Brand Visibility in Packaging: A Deep Learning Approach for Logo Detection, Saliency-Map Prediction, and Logo Placement Analysis

Alireza Hosseini[a], Kiana Hooshanfar[a], Pouria Omrani[b], Reza Toosi[c], Ramin Toosi[a], Zahra Ebrahimian[a], Mohammad Ali Akhaee[a,*]

[a] *School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran*
[b] *Faculty of Electrical Engineering, K. N. Toosi University of Technology, Tehran, Iran*
[c] *Department of Computer Engineering, Faculty of Engineering, Golestan University, Gorgan, Iran*

Abstract

In the highly competitive area of product marketing, the visibility of brand logos on packaging plays a crucial role in shaping consumer perception, directly influencing the success of the product. This paper introduces a comprehensive framework to measure the brand logo's attention on a packaging design. The proposed method consists of three steps. The first step leverages YOLOv8 for precise logo detection across prominent datasets, FoodLogoDet-1500 and LogoDet-3K. The second step involves modeling the user's visual attention with a novel saliency prediction model tailored for the packaging context. The proposed saliency model combines the visual elements with text maps employing a transformers-based architecture to predict user attention maps. In the third step, by integrating logo detection with a saliency map generation, the framework provides a comprehensive brand attention score. The effectiveness of the proposed method is assessed module by module, ensuring a thorough evaluation of each component. Comparing logo detection and saliency map prediction with state-of-the-art models shows the superiority of the proposed methods. To

*Corresponding author.
Email addresses: arhosseini77@ut.ac.ir (Alireza Hosseini), k.hooshanfar@ut.ac.ir (Kiana Hooshanfar), pouriaomrani@ieee.org (Pouria Omrani), rtoosi81@gmail.com (Reza Toosi), r.toosi@ut.ac.ir (Ramin Toosi), z.ebrahimian@ut.ac.ir (Zahra Ebrahimian), akhaee@ut.ac.ir (Mohammad Ali Akhaee)

investigate the robustness of the proposed brand attention score, we collected a unique dataset to examine previous psychophysical hypotheses related to brand visibility. the results show that the brand attention score is in line with all previous studies. Also, we introduced seven new hypotheses to check the impact of position, orientation, presence of person, and other visual elements on brand attention. This research marks a significant stride in the intersection of cognitive psychology, computer vision, and marketing, paving the way for advanced, consumer-centric packaging designs.[1]

Keywords: Brand Attention, Neuro Marketing, Logo Detection, Saliency Prediction

1. Introduction

In today's dynamic business world, having a strong brand presence is crucial. The visibility of the brand is incredibly important for keeping up with consumer trends and staying competitive. Consumers often shape their perceptions of brands by considering factors such as visual attractiveness, functionality, and the social significance they convey, predominantly relying on visual cues (Bloch, 1995). For companies striving to establish and maintain a strong market presence, the packaging of their products, as an interface between the brand and the consumer, significantly influences the purchasing process (Ampuero & Vila, 2006).

The visual appeal of packaging, along with the prominent display of design elements, contributes to creating a lasting impression on the consumer and nurturing brand recognition. As consumers navigate the diverse market landscape, a well-designed package captures attention and effectively conveys the brand's values and identity, playing a key role in influencing the purchasing decision (Méndez et al., 2011).

Several studies in marketing and consumer behavior have emphasized the

[1] The source code and dataset are available at: `https://github.com/Arhosseini77/Brand_Attention`

role of effective packaging design in promoting brand recognition (Stewart, 1995). A well-designed packaging has been shown to significantly enhance brand awareness, purchase intent, and sales (Shukla et al., 2022; **?**). These investigations thoroughly explore various aspects of packaging design, conducting a detailed examination of elements such as packaging's shape, texture, and color (Riaz & Ghafoor, 2019; Dong & Gleim, 2018; Rebollar et al., 2015; Piqueras-Fiszman et al., 2013; Raheem et al., 2014). Additionally, they explore the strategic considerations of precise positioning of design elements such as logos, aiming to uncover the subtle interactions between these factors and their impact on consumer perception and brand recognition.

Recognizing the impact of visual elements in packaging, particularly logos, on shaping brand recognition and recall is crucial. This visual aspect influences consumer responses, ultimately playing an important factor in determining the success of a product. Logo, as a fundamental visual element, plays an essential role in packaging design, significantly influencing how consumers perceive and remember a brand (Girard et al., 2013). A visually appealing package not only captures the consumer's attention but also enhances the visibility of the brand logo. On the flip side, weaknesses in design can hinder logo visibility, diminishing its potential impact on consumer awareness (Krishna et al., 2017; Otterbring et al., 2013).

Understanding the crucial influence of logo visibility on brand awareness highlights the importance of implementing effective methods to enhance logo visibility. Enhancing logo visibility is linked to understanding its strategic placement on the packaging. The positioning of a logo profoundly impacts its visibility, influencing its interaction with other design elements and resonance with consumers. Thus, it is imperative to focus on optimizing logo placement through strategic positioning on packaging. To this end, implementing advanced machine vision techniques to measure logo visibility becomes crucial to amplifying visibility. Identifying the logo's position within an image is the initial stage in assessing logo visibility (Hou et al., 2023). The pursuit of brand visibility does not conclude with knowing the location of the logo; it extends to under-

standing the attention it commands within the consumer's visual field. This is where saliency prediction (Borji & Itti, 2012) emerges as a pivotal metric. Saliency prediction involves forecasting the perceptual prominence of the logo within the overall visual composition of packaging. Understanding the saliency prediction of the logo enables us to quantify its presence and visual impact, offering a detailed understanding of how much attention the brand attracts on the visual journey of consumers.

The proposed method is positioned to provide a comprehensive framework that includes automated logo detection and a thorough analysis of saliency prediction. This approach is crafted to provide businesses with actionable insights aimed at optimizing logo visibility and creating engaging packaging designs that effectively connect with their target audience. It is composed of three key modules. The initial module of our design is the brand logo detection, leveraging the cutting-edge YOLOv8 architecture. This crucial step helps identify and precisely locate brand logos in visual content. Subsequently, the second module, utilizing a CNN-Transformer-based model generates saliency maps, a crucial element of our methodology. These maps highlight specific regions within the visuals that command the highest visual attention. These insights provide valuable information regarding viewer perception and cognitive responses. The third and concluding module efficiently integrates the outcomes of both logo detection and saliency map generation. This integration yields a score that quantifies the attention that the brand logo attracts within packaging or advertising visuals. Furthermore, it is noteworthy to mention that this approach has been validated against existing psychophysical studies related to brand logos in packaging. This validation underscores the capability of the model to simulate human visual attention on brand logos within packaging and advertising imagery accurately. Consequently, this positions our model as a tool for investigating unexplored experiments regarding brand logos in packaging and advertising contexts. Through this approach, the proposed model provides a comprehensive analysis of brand visual attention, enabling businesses to make informed decisions to enhance their brand presence and impact. Our main con-

tributions are as follows:

- We utilized the YOLOv8 framework to create a logo detection model that outperforms others in accuracy and efficiency.
- A new saliency prediction model, specifically designed for advertising images and packaging considering text maps, is proposed. This model surpasses state-of-the-art models in saliency prediction.
- An advanced framework is introduced, designed to measure the level of attention directed towards the brand logo in both packaging and promotional images.
- A novel brand attention dataset is introduced to be generated based on a cognitive perspective, exploring 12 different hypotheses.

The rest of this work is organized into four sections. Section 2 delves into related work in the field, specifically focusing on optimizing logo placement through eye-tracking, brand logo detection, and saliency map prediction. Section 3 outlines the materials, methods, and modeling procedures employed in the research. Section 4 is dedicated to discussing the experiments conducted and the results obtained. Finally, section 5 presents the main conclusions of the work, while proposing future directions and potential enhancements for the introduced architecture.

2. Related Works

In this section, we will go through the domain of artificial intelligence (AI) and its applications in the field of marketing, specifically focusing on logo placement design in advertising images and packaging. We will also explore the techniques of brand logo detection and saliency map prediction, discussing their relevance to enhancing brand recognition and optimizing advertising effectiveness.

2.1. Optimizing Logo Placement with Eye-Tracking

Neuromarketing, an increasingly influential field of study, uniquely utilizes neuroscience knowledge to directly assess product packaging, eliminating the need to depend on consumers' self-reported preferences (Hubert et al., 2008). By incorporating advanced methodologies like neuroimaging and physiological measurements, neuromarketing employs a more direct and objective approach to assessing consumer responses. This represents a notable shift away from traditional survey-based approaches. A key methodology in neuromarketing is eye tracking (Alvino et al., 2021), providing a detailed examination of visual attention patterns. By studying where and how consumers focus their gaze, researchers gain valuable insights into elements that capture attention and drive perception, uncovering processes beyond conscious awareness (Maynard et al., 2018; Gofman et al., 2009).

Specific parameters govern visual behavior, with fixations playing a central role in this context. Fixations, characterized by eye movement, represent moments when the visual system actively acquires information (Pertzov et al., 2009). Numerous studies exploring eye movements, the attention mechanism, and consumer behavior have consistently emphasized the importance of analyzing fixations based on their frequency and duration (Nagel et al., 2011). By understanding the patterns and characteristics of fixations, researchers gain insights into how individuals allocate their visual attention and engage with stimuli. This knowledge proves particularly valuable in fields such as neuromarketing, where assessing consumer responses relies on a detailed understanding of visual attention dynamics.

Employing eye-tracking techniques, previous researches underscore the critical role of packaging design, investigating the influence of specific attributes like color, shape, and labeling on consumer perceptions of the product (Ares & Deliza, 2010). Strategic positioning of packaging design components is central to practical marketing efforts (Rettie & Brewer, 2000). Inadequate placement may cause crucial design elements to go unnoticed, impacting product evaluation (Krishna et al., 2017; Otterbring et al., 2013). An important study reveals

a consumer preference for high-power brands when the brand logo is positioned on the upper side of the packaging, contrasting with diminished appeal when placed on the lower side (Riaz & Ghafoor, 2019; Dong & Gleim, 2018). The effectiveness of capturing participants' attention by placing packaging content at the top is emphasized by Rebollar et al. (2015). Building on existing research, Piqueras-Fiszman et al. (2013) explored the impact of packaging shape and images on consumer attention, with a focus on the logo. Their findings showed that squared-shaped packaging significantly heightened attention toward the logo. Additionally, the study demonstrated the substantial influence of incorporating images on capturing consumer attention. This highlights the complex balance required in packaging design to ensure that attention is not only captured but also sustained, emphasizing the need for strategic placement and thoughtful integration of visual elements to prevent essential components from being marginalized.

2.2. Brand Logo Detection

Logo detection, a subfield of object detection, has witnessed substantial advancements over the years. In its initial stages, logo detection heavily relied on manually crafted visual attributes, including the Scale-Invariant Feature Transform (SIFT) and the Histogram of Oriented Gradients (HOG), combined with traditional classification models like Support Vector Machines (SVM)(Boia et al., 2015; Sahbi et al., 2013; Revaud et al., 2012). However, these approaches faced notable constraints. They were time-consuming because of their region-selective search method using sliding windows. They also struggled to handle different types of logos and were not very efficient at adapting to new situations (Hou et al., 2023). In recent years, deep learning has emerged as the prevailing paradigm for logo detection. These approaches can be categorized into different strategies, including Region-based Convolutional Neural Network (R-CNN) models and YOLO-based models. R-CNN models (Girshick et al., 2014), Fast R-CNN (Girshick, 2015), and Faster R-CNN (Ren et al., 2015) have made noteworthy contributions to the field of logo detection. Hoi et al. (2015) introduced

the Deep Logo-DRCN scheme, which investigated various techniques within the field of deep region-based convolutional networks (DRCN) for improved logo detection. Similarly, Oliveira et al. (2016) proposed an automatic graphic logo detection system based on Fast R-CNN, known for its robustness under unconstrained imaging conditions. Their approach involved utilizing transfer learning and data augmentation to train a CNN model, enabling multiple detection of potential regions containing objects. Additionally, Li et al. (2017) developed Faster R-CNN for logo detection, incorporating transfer learning, data augmentation, and clustering to optimize hyper-parameters and anchor precision in the Region Proposal Network (RPN), resulting in a significant improvement in detection accuracy.

Feature Pyramid Networks (FPN) are crucial in addressing the multi-scale problem in object detection (Lin et al., 2017). FPN notably enhances small object detection without escalating computational demands. Recent works have employed FPN to improve logo detection. Meng et al. (2021) proposed OSF-Logo, incorporating the Regulated Deformable Convolution (RDC) module in a specific layer of FPN. This integration allows adaptive adjustments of convolution kernel positions, facilitating geometric adaptations to logos. In addition, Jin et al. (2020) developed Brand Net, utilizing FPN to extract multi-scale features for logo recognition. To enhance small object detection in the context of logo recognition, FPN have also been integrated into Detection Transformers (DETR) (Carion et al., 2020). Velazquez et al. (2021) integrated FPN into DETR, enhancing small object detection. Nevertheless, this approach results in an increased computational load during backward propagation. More recently, Hou et al. (2021) proposed the Multi-Scale Feature Decoupling Network (MFDNet) to distinguish between multiple logo categories. MFDNet incorporates a Balanced Feature Pyramid (BFP) for merging multi-scale features and a Feature Offset Module (FOM) with an anchor region proposal network for the optimal selection of logo features.

Driven primarily by the compelling demand for speed and real-time object detection applications, You Only Look Once (YOLO) was developed (Redmon

et al., 2016). YOLO models, known as single-stage detectors, have played a central role in revolutionizing object detection for their ability to achieve both accuracy and speed. Early versions of YOLO, such as YOLOv2 (Redmon & Farhadi, 2017) and YOLOv4 (Bochkovskiy et al., 2020), set new benchmarks in the field. More recent iterations, including YOLOv7 (Wang et al., 2023) and YOLOv8 (Jocher et al.), represent the current state-of-the-art in object detection. YOLO models are widely employed, particularly in the domain of logo detection. Paleček & Chaloupka (2021) presented Scaled YOLOv4, outperforming traditional two-stage models such as Faster R-CNN in both speed and accuracy. It achieved a relative improvement of up to 46%, running up to twice as fast. Notably, logo detectors utilizing YOLOv7 and YOLOv8 remain unexplored, presenting an opportunity for potential improvements in balancing accuracy and speed, potentially reaching the state-of-the-art in logo detection.

2.3. Saliency Map Prediction

Saliency prediction in computer vision involves the identification and anticipation of the most significant or salient regions within an image or video frame, likely to capture human attention. This process holds practical utility in various applications. CNNs are commonly used for saliency prediction tasks. (Kroner et al., 2020) introduced an encoder-decoder framework that incorporates several convolutional layers, each set at various dilation rates, to effectively grasp features on multiple scales. (Jia & Bruce, 2020) used deep CNN models to extract more useful visual features for saliency prediction. TempSal (Aydemir et al., 2023) enables sequential saliency map generation through a temporal information-based model, astutely exploiting human temporal attention patterns. The incorporation of transfer learning principles amplifies the potential of CNN models in the domain of saliency prediction, as exemplified in the works of (Kümmerer et al., 2016) and (Linardos et al., 2021). The fusion of RNN with CNN represents a hybrid approach in the field of both image and video saliency prediction, as introduced in the work of (Droste et al., 2020).

Researchers have been inspired by the achievements of attention in natural

language processing (NLP) and have started applying these models to computer vision tasks such as saliency prediction. (Cao et al., 2020) proposed a saliency prediction method named VGG-SSM. Their pipeline consists of three parts: feature extraction, multi-level integration, and a self-attention module. They demonstrated that refining global information from deep layers through a self-attention mechanism, in coordination with fine details in distant portions of a feature map, yields a comprehensive data enhancement process. Additionally, (Lou & et al., 2022) developed a transformer-based method with both DenseNet and ResNet backbones.

The works mentioned earlier were created for general use, while numerous other works have been suggested specifically for advertising purposes. (Lévêque & Liu, 2019) collected an eye-tracking database of video advertising and evaluated their analysis with state-of-the-art deep learning-based saliency models. (Liang et al., 2021) compiled an eye-tracking dataset comprising 1000 advertising images. Subsequently, they introduced a method that incorporates text features within advertising images, which considers the interaction between text region and pictorial region. (Kou et al., 2023) proposed confidence scores fusion for saliency prediction in advertising images, which is helpful to improve the robustness and performance. Another study, conducted by (Jiang et al., 2022), introduces the concept of salient Swin-Transformers. In this work, the researchers initially curated a dataset of e-commerce images for saliency prediction tasks. Subsequently, they proposed a novel multi-task learning framework that demonstrated state-of-the-art performance in e-commerce scenarios.

3. Proposed Method

The primary aim of our research is to design a system for a comprehensive evaluation of the visual prominence of brand logos within the context of packaging or advertising images. To achieve this objective, the proposed methodology encompasses three distinct modules, each designed to address specific aspects of this assessment, as illustrated in Figure1. The first module is dedicated to

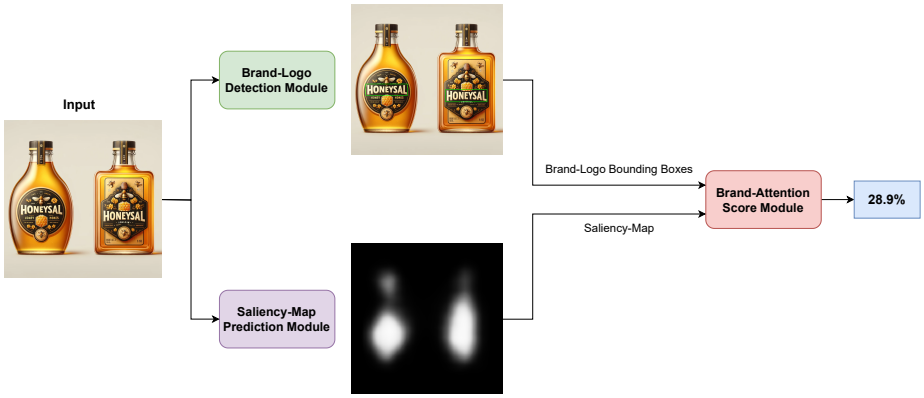


Figure 1: Overview of the proposed brand-attention method

brand logo detection and is supported by the state-of-the-art object detection model, *YOLOv8*. This module identifies and locates brand logos within the imagery, forming the foundational basis for subsequent analysis. Then, the second module focuses on generating saliency maps, a critical aspect of our approach. The saliency maps illuminate the regions within the image that command the highest degree of visual attention, providing valuable insights into viewer perception and cognition. The final module consolidates the outcomes of the brand logo detection and saliency map generation modules. This approach gives a score that measures how much attention the brand logo gets in the packaging or advertising image. This combination of techniques offers valuable insights for businesses aiming to optimize the visual prominence of their brand logos in marketing materials.

3.1. Brand Logo Detection

In the initial stage of the proposed method, our focus lies on brand logo detection. For this task, we employ the YOLOv8 model, specifically trained for logo detection purposes. When presented with an input image I with spatial dimensions $H \times W$ and C color channels, our Logo YOLOv8 model processes this image. The output of this model consists of a $1D$ list of bounding boxes, denoted as B, where each bounding box (b) is represented as a tuple containing

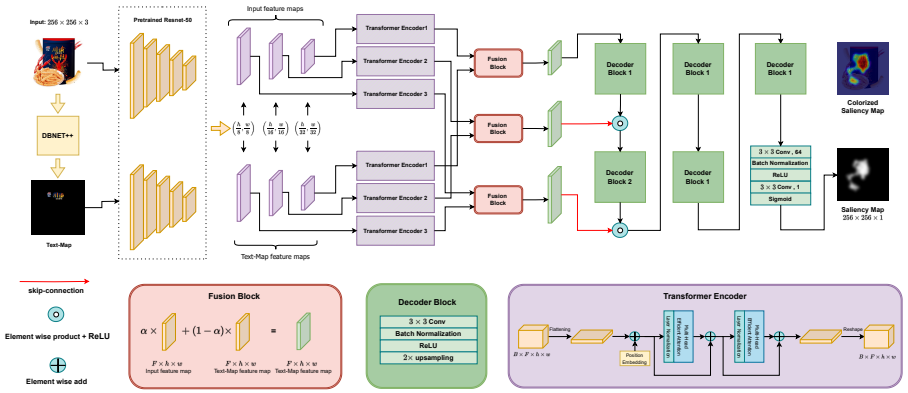


Figure 2: The block diagram of the proposed saliency model.

the coordinates $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$

$$B = \text{LOGO\_YOLOv8}(I) = [b_1, b_2, \ldots, b_n] \tag{1}$$

The number of logo boxes detected in the image is represented by n. This detection is the first fundamental step in our brand attention system.

3.2. Saliency Map Prediction

Our primary objective in the second stage is to generate saliency maps for images, with a specific focus on advertising and packaging designs. Previous studies have shown that text is just as important as other visual elements in packaging and advertising. These studies found that text is instrumental in capturing people's attention, and they used eye-tracking data to confirm this (Jiang et al., 2022), (Liang et al., 2021). Recognizing the paramount significance of text in advertising images, we introduce a novel saliency map prediction model tailored to address the unique requirements of both advertisements and packaging images. This model is inspired by the TranSalNet network (Lou & et al., 2022), with major improvements made to boost its efficiency and performance. In the proposed model, we initiate the process by detecting text within the image. To achieve this, a pre-trained text detection model is employed (Liao & et al., 2022), which outputs a text map. Both the text map and the original

image are subsequently processed through a CNN decoder, resulting in multiple feature maps. To efficiently capture and process information from feature maps, we apply transformation through Transformer encoders. This enables the model to consider complex relationships and dependencies within visual content. To ensure a seamless integration of these elements, we introduce a pivotal component of the model: the *Fusion Block*. This block is strategically designed to merge the feature maps derived from both the text map and the original image. By doing so, it enables the simultaneous utilization of visual and text-map features, thereby enhancing the overall interpretative capabilities of the proposed model. After the fusion block, we use a CNN decoder, which is supported by skip connections coming from the encoder section. This integrated process ensures the restoration of long-range context-enhanced feature maps obtained from the fusion block. These enhanced feature maps serve as the foundation for constructing the final saliency map, capturing the regions of the image that attract the most visual attention. Figure 2 illustrates the proposed saliency model, providing a visual representation of its architecture and the various components that comprise our refined saliency map prediction system. As depicted, the model comprises five principal components, each of which will be explained in more detail in subsequent sections of this paper.

3.2.1. Text Detector

We leverage the cutting-edge DBNet++ network (Liao & et al., 2022), which has emerged as a front-runner in the domain of text detection, consistently achieving state-of-the-art accuracy across a spectrum of five scene text detection benchmarks. These benchmarks cover a diverse range of challenges, from handling horizontal and multi-oriented text to curved text, demonstrating the versatility and performance of DBNet++. The DBNet++ operates on images with spatial dimensions of $H \times W$ and C channels, allowing it to accurately identify text regions within these images. By deploying this innovative network, we can precisely extract and isolate text from non-textual information, ultimately generating text maps. Given an input image $I \in \mathbb{R}^{H \times W \times C}$, the DBNet++

detects text regions denoted as R. As a consequence, a text map, denoted as $t_{\text{map}} \in \mathbb{R}^{H \times W \times C}$, is generated as follows:

$$t_{\text{map}}(x, y, c) = \begin{cases} I(x, y, c) & \text{if } (x, y, c) \in \text{R} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

3.2.2. CNN Encoder

A CNN encoder is employed as our feature extractor. The primary objective of this CNN encoder is to extract essential features from both the image and the text-map while ensuring that the spatial information is distinctly preserved. To achieve this, three sets of convolutional layers is used, each designed to capture features at different spatial scales. Specifically, we extract feature maps with spatial dimensions of $(w/8, h/8)$, $(w/16, h/16)$, and $(w/32, h/32)$. For the image and text-map image feature extraction, the ResNet-50 architecture is utilize (He et al., 2016).

3.2.3. Transformer Encoder

After the initial CNN Encoder stage, which focuses on enhancing long-range and contextual information within our data, we deploy three distinct transformer encoders designed to efficiently capture and process this enriched information. In the pipeline, transformer encoders are integrated to handle the unique characteristics of both original images and text-maps. Specifically, three sets of multi-scale feature maps, denoted as i_1, i_2, and i_3, are derived from the image data. These sets have spatial dimensions of $(w/32, h/32)$, $(w/16, h/16)$, and $(w/8, h/8)$, respectively. Each set is then fed into its respective transformer encoder. To adapt the input size of the transformer encoder and reduce computational complexity, we employ 1×1 convolution layers (conv1×1) with a stride of one. These convolution layers are applied to the input tensors, including i_1, i_2, and i_3, to decrease their channel dimensions while preserving spatial dimensions. The conv1×1 operation specifically reduces the dimensions of i_1, i_2, and i_3 from 2048, 1024, and 512 to 768, 768, and 512, respectively. This

dimension reduction streamlines the data for subsequent processing within the transformer encoder, aligning it with the required input dimensions and optimizing computational efficiency. Likewise, the textual components of the data, denoted as t_1, t_2, and t_3, undergo dimension reduction through conv1$\times$1 layers employing the same filter size and stride. This process ensures their alignment with the reduced dimensions of the visual components.

To facilitate position awareness and optimize the transformer encoders for effective processing of spatial information within these feature maps, we integrate position embeddings (PE) (Dosovitskiy et al., 2021) into the input before feeding it into the transformer encoders. Each transformer encoder in the proposed model consists of two identical layers featuring Multi-Head Efficient Attention (MEA) (Shen & et al., 2021) and multi-layer perceptron (MLP) blocks. Notably, the model's design deviates from Transalnet regarding the number of heads and layers in each transformer encoder. Specifically, transformer encoders employ one efficient attention head and a 2-layer MLP. These tailored configurations are designed to meet the specific requirements of our model, ensuring the efficient processing of the enriched feature maps. Additionally, the MLP block in each transformer encoder consists of two layers with a GELU activation function. Layer normalization (LNorm) and residual connections are applied before and after each block, ensuring stable and effective feature processing.

The introduced methodology distinguishes itself through the adoption of efficient attention, as proposed by Shen & et al. (2021), diverging from the conventional self-attention mechanism. Traditional self-attention is mathematically represented as

$$s(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

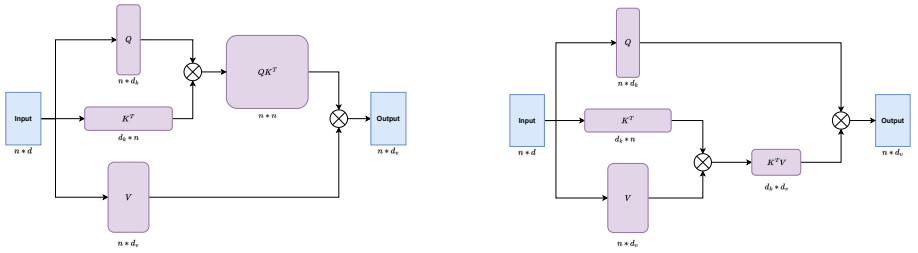
In this formula, Q, K, and V are the query, key, and value vectors, while d_k is the embedding dimension. However, this approach is limited by its $O(N^2)$ computational complexity, which presents major challenges when processing high-resolution images.

Efficient attention, on the other hand, optimizes this process by normalizing

the keys and queries before their interaction. Represented as,

$$E(Q, K, V) = \rho_q(Q)(\rho_k(K)^T V) \tag{4}$$

where ρ_q and ρ_k are normalization functions. This approach addresses the redundancy in the context matrix generation of standard self-attention. It reduces the computational complexity to $O(d^2 n)$, with a memory complexity of $O(dn + d^2)$, assuming $d_v = d$ and $d_k = d/2$. Here, d represents the embedding dimension. This model's efficient attention mechanism prioritizes a comprehensive understanding of the input feature, avoiding the computation of pairwise similarities. By treating keys as attention maps k_j^T and focusing on semantic information rather than positional similarities, it achieves a significant computational efficiency improvement without sacrificing representation richness. The diagram depicting the efficient attention mechanism discussed above is presented in Figure 3 (Shen & et al., 2021).



(a) Dot-Product Attention

(b) Efficient Attention

Figure 3: Architecture of dot-product and efficient attention(Shen & et al., 2021)

It can be summarized that for a given sample input m consisting of t_1 to t_3 (representing textual content) and i_1 to i_3 (representing image-based features), the transformer encoder process can be mathematically described as follows:

$$z_0 = \text{conv}_{1\times1}(m) \oplus \text{PE} \tag{5}$$

$$z_l^{'} = MEA(\text{LNorm}(z_{l-1}) \oplus z_{l-1}) \tag{6}$$

$$z_l = \text{MLP}(\text{LNorm}(z_l^{'}) \oplus z_l^{'}) \tag{7}$$

where z_l represents the output feature maps of the l-th layer in the transformer encoder. The feature maps that go through transformer encoders 1, 2, and 3 are contextually enhanced and are referred to as i_j^* for $j = 1$ to 3 for image and t_j^* for $j = 1$ to 3 for text map image.

3.2.4. Fusion Block

After generating enhanced visual features for the image and text map image, it is imperative to merge these features effectively. The fusion process involves assigning weights to the visual and textual modalities. We introduce weighting factors, denoted as α, which determine the influence of visual and textual data, respectively.

$$i_{f_j}^* = \sigma(\alpha) \cdot i_j^* + (1 - \sigma(\alpha)) \cdot t_j^* \tag{8}$$

In this equation, i_f^{*j} represents the final feature representation after the fusion process. The selection of the α parameter is of paramount importance since it governs the equilibrium between the visual and textual modalities. In the proposed model, we treat α as a learnable parameter, enabling the model to determine the optimal value for this factor. This dynamic approach allows the model to adapt and effectively combine visual and textual information based on the unique demands of the task at hand, thereby enhancing the overall performance and versatility of the model. To ensure that α remains within the valid range [0, 1] after optimization, a sigmoid function is applied. The sigmoid function, denoted as $\sigma(\cdot)$, maps real-valued inputs to the interval [0, 1], making it an ideal choice for constraining the α parameter.

3.2.5. CNN Decoder

The CNN decoder plays a key role in integrating and restoring long-range context-enhanced feature maps obtained from the fusion block. Its primary objective is to reconstruct the saliency maps while restoring the original image resolution. The suggested CNN decoder is designed to facilitate efficient and effective pixel-level classification, enabling the prediction of saliency maps. Within

the network, several key operations are performed to enhance the model's performance. After each 3×3 convolution operation (`Conv3×3`), the batch normalization (`BNorm`) is applied to promote convergence. Besides, the activation function ReLU is used in all blocks, with Sigmoid employed in the final block. After initial down-sampling of the input image to a 32-scale by the encoder network, a pivotal process in the CNN Decoder involves a 2-scale up-sampling. This method uses nearest-neighbor interpolation and happens in the first five decoding stages. It creates the saliency map that has the same size as the original input image.

To improve the feature map's long-range and multi-scale context during the decoding process, the up-sampled feature map is fused with the output from the fusion blocks, denoted as $i^*_{f_j}$ for $j = 1$ to 3. This fusion is acquired through the corresponding skip-connection, using an element-wise product operation and ensures that the model benefits from comprehensive contextual information at different scales.

The operations within each CNN decoder block can be represented as follows:

$$O_i = \begin{cases} i^*_{f_1}, & i = 1 \\ \text{ReLU}\left(\text{2X\_Upsample}(O_{i-1}) \cdot i^*_{f_i}\right), & i = 2, 3 \\ \text{2X\_Upsample}(O_{i-1}), & i = 4, 5, 6 \end{cases} \tag{9}$$

$$O^*_i = \text{ReLU}\left(\text{BNorm}(\text{Conv}_{3\times3}(O_i))\right), \quad \text{for } i : 1 \text{ to } 6 \tag{10}$$

$$S = \text{Sigmoid}(\text{Conv}_{3\times3}(O^*_6)) \tag{11}$$

where **S** represents the final saliency map predicted by the proposed model.

3.2.6. Loss Function

Drawing inspiration from established conventions in the domain of saliency map prediction models and referencing other saliency prediction frameworks ((Droste et al., 2020), (Che et al., 2020; Lou & et al., 2022)), our model employs a composite loss function. This function combines three metrics: Kullback-Leibler divergence (KL), Linear Correlation Coefficient (CC) and Mean Squared Error (MSE) loss.

Let g^s represent the ground truth of the saliency map, g^f denote the ground truth of the saliency fixation map, and S denote the network's predicted saliency map. The overarching loss function is defined as:

$$\text{Loss} = \lambda_1 \cdot \text{KL}(g^s, S) \quad + \lambda_2 \cdot \text{CC}(g^s, S) + \lambda_3 \cdot \text{MSELoss}(g^s, S) \tag{12}$$

where each component is elucidated as follows:

- **KL divergence**: A standard measure of dissimilarity between probability distributions, is expressed as:

$$\text{KL}(g^s, S) = \sum_{i=1}^{n} g_i^s \log\left(\epsilon + \frac{S_i}{g_i^s + \epsilon}\right) \tag{13}$$

  Here, ϵ serves as a regularization constant, set to 2.2×10^{-16}.

- **CC**: CC is defined as the ratio of the covariance between g^s and S to the product of their standard deviations, signifying similarity. The formula is presented as:

$$\text{CC}(g^s, S) = \frac{\text{cov}(g^s, S)}{\sigma(g^s) \cdot \sigma(S)} \tag{14}$$

  Here, $\sigma(.)$ designates the standard deviation, and cov(.) stands for the covariance.

The objective of this loss function is to minimize the KL and MSELoss while concurrently maximizing the value of CC. This dynamic balance is achieved through the fine-tuning of the coefficients λ_i, where i ranges from 1 to 3. By employing the Optuna framework (Akiba et al., 2019), we have systematically determined the values for these coefficients to achieve optimal training. The values of the coefficients are: $\lambda_1 = 10$, $\lambda_2 = -3$ and $\lambda_3 = 5$. Based on the achieved experiments, these coefficients have been chosen to optimize the model's performance, with a specific focus on reducing KL while concurrently enhancing CC, aligning closely with the intended outcome of the proposed model.

In our comprehensive evaluation framework, we use three additional metrics—Similarity (SIM), Normalized Scan-path Saliency (NSS) and Area under ROC Curve(AUC)—to provide an assessment of the model's performance.

While these metrics are not directly embedded within the training loss function, they play an important role in the evaluation phase.

- **SIM**: SIM gauges the linear relationship between the elements of g^s and S, where the minimum value at each position is summed to calculate the coefficient:

$$\text{SIM}(g^s, S) = \sum_{i=1}^{n} \min(g_i^s, S_i) \tag{15}$$

- **NSS**: NSS measures the similarity between the predicted S and g^f by comparing the fixations with the saliency map values:

$$NSS(g^f, S) = \frac{1}{\sum_i (g_i^f)} \sum_i \left(\frac{S_i - \mu(S)}{\sigma(S)} \right) g_i^f \tag{16}$$

  where, $\sigma(.)$ designates the standard deviation, and $\mu(.)$, cov(.) stands for the mean and covariance, respectively.

3.3. Brand-Attention Score

After localizing brand bounding boxes (B) and generating the saliency maps for both packaging and advertising images, we can quantitatively assess the prominence of the brand within an image. The fundamental concept involves converting the saliency map image into a list of pixel probabilities, ensuring that the cumulative probability sums to 1. Subsequently, we calculate the sum of probabilities associated with pixels contained within the image region.

The pseudo-code for calculating the brand attention score is presented in Algorithm 1. This pseudo-code outlines the procedure for computing the brand attention score based on the provided saliency map and bounding boxes. It involves removing saliency map values below a threshold, normalizing the remaining values to probabilities, and then calculating the score by summing the normalized values within the specified bounding box regions. Using the saliency map and this algorithm, we can obtain an attention score for every object or text (not only the brand logo) for which bounding boxes are provided or selected by users.

Algorithm 1: Brand-attention score calculation

Data: B, S

Result: Brand-Attention Score

S[S <Threshold] = 0 ;

SNorm = S / sum(S) ;

if B *is None* **then**

    return 0

else

    Brand-Attention Score = 0 ;

    for *b in* **B do**

        $x_{\min}, y_{\min}, x_{\max}, y_{\max}$ = b ;

        for *y in range*$(y_{min}, y_{max} + 1)$ **do**

            for *x in range*$(x_{min}, x_{max} + 1)$ **do**

                Brand-Attention Score += SNorm[x, y] ;

    return Brand-Attention Score ;

4. Experiments and results

In this section, we go through the datasets, training setup, and result analysis for both logo detection and saliency prediction. Moreover, the outcomes underscore the enhanced efficacy of the proposed technique compared to leading-edge methods across diverse evaluation metrics. The following part introduces the brand attention module and the proposed dataset. The brand attention module is then validated based on earlier hypotheses, with results thoroughly analyzed using feedback from human participants. The section concludes by proposing and discussing new hypotheses regarding brand visibility in packaging. The computational tasks described in this section were executed using the PyTorch framework on a workstation equipped with an Intel Core i-9 CPU and an NVIDIA GeForce RTX3090 GPU..

4.1. Logo Detection

4.1.1. Datasets

Over recent periods, the development of specialized datasets for logo detection tasks has garnered considerable focus within the domain of computer vision, providing a valuable resource for various applications (Hou et al., 2023). To address the challenge of logo detection within the context of diverse packaging products, this research has carefully curated a selection of datasets that align with our specific objectives.

Our attention is focused on two logo detection data sets: FoodLogoDet-1500 (Hou et al., 2021), and LogoDet-3K (Wang et al., 2022). These datasets were chosen for their unique attributes, making them well-suited for this research and the complexities associated with logo detection in the context of product packaging. A summary of the selected datasets is provided in Table 1. Additionally, to provide a visual perspective, Figure 4 presents a few samples of the FoodLogoDet-1500 and LogoDet-3K datasets.

Table 1: Summary of selected logo detection datasets

Dataset	**#Images**	**#Objects**	**#Logos**
FoodLogoDet-1500	99,768	145,400	1,500
LogoDet-3K	158,652	194,261	3,000

4.1.2. Training Setup

The dataset used for logo detection contains numerous classes, which are not essential for our specific case. Therefore, all classes have been aggregated into one for logo detection. Due to the inability of the proposed model to converge on large-scale datasets, a two-stage fine-tuning process has been implemented. In the first stage, the *small* version of the YOLOv8 model is fine-tuned, initially pre-trained on the COCO dataset, over the FoodLogoDet-1500 dataset. This initial fine-tuning serves as a crucial step to help the model adapt to the characteristics of the data and mitigate convergence issues. The fine-tuning process in this stage is carried out using the Adam optimizer across 100 epochs, with a batch size set to 32, and involves specifying a learning rate of 10^{-2} and a momentum of 0.9. During the second stage, we continued the fine-tuning process on both the FoodLogoDet-1500 and the larger LogoDet-3k datasets. This approach ensures that the model further adapts to a broader range of data patterns. The second-stage fine-tuning is conducted for 50 epochs with a batch size of 64, using the same hyperparameters as in the first stage. This two-stage fine-tuning strategy has proven effective in addressing the model convergence challenge. The entire process takes approximately 60 hours to complete.

4.1.3. Result Analysis and Comparison

We compared the proposed logo detection model with two state-of-the-art methods, including YOLOv7, which was employed as a base line for logo detection, and MFDnet Hou et al. (2023). Results are shown in Table 2 and Table 3. As can be observed, YOLOv8 significantly outperforms both YOLOv7 and MFDNet across various metrics, such as mAP50, mAP50-95, precision, and

(a) FoodLogoDet-1500

(b) LogoDet-3K

(c) SalECI

Figure 4: Sample images from FoodLogoDet-1500, LogoDet-3K, and SalECI datasets

recall, in both stages of evaluation.

Table 2: Metrics on models fine-tunned over Foodlogo-det-1500

Method	mAP_{50}	mAP_{50-95}	**Precision**	**Recall**
MFDNet	0.879	0.635	0.836	0.811
YOLOv7	0.932	0.698	0.90	0.866
YOLOv8	**0.936**	**0.704**	**0.904**	**0.879**

Table 3: Metrics on models which is pretrained on FoodLogo and fine-tunned over FoodLogoDet-1500+LogoDet3k dataset.

Method	mAP_{50}	mAP_{50-95}	**Precision**	**Recall**
MFDNet	0.87	0.62	0.82	0.8
YOLOv7	0.88	0.61	0.84	0.81
YOLOv8	**0.94**	**0.71**	**0.91**	**0.88**

4.2. Saliency Map Prediction

4.2.1. Dataset

In the domain of saliency map prediction tasks, various general-purpose datasets, including SALICON (Jiang et al., 2015), CAT2000 (Borji & Itti, 2015), MIT1003 (Judd et al., 2009), and MIT300 (Judd et al., 2012) have been established. However, this paper uniquely centers its focus on commercial and advertisement images. To address this specific focus, we leverage the Saliency E-commerce Images (SalECI) dataset introduced by Jiang *et al.* (Jiang et al., 2022). The SalECI dataset comprises 257,302 fixations obtained through eye-tracking experiments involving 25 subjects. The dataset comprises 972 e-commerce images, each paired with corresponding fixation maps and text boundaries. This dataset acts as an important tool for exploring saliency within the realm of commercial and advertising stimuli.

4.2.2. Training Setup

The proposed method utilizes 66 million parameters, a lower number compared to TranSalNet (Lou & et al., 2022). This reduction is achieved through the incorporation of efficient attention mechanisms and a decrease in the number of attention heads and layers. Notably, despite the inclusion of the text detector and fusion block, the overall parameter count remains lower than that of TranSalNet. During training, the proposed method was trained over the SalECI dataset (Jiang et al., 2022) using a step learning rate scheduler with a step size of 4 and a gamma value of 0.1. The initial learning rate was set to

5×10^{-4}, and weight decay was applied at a rate of 10^{-4}. The Adam optimizer is used for training.

4.2.3. Result Analysis and Comparison

Consistent with the importance of the text map to the proposed saliency prediction method for advertising purposes, a parameter named α is introduced. The saliency model is initially assessed without a text detector, followed by an evaluation with a fixed α parameter. Subsequently, the model undergoes another evaluation with the integration of a text detector, maintaining the fixed α value. Finally, the third evaluation is conducted with a learnable α parameter, where the initial value of 0.5 dynamically adjusts to 0.659 during the training process. It is noteworthy that the best result is achieved with the α value of 0.659, which is obtained through the training process.

Table 4 provides a comparative analysis of the saliency prediction accuracy for ten different state-of-the-art methods using the SalECI dataset. The saliency prediction results of models on the SalECI dataset are visualized in Figure 5. As seen in Table 4, the proposed method emerges as the leading method in this comparison, excelling in correlation coefficient, KL divergence, NSS, and similarity, demonstrating its effectiveness in predicting saliency in the SalECI dataset. These results provide valuable insights into the performance of these methods in the context of saliency prediction.

4.3. Brand Attention

In this section, we evaluate the effectiveness of the proposed brand attention module by comparing it with the observations in psychophysical studies. To test the model, we have designed a dataset where each group of images is the same in every way, apart from one particular logo feature it is examining. Wrapping up this section, we introduce some new hypotheses in this field that have not been explored yet.

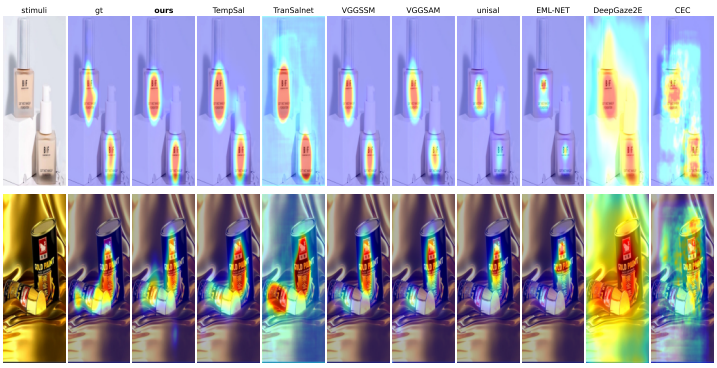


Figure 5: Comparison of the saliency maps of different models over SALECI

Table 4: Comparing the saliency prediction accuracy for the proposed and nine other state-of-the-art methods over SalECI.

Method	**CC ↑**	**KL ↓**	**AUC ↑**	**NSS ↑**	**SIM ↑**
Contextual Encoder-Decoder	0.459±0.136	1.1346±0.23	0.76±0.066	0.925±0.268	0.373±0.06
DeepGazeIIE	0.561±0.124	0.995±0.215	0.842±0.055	1.327±0.318	0.399±0.065
UNISAL	0.6±0.15	0.768±0.262	0.845±0.056	1.574±0.522	0.514±0.094
EML-Net	0.510±0.16	1.227±0.903	0.807±0.062	1.232±0.407	0.536±0.103
VGGSAM	0.691±0.126	0.682±0.259	0.815±0.048	1.324±0.362	0.58±0.091
Transalnet	0.717±0.061	0.873±0.079	0.824±0.054	1.723±0.203	0.534±0.043
VGGSSM	0.728±0.121	0.599±0.237	0.829±0.043	1.396±0.359	0.611±0.089
Temp-SAL	0.719±0.065	0.712±0.126	0.813±0.077	1.768±0.182	0.629±0.048
SSwin transformer	0.687±0.175	0.652±0.478	0.868±0.072	1.701±0.497	0.606±0.101
ours(without text detection)	0.741±0.061	0.6958±0.071	0.848±0.071	1.86±0.234	0.635±0.054
Ours	**0.75±0.050**	**0.578±0.117**	**0.892±0.033**	**1.89±0.204**	**0.645±0.040**

4.3.1. Dataset

While aiming to validate various hypotheses concerning logo placement and packaging design, we have created a dataset comprising 650 images. This collection is a systematically designed platform for testing various ideas connected to packaging design and how people perceive brands. To ensure a rich and varied base for our study, 95% of the images in this dataset are sourced from the Internet templates, complemented by those generated by DALL-E, an advanced

AI image generation tool. Each image has been carefully modified to align with specific research questions, with alterations ranging from subtle logo repositioning to more substantial design transformations. Our dataset is organized into 12 hypotheses, each examining different aspects of design and branding. For each hypothesis, we analyze 18 ± 3 images per hypothesis. In each hypothesis image set, all logo characteristics are fixed, except the one under experiment. This setup provides us with an in-depth insight into the influence of packaging design and logo placement on brand perception.

4.3.2. Previous Hypothesises Analysis

We evaluate the effectiveness of our brand attention model by comparing its output to data from human observers who have studied logo attention. This comparison involves aligning the model's predictions with findings from psychophysical studies, ensuring its accuracy in predicting how humans notice logos for real-world applications. The following subsequent items provide a summary of the studies that form the basis of this comparative analysis, showcasing their relevance in the context of brand logo attention:

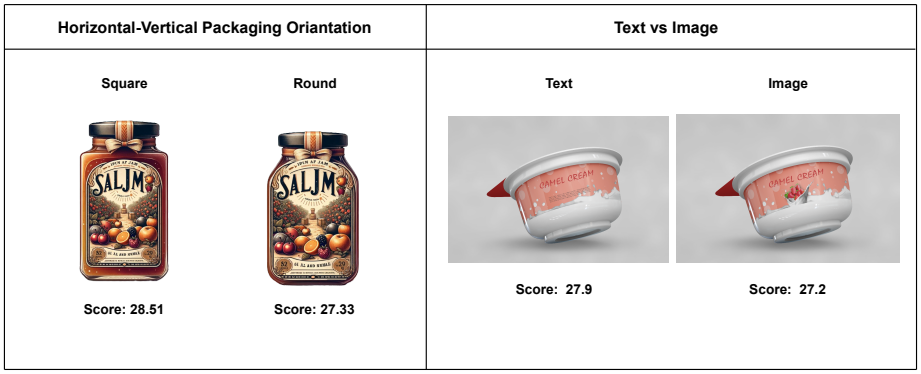


Figure 6: Sample images illustrating the influence of packaging shape (left) and the presence of an image on directing attention to different elements, such as the logo (right).

1. **Study 1 (Piqueras-Fiszman et al., 2013)**

   This study examined the influence of packaging shape and the presence of an image on attention to different elements, like the logo. It was found

that a squared shape, as opposed to a rounded one, drew more attention to the logo. They have also demonstrated that the photo element on product packaging was highly influential in drawing consumer interest rather than text. Backing these ideas, our initial results, as displayed in Table 5, suggest that packaging with a squared shape indeed garnered more attention to the logo compared to rounded shapes. Notably, the obtained results align with existing research, underscoring the substantial influence of incorporating visual elements, particularly photos, on product packaging in capturing consumer attention. Figure 6 showcases a sample of images created for testing this thesis.

2. **Study 2 (Dong & Gleim, 2018; Riaz & Ghafoor, 2019)**

   The findings from these two studies underscore a noteworthy connection between logo placement and consumer purchase intention. The research reveals a clear preference for high-power brands when logos are positioned at the top, whereas low-power brands garner favor when logos are placed lower. The study explores strategic logo placement using the concept of power metaphors. It suggests that powerful brands should choose top-of-packaging logo placement, aligning with the theory that associates higher placement with increased perceived power. This research holds significance for its implications for rebranding efforts, offering insights into strategic logo placement tailored to brand power dynamics.

   The results of the proposed model, as detailed in Table 5, support the observed relationship between logo placement and brand power, reinforcing the practical applicability of these findings in marketing and brand strategy. To further illustrate these concepts, Figure 7 presents visual examples specifically developed for this thesis.

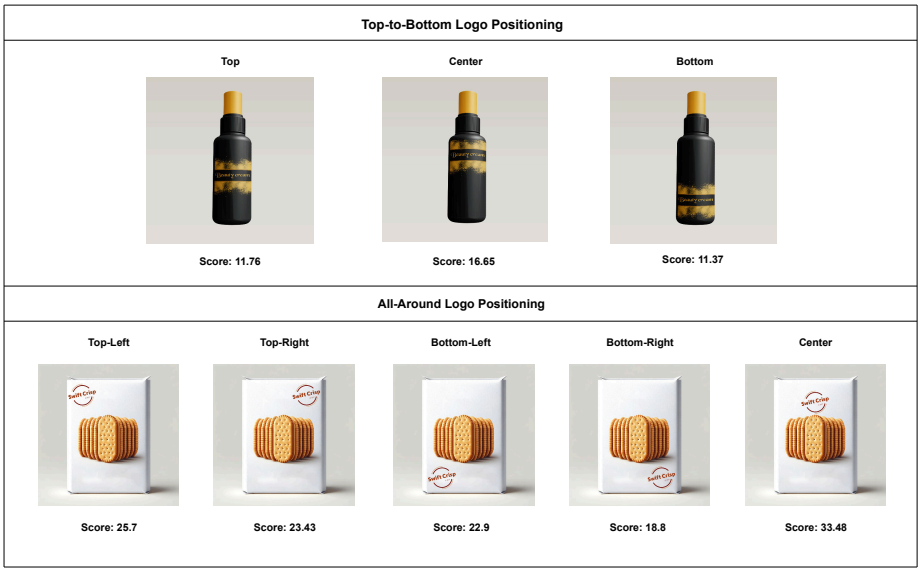


Figure 7: Testing images to demonstrate how logo position impacts brand attention. Top-to-bottom logo positioning (top) and all-around logo positioning (bottom).

Table 5: comparing the impact of top-to-bottom logo positioning, text vs. image, and square-round packaging orientation hypotheses on brand attention score.

Hypothesis	**Position**	**Mean**	**SE**
Top-to-Bottom Logo positioning	Down	28.89	5.19
	UP	34.05	5.64
Text vs Image	Image	31.71	4.61
	Text	37.23	4.62
Square-Round Packaging Orientation	Round	25.82	4.86
	Square	27.02	4.01

4.3.3. Proposed Hypothesises

Similarly, as outlined in the preceding section, the proposed brand attention method serves as a robust foundation for exploring brand marketing and visual analytics. Beyond the ongoing studies, we have introduced several new hypotheses that investigate aspects not extensively covered in the existing lit-

erature. These hypotheses represent unexplored territories as we strive for a comprehensive understanding of brand perception and consumer behavior. This exploration guides future psychophysical tasks, providing a framework for new investigations in the field.

- **Positioning of Brand Logos**

  Previous studies have mostly looked at where logos are placed, either at the top or bottom of packaging (Dong & Gleim, 2018; Riaz & Ghafoor, 2019) and explored the impact of these positions. However, there exists a notable gap in research, specifically addressing central logo placement or exploring the upper-left, bottom-left, upper-right, and lower-right areas. We conducted experiments to investigate the impact of these different placements on brand attention. the proposed model predicts that positioning the brand logo at the center of the packaging significantly enhances brand attention when compared to alternative positions, as outlined in Table 6. Additionally, the model proposed that upper placements generally are better than lower ones, with the upper-left being better than the upper-right, likely because people start reading from the top-left (Lautenbacher, 2012). Similarly, among the bottom positions, the bottom-left is more effective than the bottom-right.

- **Bold Distinction in Packaging**

  Many packaging designs incorporate bold text or objects, yet the impact of these elements on brand logo attention has been under-explored. To investigate the impact of emphasized elements other than the brand logo on consumer attention, we conducted experiments and analyzed the findings, as outlined in Table 6. In this experiment, we bolded and emphasized certain non-logo text or objects on the packaging to observe their effect on consumer attention. The proposed model predicts that when elements other than the brand logo are highlighted and emphasized on the packaging, it can lead to a reduction in brand attention.

- **Presence of Person in Packaging**

  It is well known that human faces instinctively capture visual attention (Cerf et al., 2007). However, the effect of this on brand attention within packaging contexts has not been thoroughly investigated. Exploring the impact of incorporating a person or face on packaging, the introduced model shows a considerable decrease in attention toward the brand. Results in Table 6 hint at a diminishing focus on brand logos when human faces are included in the packaging design.

- **Multi Packaging** The influence of presenting multiple packages of a brand in a single image has not been studied, particularly concerning its impact on brand logo visibility and attention. Our model predicts that images containing multiple packages simultaneously of a brand are more effective at absorbing brand attention compared to single-package images as demonstrated in 6. This could lie in the increased likelihood of detecting the brand logo when presented multiple times in diverse packaging formats.

- **Multi Objects in Packaging**

  The effect of featuring multiple objects in packaging design (for example, featuring either a single orange or multiple oranges within the packaging), as opposed to just one image, has not been thoroughly investigated in the context of brand logo attention. The proposed model predicts that packaging designs featuring multiple objects divert attention from the brand logo, as illustrated in Table 6. The presence of multiple objects in the visual field might cause distraction, leading to a diminished focus on the brand logo.

  This proposition finds support in our experimental results, indicating that simpler packaging designs, featuring only a single object, are more effective in maintaining higher brand logo attention.

- **Horizontal-Vertical Packaging Orientation**

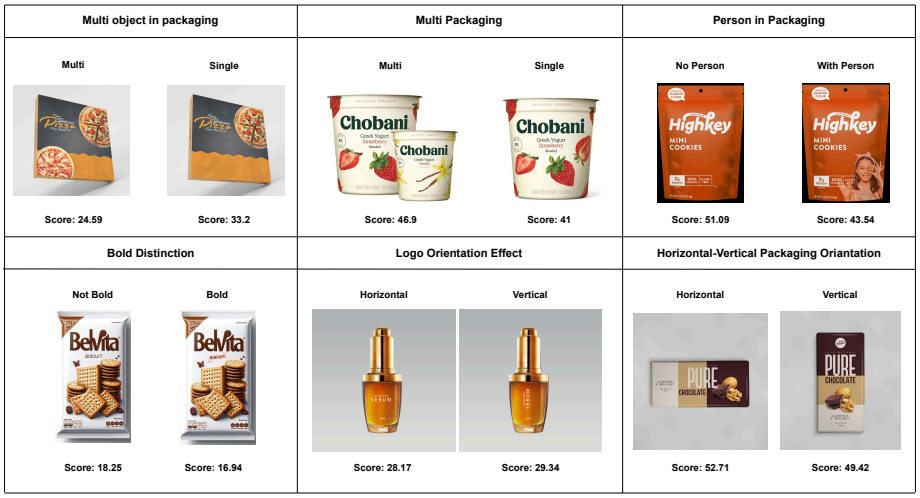


Figure 8: Sample images for assessing the proposed hypotheses: multi objects (top left), multiple packaging (top center), presence of a person (top right), bold distinction (bottom left), horizontal-vertical brand logo orientation (bottom center), and horizontal-vertical packaging orientation (bottom right).

In terms of the horizontal vs vertical packaging view, the proposed model predicts that horizontally oriented packaging enhances brand logo attention more effectively than vertically oriented packaging. This assumption could be based on the premise that horizontal orientation allows for a larger scale of the brand logo, making it more conspicuous and easily recognizable. The experimental results, as shown in Table 6, indicate a clear preference in brand attention for horizontal packaging designs.

- **Horizontal-Vertical Brand Logo Orientation**

  Next, we investigate whether the orientation of the logo itself, horizontally or vertically while keeping other packaging elements constant, would influence the logo attention score. Exploring different orientations, the proposed model suggests that vertical logos grab more attention. This could be from the extra time it takes to read and process vertical text, boosting engagement with the logo (Yu et al., 2010). Even when keeping the logo size consistent in both orientations, the proposed model indicates

a notable increase in brand attention for vertical logos, as outlined in Table 6.

Table 6: Comparing the impact of top-to-bottom logo positioning, all-around logo positioning, bold distinction, horizontal-vertical brand logo orientation, horizontal-vertical packaging orientation, presence of person, multi-object and multi packaging on brand attention score.

Hypothesis	**Position**	**Mean**	**SE**
Top-to-Bottom Logo positioning	Down	28.89	5.19
	UP	34.05	5.64
	Center	40.02	7.06
All-Around Logo Positioning	Down-Right	15.05	3.05
	Down-Left	18.8	3.41
	UP-Right	16.51	3.05
	UP-Left	20.24	3.34
	Center	24.92	4.12
Bold Distinction	Boldness	19.98	2.27
	Not Bold	21.1	2.35
Horizontal-Vertical Brand logo Orientation	Horizontal	29.91	4.05
	Vertical	34.54	4.8
Horizontal-Vertical Packaging Orientation	Vertical	27.92	4.92
	Horizontal	36.92	5.59
Person in Packaging	With Person	32.26	5.76
	No Person	36	6.16
Multi Object in Packaging	Multi	32.5	5
	One	40.95	5.29
Multi Packaging	Single	31.64	4.16
	Multi	39.52	4.73

- **Brand Logo Color** The influence of different colors of the packaging on attracting attention has been a subject of interest in marketing research (Raheem et al., 2014). However, comprehensive studies comparing a wide

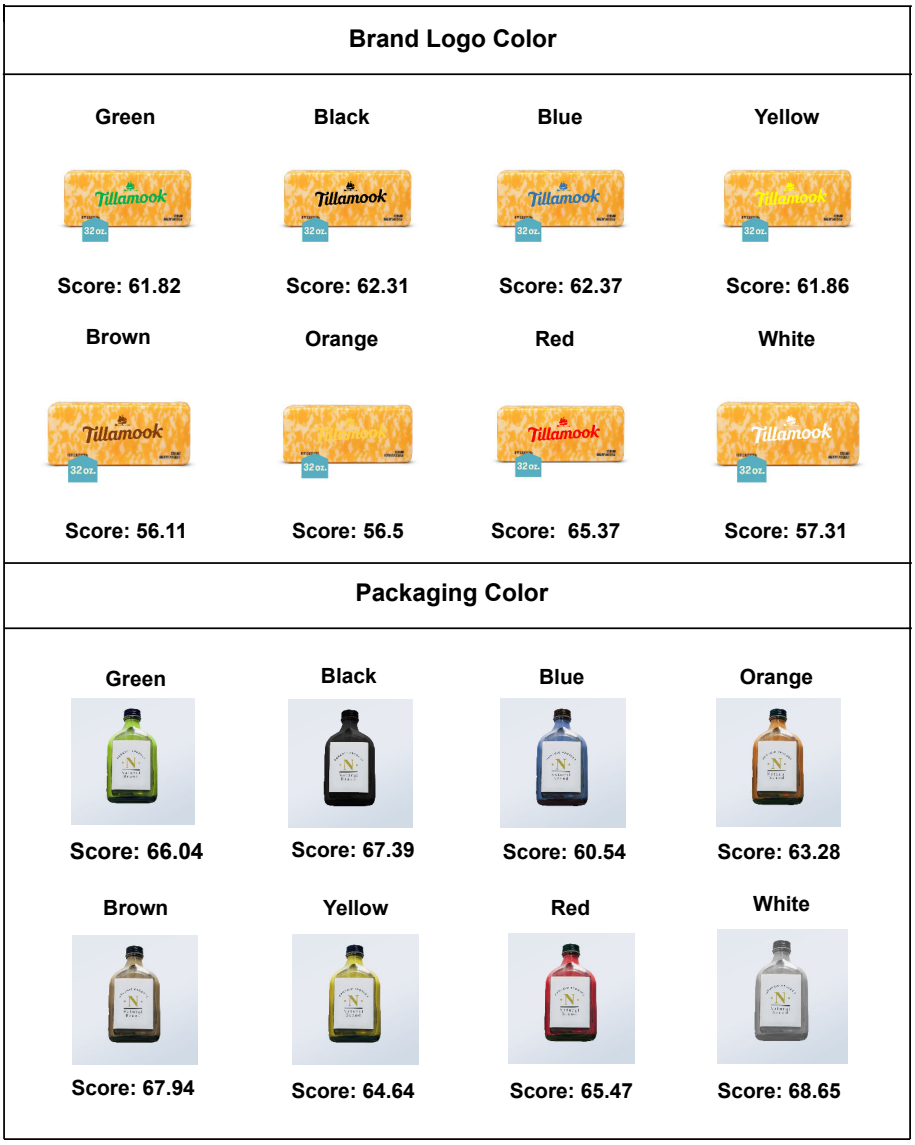


Figure 9: Visualization of the brand-logo color and packaging color influence on brand attention

range of colors, particularly in the context of brand logos, are less common. This study indicates that logo color influences brand attention, examining colors like white, brown, orange, yellow, green, blue, black, and red. The

results, detailed in a table 7, show that red is the most effective for drawing attention. This aligns with red's psychological associations with alertness and prominence (Singh, 2006), making it a strategic choice for logos to capture consumer interest.

- **Packaging Color**

  Packaging color has a substantial impact on brand visual attention, yet there exists limited research comparing various colors. We examined how different packaging colors, including white, brown, orange, yellow, green, blue, black, and red affect the visibility of the brand logo. the proposed model predicts that packaging color greatly affects brand attention, with less intense, warmer, and simpler colors being more effective. The achieved results, as shown in the table 7, indicate that white as a packaging color enhances brand attention the most, likely because its neutral and unobtrusive nature allows the logo to be more noticeable.

Figures 7 , 8 and 9 present samples from the brand attention dataset, serving as empirical evidence for the evaluation of our proposed hypotheses.

5. Conclusion

The importance of logos within packaging emerges as an influential visual cue, profoundly shaping consumer perception and promoting brand recognition. This paper introduces a module specifically designed to model human attention to brands in packaging. The module comprises three main components: fine-tuned Yolov8 logo detection, a novel CNN-Transformer-based saliency map prediction model that surpasses existing models in predicting visual attention, and brand attention score. To verify the proposed method, we employed existing psychophysical studies. The proposed method was in line with all previous studies on brand attention, which shows the robustness of the suggested brand attention score. Next, by utilizing the capabilities of this module, it becomes possible to model human visual attention to brand logos within the packaging, opening new opportunities for testing unexplored hypotheses in this field.

Table 7: Comparing the impact of packaging color and brand logo color on brand attention score

Hypothesis	**Position**	**Mean**	**SE**
Packaging Color	black	36.82	5.65
	Brown	37.85	5.62
	Orange	37.46	5.46
	Yellow	37.45	5.61
	Green	36.38	5.55
	Blue	37.51	5.66
	Red	38.23	5.73
	White	40.84	5.89
Brand Logo Color	White	31.08	4.33
	Brown	34.54	4.4
	Orange	33.2	4.63
	Yellow	32.93	4.66
	Green	32.16	4.93
	Blue	33.13	4.87
	Black	36.56	4.7
	Red	37.44	4.79

Therefore, using the brand attention score, seven new hypothesis, that have never been studied before, are examined. For example, our model suggests that positioning the brand logo at the center or upper left of the packaging increases its visibility. Additionally, the model predicts that employing a red color for the brand logo and a white color for the packaging will enhance the brand's attention score.

The practical utility of the proposed module is especially significant for designers in the fields of advertising and packaging. By quantifying how customers perceive the prominence of a brand, our tool empowers designers to make data-driven decisions regarding logo placement and packaging design. Moreover, the

versatility of our module goes beyond brand logos, as designers can use the saliency map and our algorithm to calculate attention scores for any object or text in an image, as long as they designate or select the bounding boxes. This capability enhances the module's applicability, rendering it a valuable tool across various design and marketing applications. Ultimately, this research not only enriches academic discussions on brand visibility but also provides pragmatic tools for enhancing consumer engagement in the dynamic landscape of advertising and packaging. As the future work, the proposed method can be utilized as a video saliency predictor, considering both spatial and temporal features to enhance its applicability in dynamic contexts. Additionally, the module has the potential to serve as a discrimination network within packaging generative networks, guiding the optimization of logo placement and packaging design. These advancements promise to further revolutionize the field, offering novel insights and tools for effective brand visualization in an ever-evolving digital landscape.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Alvino, L., Constantinides, E., & van der Lubbe, R. H. (2021). Consumer neuroscience: Attentional preferences for wine labeling reflected in the posterior contralateral negativity. *Frontiers in psychology*, *12*, 688713.

Ampuero, O., & Vila, N. (2006). Consumer perception of product packaging. *Journal of Consumer Marketing*, *23*, 100–112. doi:10.1108/07363760610655032.

Ares, G., & Deliza, R. (2010). Studying the influence of package shape and colour on consumer expectations of milk desserts using word association

and conjoint analysis. *Food Quality and Preference*, *21*, 930–937. doi:10.1016/j.foodqual.2010.03.006.

Aydemir, B., Hoffstetter, L., Zhang, T., Salzmann, M., & Süsstrunk, S. (2023). Tempsal-uncovering temporal information for deep saliency prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6461–6470).

Bloch, P. (1995). Seeking the ideal form: Product design and consumer response. *Journal of Marketing*, *59*, 16–29. doi:10.2307/1252116.

Bochkovskiy, A., Wang, C. Y., & Liao, H. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, .

Boia, R., Florea, C., & Florea, L. (2015). Elliptical asift agglomeration in class prototype for logo detection. In *Proceedings of the British Machine Vision Conference* (pp. 115.1–115.12).

Borji, A., & Itti, L. (2012). State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, *35*. doi:10.1109/TPAMI.2012.89.

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

Cao, G., Tang, Q., & Jo, K.-h. (2020). Aggregated deep saliency prediction by self-attention network. In *Intelligent Computing Methodologies: 16th International Conference, ICIC 2020, Bari, Italy, October 2–5, 2020, Proceedings, Part III 16* (pp. 87–97). Springer.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European Conference on Computer Vision* (pp. 213–229).

Cerf, M., Harel, J., Einhaeuser, W., & Koch, C. (2007). Predicting human gaze using low-level saliency combined with face detection. In

J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc. volume 20. URL: https://proceedings.neurips.cc/paper_files/paper/2007/file/708f3cf8100d5e71834b1db77dfa15d6-Paper.pdf.

Che, Z., Borji, A., Zhai, G., Min, X., Guo, G., & Callet, P. L. (2020). Why is gaze influenced by image transformations? dataset and model. *IEEE Transactions on Image Processing*, *29*, 2287–2300.

Dong, R., & Gleim, M. (2018). High or low: The impact of brand logo location on consumers product perceptions. *Food Quality and Preference*, *69*. doi:10.1016/j.foodqual.2018.05.003.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., & et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *2021 International Conference on Learning Representations (ICLR)*.

Droste, R., Jiao, J., & Noble, J. A. (2020). Unified image and video saliency modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16* (pp. 419–435). Springer.

Girard, T., Anitsal, M. M., & Anitsal, I. (2013). The role of logos in building brand awareness and performance: Implications for entrepreneurs. *The Entrepreneurial Executive*, *18*, 7.

Girshick, R. B. (2015). Fast r-cnn. In *IEEE International Conference on Computer Vision* (pp. 1440–1448).

Girshick, R. B., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 580–587).

Gofman, A., Moskowitz, H., Fyrbjork, J., Moskowitz, D., & Mets, T. (2009). Extending rule developing experimentation to perception of food packages

with eye tracking. *The Open Food Science Journal*, *3*, 66–78. doi:10.2174/1874256400903010066.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778).

Hoi, S. C. H., Wu, X., Liu, H., Wu, Y., Wang, H., Xue, H., & Wu, Q. (2015). Logo-net: Large-scale deep logo detection and brand recognition with deep region-based convolutional networks. *arXiv preprint arXiv:1511.02462*, .

Hou, Q., Min, W., Wang, J., Hou, S., Zheng, Y., & Jiang, S. (2021). Foodlogodet-1500: A dataset for large-scale food logo detection via multi-scale feature decoupling network. In *Proceedings of the 29th ACM International Conference on Multimedia* (pp. 4670–4679).

Hou, S., Li, J., Min, W., Hou, Q., Zhao, Y., Zheng, Y., & Jiang, S. (2023). Deep learning for logo detection: A survey. *ACM Transactions on Multimedia Computing, Communications and Applications*, *20*, 1–23.

Hubert, M., Baecke, S., & Kenning, P. (2008). What they see is what they get? an fmri-study on neural correlates of attractive packaging. *Journal of Consumer Behaviour*, *7*, 342 – 359. doi:10.1002/cb.256.

Jia, S., & Bruce, N. D. (2020). Eml-net: An expandable multi-layer network for saliency prediction. *Image and vision computing*, *95*, 103887.

Jiang, L., Li, Y., Li, S., Xu, M., Lei, S., Guo, Y., & Huang, B. (2022). Does text attract attention on e-commerce images: A novel saliency prediction dataset and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2088–2097).

Jiang, M., Huang, S., Duan, J., & Zhao, Q. (2015). Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1072–1080).

Jin, X., Su, W., Zhang, R., He, Y., & Xue, H. (2020). The open brands dataset: Unified brand detection and recognition at scale. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4387–4391).

Jocher, G., Chaurasia, A., & Qiu, J. (). Yolo by ultralytics. URL: `https://github.com/ultralytics/ultralytics`.

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

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision* (pp. 2106–2113). IEEE.

Kou, Q., Liu, R., Lv, C., Jiang, H., & Cheng, D. (2023). Advertising image saliency prediction method based on score level fusion. *IEEE Access*, *11*, 8455–8466. doi:`10.1109/ACCESS.2023.3236807`.

Krishna, A., Cian, L., & Aydınoğlu, N. (2017). Sensory aspects of package design. *Journal of Retailing*, *93*, 43–54. doi:`10.1016/j.jretai.2016.12.002`.

Kroner, A., Senden, M., Driessens, K., & Goebel, R. (2020). Contextual encoder–decoder network for visual saliency prediction. *Neural Networks*, *129*, 261–270.

Kümmerer, M., Wallis, T. S., & Bethge, M. (2016). Deepgaze ii: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*, .

Lautenbacher, O. P. (2012). From still pictures to moving pictures. *Eye-tracking in Audiovisual Translation*, (pp. 135–155).

Li, Y., Shi, Q., Deng, J., & Su, F. (2017). Graphic logo detection with deep region-based convolutional networks. In *IEEE Visual Communications and Image Processing* (pp. 1–4).

Liang, S., Liu, R., & Qian, J. (2021). Fixation prediction for advertising images: Dataset and benchmark. *Journal of Visual Communication and Image Representation*, *81*, 103356.

Liao, M., & et al. (2022). Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*, 919–931.

Lin, T. Y., Dollar, P., Girshick, R. B., He, K., Hariharan, B., & Belongie, S. J. (2017). Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 936–944).

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).

Lou, J., & et al. (2022). Transalnet: Towards perceptually relevant visual saliency prediction. *Neurocomputing*, *494*, 455–467.

Lévêque, L., & Liu, H. (2019). An eye-tracking database of video advertising. In *2019 IEEE International Conference on Image Processing (ICIP)* (pp. 425–429). doi:10.1109/ICIP.2019.8802989.

Maynard, O., McClernon, F., Oliver, J., & Munafò, M. (2018). Using neuroscience to inform tobacco control policy. *Nicotine & tobacco research : official journal of the Society for Research on Nicotine and Tobacco*, *21*. doi:10.1093/ntr/nty057.

Meng, Y., Hou, S., Wang, J., Jia, W., Zheng, Y., & Karim, A. (2021). An adaptive representation algorithm for multi-scale logo detection. *Displays*, *70*, 102090.

Méndez, J., Oubiña, J., & Rubio, N. (2011). The relative importance of brand-packaging, price and taste in affecting brand preferences. *British Food Journal*, *113*, 1229–1251. doi:10.1108/00070701111177665.

Nagel, R., Reutskaja, E., Camerer, C., & Rangel, A. (2011). Search dynamics in consumer choice under time pressure: An eye-tracking study. *American Economic Review*, *101*, 900–926. doi:10.1257/aer.101.2.900.

Oliveira, G., Frazao, X., Pimentel, A., & Ribeiro, B. (2016). Automatic graphic logo detection via fast region-based convolutional networks. In *International Joint Conference on Neural Networks* (pp. 985–991).

Otterbring, T., Shams, P., Wästlund, E., & Gustafsson, A. (2013). Left isn't always right: Placement of pictorial and textual package elements. *British Food Journal*, *115*. doi:10.1108/BFJ-08-2011-0208.

Paleček, K., & Chaloupka, J. (2021). Logo detection and identification in system for audio-visual broadcast transcription. In *2021 44th International Conference on Telecommunications and Signal Processing (TSP)* (pp. 357–360).

Pertzov, Y., Avidan, G., & Zohary, E. (2009). Accumulation of visual information across multiple fixations. *Journal of Vision*, *9*, 2–2. URL: https://doi.org/10.1167/9.10.2. doi:10.1167/9.10.2.

Piqueras-Fiszman, B., Velasco, C., Salgado-Montejo, A., & Spence, C. (2013). Using combined eye tracking and word association in order to assess novel packaging solutions: A case study involving jam jars. *Food Quality and Preference*, *28*, 328–338. URL: https://www.sciencedirect.com/science/article/pii/S0950329312002005. doi:https://doi.org/10.1016/j.foodqual.2012.10.006.

Raheem, A. R., Vishnu, P., & Ahmed, A. M. (2014). Impact of product packaging on consumer's buying behavior. *European journal of scientific research*, *122*, 125–134.

Rebollar, R., Lidón, I., Martin Vallejo, F., & Puebla, M. (2015). The identification of viewing patterns of chocolate snack packages using eye-tracking

techniques. *Food Quality and Preference*, *39*, 251–258. doi:10.1016/j.foodqual.2014.08.002.

Redmon, J., Divvala, S. K., Girshick, R. B., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 779–788).

Redmon, J., & Farhadi, A. (2017). Yolo9000: Better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6517–6525).

Ren, S., He, K., Girshick, R. B., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*, 1137–1149.

Rettie, R., & Brewer, C. (2000). The verbal and visual components of package design. *Journal of Product & Brand Management*, *9*. doi:10.1108/10610420010316339.

Revaud, J., Douze, M., & Schmid, C. (2012). Correlation-based burstiness for logo retrieval. In *Proceedings of the 20th ACM International Conference on Multimedia* (pp. 965–968).

Riaz, T., & Ghafoor, M. (2019). Strategic logo placement on packaging - using conceptual metaphors of power in packaging – evidence from pakistan. *Procedia Computer Science*, *158*, 582–589. doi:10.1016/j.procs.2019.09.092.

Sahbi, H., Ballan, L., Serra, G., & Bimbo, A. (2013). Context-dependent logo matching and recognition. *IEEE Transactions on Image Processing*, *22*, 1018–1031.

Shen, Z., & et al. (2021). Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*.

Shukla, P., Singh, J., & Wang, W. (2022). The influence of creative packaging design on customer motivation to process and purchase decisions. *Journal of Business Research*, *147*, 338–347. doi:10.1016/j.jbusres.2022.04.026.

Singh, S. (2006). Impact of color on marketing. *Management decision*, *44*, 783–789.

Stewart, B. (1995). *Packaging as an effective marketing tool*. CRC Press.

Velazquez, D. A., Gonfaus, J. M., Rodríguez, P., Roca, F. X., Ozawa, S., & Gonzalez, J. (2021). Logo detection with no priors. *IEEE Access*, *9*, 106 998–107 011.

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2023). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wang, J., Min, W., Hou, S., Ma, S., Zheng, Y., & Jiang, S. (2022). Logodet3k: A large-scale image dataset for logo detection. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *18*, 1–19.

Yu, D., Park, H., Gerold, D., & Legge, G. E. (2010). Comparing reading speed for horizontal and vertical english text. *Journal of vision*, *10*, 21–21.