT/Tuebingen Saliency Benchmark. (s. f.). https://saliency.tuebingen.ai/results_CAT2000.html

[14] Herman Kamper. (2023, April 21). What is the difference between negative log likelihood and cross entropy? (in neural networks) [Video]. YouTube. https://www.youtube.com/watch?v=ziq967YrSsc

[15] Machine Leraning Mastery. (2020, December 10). https://machinelearningmastery.com/information-gain-and-mutual-information/

[16] Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2018). What do different evaluation metrics tell us about saliency models?. IEEE transactions on pattern analysis and machine intelligence, 41(3), 740-757.

[17] Matthias-K. (s. f.). GitHub - matthias-k/pysaliency: Python Framework for Saliency Modeling and Evaluation. GitHub. https://github.com/matthias-k/pysaliency?tab=readme-ov-file

[18] ReduceLROnPlateau — PyTorch 2.5 documentation. (s. f.). https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html

[19] Rizzo, M., Anderson, S. W., Dawson, J., Myers, R., & Ball, K. (2000). Visual attention impairments in Alzheimer's disease. Neurology, 54(10), 1954-1959.

[20] Treue, S. (2003). Visual attention: the where, what, how and why of saliency. Current opinion in neurobiology, 13(4), 428-432.

[21] SR Research Ltd. (2024, November 15). EyeLink 1000 Plus - Fast, accurate, reliable eye tracking. Fast, Accurate, Reliable Eye Tracking. https://www.sr-research.com/eyelink-1000-plus/

[22] Kümmerer, M., & Bethge, M. (2023). Predicting visual fixations. Annual Review Of Vision Science, 9(1), 269-291. https://doi.org/10.1146/annurev-vision-120822-072528

[23] Matthias-K. (s. f.-b). GitHub - matthias-k/saliency-benchmarking: Code for evaluating models in the MIT/Tuebingen saliency benchmark. GitHub. https://github.com/matthias-k/saliency-benchmarking

[24] Jetley, S., Murray, N., & Vig, E. (2016). End-to-end saliency mapping via probability distribution prediction. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5753-5761).