Deciphering Brand Identity from package: Visual Feature Analysis through Convolutional Neural Networks

Asaya Shimojo and Shoichi Uratani
KONICA MINOLTA, Inc.
Tokyo, Japan
email: asaya.shimojo@konicaminolta.com, shiyouichi.uratani@konicaminolta.com

***Abstract*— Many brands traditionally rely on qualitative methods to design their product packaging, leaving uncertainties about the consistency of brand identity across different packages. This study leverages machine learning to quantitatively extract and analyze design elements that resonate with consumers' perception of brand identity. Specifically, we employed Grad-CAM, an interpretative method for Convolutional Neural Networks (CNNs), to identify crucial visual features—termed Visual Identities—within the middle layers of a model trained on specific brand package images. These features were analyzed to determine their influence on package classification and their alignment with human perception of brand identity. Our findings demonstrate that the machine learning approach approximates human perception closely, providing a novel quantitative method to enhance and maintain brand identity. Additionally, we quantified the contribution of each identified Visual Identity to overall brand recognition, offering a more systematic approach to understanding and preserving a brand's distinctiveness that has traditionally been handled qualitatively.**

Keywords- Grad-CAM; Brand Identity; Visual Identity; Package Design; Consumer Recognition.

I. INTRODUCTION

A product's packaging design reflects its Brand Identity (BI), significantly impacting consumer perception. Previous research has explored which elements of package design contribute to BI. However, this evaluation process can burden evaluators and vary individually, focusing on limited and abstract aspects like color tones and fonts, reducing its practicality. The purpose of this study is to utilize the latest machine learning technologies to automatically extract design elements from packaging and quantitatively analyze their impact on consumer brand recognition. This approach assists brand managers in developing more effective packaging design strategies.

A. The Aim of This Study

The study visualized intermediate model layers to extract packaging design elements. These elements are then compared with factors that humans consider indicative of BI, to validate the relevance of the former. Through this process, we develop a simple yet precise method that compares human-detected BI features with those identified by the model.

B. Related Works

BI is the embodiment of a company's values and characteristics [1]. Specifically, BI is composed of visual elements such as logos, colors, and design styles [2]. Thus, design elements that visualize the values and concepts of a Brand and symbolize the brand are called Visual Identity (VI). For example, package design influences consumers' emotions and perceptions through its VI, including its colors, shapes, materials, logos, and text [2]. It has also been found that the visual attractiveness of a package draws consumers' attention and forms a favorable impression of the product, as well as a high perception of the product's quality and value [3]. Therefore, the VI of the packages is considered to play a central role in BI communication [4]. Thus, most brands need to express their identity through their VI and keep consumers consistently identifying with the brand. In fact, VI consistency has been found to improve purchase intent and brand loyalty [5].

On the other hand, the specific visual elements used to represent the brand need to change with time and trends. However, this idea is inconsistent with maintaining consistency in the VI. In contrast, when interviews were conducted on the consistency of the BI with successive changes in the VI, art directors and other professionals had a narrow range of acceptance of consistency, and some consumers with a high aesthetic sense were also sensitive to changes in the VI [6]. However, while art directors emphasized the complexity of the VI construct, it was also clear that VIs are changed based on preference and emotion, making it a very inadequate method for capturing VIs in an exhaustive and quantitative manner.

In addition, quantitative studies have examined BI by focusing on logos [7][8]. However, VI requires an exhaustive study because other factors such as object edges and color contrasts are also considered to be involved in its composition [9].

C. Theoretical Background

In recent years, AI technology has advanced to the point where it is now possible to build CNN models that learn

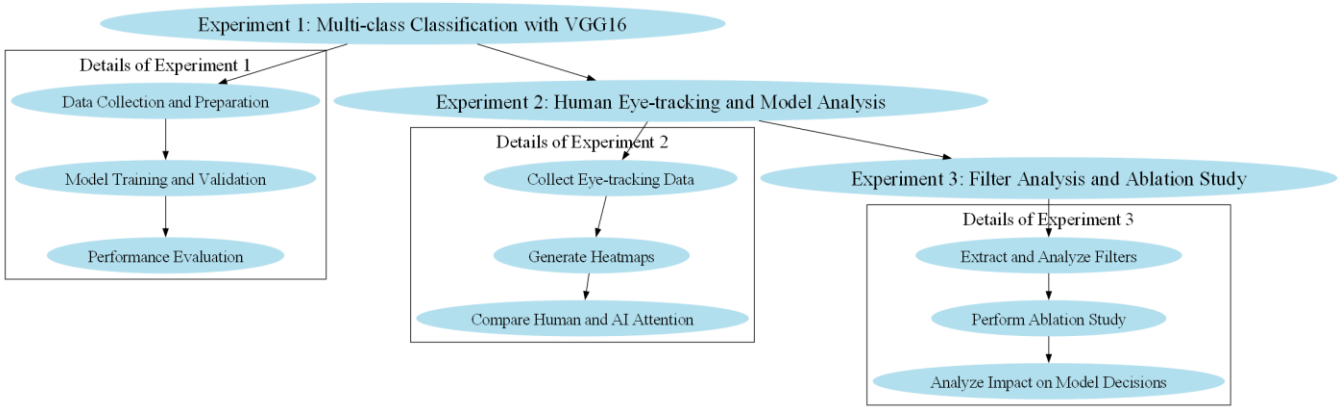

Figure 1. Flowchart of three experiments' procedures in this study.

package images of specific brands and classify whether the brand is that Brand. In fact, brand identification by CNN has been verified on brand logos and fashion show runway photos [10][11].

Moreover, a technique called Grad-CAM has also been established to visualize where in the image the model focused its attention in making decisions in classification [12]. Specifically, the gradient information for the feature map of the final convolutional layer associated with a particular output class can be used to present important regions in the input image as a heat map. The visualization of the heatmap facilitates the interpretation of the model, as this output indicates the image regions that the CNN focuses on when making a particular decision [13].

However, this method only visualizes the decision-making process of the model and does not directly reveal where humans, as consumers, focus their attention on the package to identify the brand. In fact, the comparison of the AI model's point of attention with the human visual area of attention is considered important but has not yet been validated [14]. Therefore, this study will examine the extent to which the visual factors that humans and machines focus on when identifying the packaging design of a particular brand coincide by collecting human eye-tracking data and comparing this with the results of a deep learning model using Grad-CAM. This examination would reveal the usefulness of the features automatically extracted by the model.

In summary, this study has two major contributions. One is to propose a methodology for automatically extracting the features used to determine whether a model is the relevant brand or not by training the model on the package data of a specific brand and visualizing the middle layer of the model using the above-mentioned techniques. This would enable quantitative VIs management for package design and more efficient communication of BI. The second is to verify the degree to which the trajectory of the human gaze matches the visualization of the middle layer of the learned model. In addition to validating the accuracy of the methodology, this could reveal the degree of agreement between the model's point of interest and the human visual area of interest. To realize these two contributions, three experiments were conducted in this study; their appearance was shown in Figure 1.

Specifically, in Section II, a multi-class classification model and Grad-CAM was developed to identify key design elements in package designs. In Section III, human eye-tracking data was compared with Grad-CAM visualizations to evaluate the consistency between human and machine recognition. In Section IV, the CNN model's processing of visual information was analyzed to identify design elements contributing to Brand A's BI. Finally, in Section V, we discussed the study's findings and limitations.

II. EXPERIMENT 1

This experiment built a multi-class classification model by fine-tuning VGG16 with package design data from brands A through E. Subsequently, the intermediate layers of the model were visualized using Grad-CAM to identify which design elements were instrumental in distinguishing between the package designs of the five brands.

A. Brand Selection and Dataset Construction

This experiment used a self-created dataset comprising approximately 6,000 images, specifically focusing on the packaging of brands with a long history and established BI. Based on scale [15], five writing instrument brands were selected, categorized into two luxury brands (Brands A and B), two general consumer brands (Brands C and D), and one lesser-known brand (Brand E). Figure 2 shows a thumbnail of Brand A's packaging design images of the dataset.

Figure 2. A thumbnail of some packaging design images of the dataset.

This selection was aimed at testing the generalizability across a broad consumer base. By conducting a multi-class classification that encompasses a variety of brands, the model could have the validity of this approach for tail brand with small sample sizes (e.g., Brand E). In the data collection process, considerations were made for copyright issues, acknowledging that the data has been publicly available for over 70 years. If the brand name was written on the package, it was masked in gray to hide it.

B. Construction of Classification Model based on VGG16

This study used VGG16, pre-trained on ImageNet, to develop a multi-class classifier capable of identifying package designs from Brands A through E. The model was trained with 70% of the dataset images, adjusting the output size of the final layer to 5 to classify each image into one of five brands. The remaining 30% of the dataset was used for validation. The training and validation processes were conducted 2,000 times, shuffling the data each time, and using a batch size of 32, with the number of epochs set to 10.

Performance was evaluated by varying the neuron counts in the model's dense layer from 51 (10% of full capacity) to 512 (100%). The model showed optimal performance with 512 neurons, achieving a validation accuracy of 91.18%. Models with 384 neurons also performed well, achieving an accuracy of 87.47%. However, models with fewer neurons—256, 128, and 51—all showed validation accuracies below 80%. Moreover, when the training dataset was reduced, the model with 512 neurons showed varying accuracies: 62% with 600 images (10% of full data), 77% with 1,500 images (25%), 82% with 3,000 images (50%), and 89% with 4,500 images (75%). A training dataset of at least 4,500 images was considered necessary for optimal performance. These performance changes are shown in Figure 3.

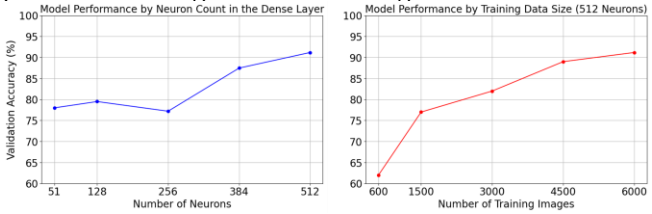

Figure 3. An Impact of neuron count or training data size on model accuracy.

This study used the largest model size and number of samples available (100% for both) because accuracy was given priority.

C. Filter Visualization with Grad-CAM

In the model, image features were transformed into a feature vector of 4,096 dimensions through the convolution and all the coupling layers. This made it difficult to directly understand which elements of the package design affected the classification results. Therefore, Grad-CAM was used to visualize the filter output of "block5_conv3," the layer closest to the output of the convolution layer (Conv2D), to determine which image regions had the most influence on the classification decision. As a result of evaluating the contribution of each layer to brand identification in our ablation study, it was found that the validation accuracy decreased by over 70% when this layer was disabled. The result indicated the layer's critical role in accurate brand recognition. Consequently, we determined that it was important to visualize this layer to identify VIs that affect brand recognition. For instance, the following package image showed a visualization of the areas in the image that contributed to being classified as Brand A (Figure 4).

Figure 4. Brand A classification contribution area acquired by Grad-CAM. "Predicted: 0" meant that the image was determined to be Brand A.

This gave the model's predictive basis in a form that was intuitively understandable to humans. In the next section, we clarified the extent to which the results of this visualization correspond to the areas that humans pay attention to when recognizing brands.

III. EXPERIMENT 2

This experiment investigated similarities that exist between human visual perception of package designs and the results of image analysis by the machine learning model obtained in the previous section. Specifically, human eye-tracking data were collected and compared with the features of the model visualized by Grad-CAM.

A. Method

Six adults (2 females, 4 males; Age: 39.8 ± 7.9 years) participated in the experiment. The participants had normal visual acuity. They were presented with 10 packages each of five brands and learned their VIs. The participants' heads were then fixed by a chinrest placed approximately 60 *cm* from the monitor screen. Images were displayed across the entire monitor screen. Then, one image from the image set was presented at random, and the participant gazed at it for 5 seconds. There were 20 images totally in the image set, which consisted of 10 packages for each brand except the package used for the learning.

The participants' eye movements were recorded while gazing at the package images. To record the trajectory, a webcam (ELECOM UCAM-C750FBBK; resolution 1920 × 1080 *px*, frame rate 30 *FPS*, angle of view 66 degrees, 1/4-*inch* CMOS sensor), a monitor (I-O DATA KH240V-B; 23.8-*inch* wide, resolution 1920 × 1080 *px* resolution, ADS panel, brightness 250 cd/m^2, response time 5 *ms*), and *GazeRecorder* [16], a line-of-sight measurement software. The experiment room brightness was kept constant at 500 *lux*.

After gazing at each image, participants responded with a confidence level ranging from 0 % (definitely not a Brand A

package) to 100 % (definitely a Brand A package) that they thought each image was a Brand A package.

B. Results and Discussion

Comparison both heat maps (one from human eye-tracking and one from Grad-CAM) quantitatively assessed eye-tracking data with an average confidence level of at least 70 % were used to perform the analysis only on package images that the participants were confident were Brand A.

This analysis used the Jaccard Index to investigate the similarity between two different image generation processes: a heatmap based on human visual tracking and a computerized Grad-CAM heatmap. The Jaccard index is a measure of similarity between sets, with values varying from 0 to 1. Here, 1 indicates that the two heatmaps match perfectly, while 0 means no overlap at all.

Specifically, heatmaps based on human visual tracking were generated from viewpoint data as participants viewed each package. The visual tracking data was processed using the built-in systems of *GazeRecorder*. In contrast, the Grad-CAM heatmaps were obtained by analyzing the same images using the Python *cv2* library within a specified CNN model. Each heatmap was visualized as a color intensity map indicating areas of visual attention. In these heatmaps, red indicated the most focused areas of attention, and blue indicated the least focused areas.

These heat maps were then overlaid at the pixel level to quantitatively assess the size and distribution of commonly noticed regions of interest. The process involved loading the images in grayscale, resizing, and binarizing them using a Python script. The intersection (common area) and union (total area) of these images were then computed to derive the Jaccard index. Figure 5 was shown as an example of the comparison.

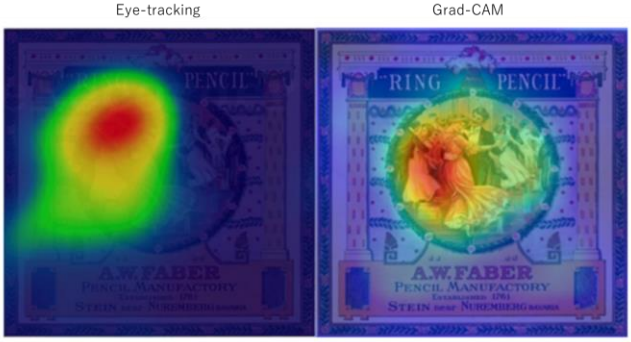

Figure 5. An example of a comparison of the features of the middle layer of the visualized CNN and the gaze trajectory.

The analysis resulted in a mean value of 0.32 (*SD*=0.12) for the obtained Jaccard index. This value indicated that although a certain degree of similarity was observed between the heat maps, differences existed in several regions. However, it was also visually readable that in most package pairs, the areas of highest attention matched.

Thus, the results suggest that there is a partial match between the AI-generated attention maps and the areas of human attention focus, supporting the possibility that the extent to which human visual attention and machine learning models' judgments are consistent. The analysis was conducted to identify similarities and differences between the human recognition process and the machine learning model's ability to recognize patterns. Note that only human AI's visual processing algorithms may be somewhat like the human visual recognition process.

IV. EXPERIMENT 3

This experiment demonstrated that the CNN model processes visual information in a manner somewhat analogous to human perception. However, delineating the specific features the model recognizes and uses for classification remains challenging. Therefore, we focused on identifying individual design elements and assessed their contribution to the BI of Brand A.

A. Extraction of Brand classification contribution filters

This experiment aimed to understand how the model extracts and interprets image features by visualizing filter activation in the middle layer of the CNN model fine-tuned in Experiment 1. Specifically, filter weights were extracted from the intermediate layer, "block5_conv3," and, activation maps for each filter were generated. The Python *TensorFlow* and *Keras* libraries were used to generate images starting with random noise for each filter and iteratively update the images in the direction of maximizing filter activation.

This analysis resulted in 512 output filters, of which 40 were significantly effective in classifying Brand A. Figure 6 showed one such filter—specifically, the 110th filter—which was particularly interpretable. Upon visualizing the activation map using the jet colormap, areas with the highest activations corresponded remarkably to objects, such as pencils, fountain pens, and geometric shapes. This suggests that the model might prioritize sharp and defined tips of objects, which are characteristic elements in the VI of Brand A. Indeed, this distinct emphasis on pointed features was visually confirmed to be more pronounced in Brand A's packaging compared to other brands' ones, aligning with the brand's distinctive aesthetic attributes. Other contributing filters also were described below.

This process allowed us to understand what is involved in the identification of Brand A by extracting the filter used by the CNN for discrimination and overlaying it on the original image in the form of a heat map to see where it is applied. In other words, the physical characteristics of Brand A, or VI, since that is where it is used to recognize Brand A. This process was useful in understanding how the model identified a specific VI and how it contributes to the identification of the brand.

B. Ablation Study for the Contribution Ratio for each VI

Based on the results of the filter visualization, an ablation study was conducted to better understand its functional importance. In this study, the weights of each

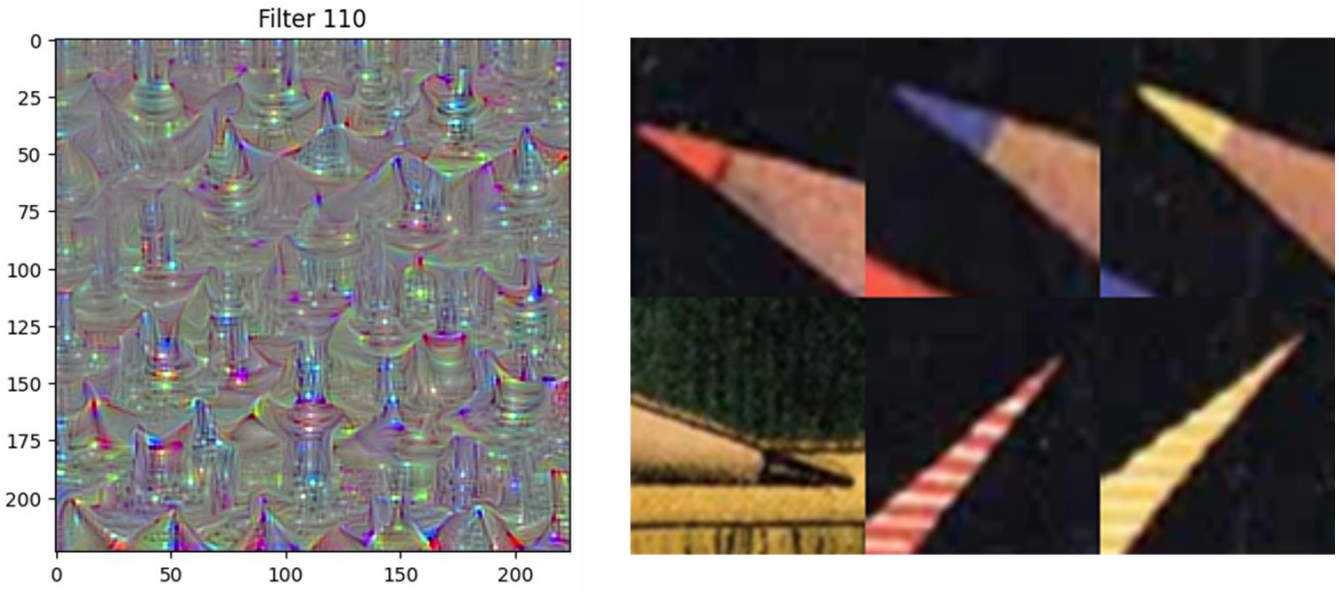

Figure 4. Activation Map of the 110th Filter Highlighting Response to Sharp Object Tips.

filter in the middle tier were individually set to zero, and the impact of these changes on the model's overall testing accuracy was systematically evaluated. By disabling each filter, we quantitatively analyzed the extent to which the filter contributed to the model's ability to make decisions.

The analysis showed that the filters' contribution ranged from a maximum of 25% to a minimum of 0.12%. Of note was the 110th filter, which, when disabled, resulted in a 13% reduction in the ability to identify Brand A. This indicates that the sixth filter plays a particularly important role in extracting the visual features of Brand A. Other contributing points to Brand A recognition detected were the strength of the curve of the product (the 93rd filter; 11.2% contribution to Brand A recognition), the light reflectance of the metal body (the 103rd filter; 8.9%), and the strength of the color contrast between the background and the product (the 106th filter; 8.1%).

Findings such as these are valuable in clarifying the key visual elements in brand identification and understanding how they affect the performance of the model. Therefore, the process of extracting VIs from the package and calculating their impact on BI could be automated.

V. CONCLUSION AND FUTURE WORK

This study used machine learning to quantify the impact of packaging design on BI. Specifically, CNN and Grad-CAM were used to explore the extent to which machine feature extraction is consistent with human BI recognition. The results show that machine learning models can effectively identify and highlight important design elements in brand recognition.

It was confirmed that the CNN model can identify brands based on specific elements of the package design, such as hue and logo, but also on detailed representations, such as the edges in an illustration. This showed that machines can capture important elements of VI and use them to make classification decisions.

Furthermore, visualization with Grad-CAM reveals that the areas that the model focuses on coincide with the areas that humans focus on when recognizing BI. This suggested that machine learning models may be able to mimic the human recognition process, indicating the existence of common ground between human and machine recognition.

The study also provided a method for quantifying the impact of individual design elements on brand recognition. This would enable brand managers to understand the specific impact of each packaging design element on BI and make more strategic design decisions.

A. Limitation and Future works

While Grad-CAM effectively highlights crucial areas within an image, it could focus on regions with high visual saliency, potentially overlooking subtler yet important features that contribute to the overall understanding of the image [12]. This phenomenon, known as the saliency bias, would raise concerns about the comprehensiveness of visual explanations provided by convolutional networks.

To address this limitation, integrating attention mechanisms that adjust focus based on the context of the entire image rather than just visual salient features has gained traction. For instance, the Transformer relies entirely on an attention mechanism, discarding the need for recurrent layers [17]. This model dynamically weights the influence of different parts of the input data, which can be particularly beneficial for understanding complex images in a more human-like manner. Furthermore, multi-dimensional scaling techniques can complement attention mechanisms by reducing high-dimensional data into a space where relationships between features are preserved, allowing for a clearer visualization of how features interact and contribute to the model's decisions [18].

These methodologies could avoid the saliency bias, which helps identify subtle yet crucial patterns that might be missed by traditional saliency-based approaches. They not only perform well but also align more closely with human cognitive processes, potentially making machine learning tools more intuitive and trustworthy for users in real-world applications.

Moreover, Experiment 2 showed that feature extraction using Grad-CAM is consistent with human visual regions of interest. This suggests that the proposed model captures visual elements like the human brand recognition process. However, due to limitations in the number of participants, further validation is required before these results can be widely generalized. It is essential that future research extensively test the generalizability and effectiveness of the proposed model through a variety of brand categories and a large set of experiments.

In addition, increasing the diversity and comprehensiveness of the data set would allow us to capture a broader range of visual elements of BI. This would include data from brands from different time periods and cultural backgrounds, which would enhance the generalizability of the model and provide a more generic quantitative method for VI management.

B. Social Contribution

This study would contribute to strategic improvements in package design in that the visibility elements of the BI could be automatically extracted. Specifically, it would help package design to maintain consistency in BI and differentiate it from other brands. As a result, brands would be able to manage their VI in package design more strategically, leading to stronger relationships with consumers and increased brand value.

REFERENCES

[1] Jean-Noël. Kapferer, "*Strategic Brand Management*." London:Kogan Page. 1992.

[2] D. A. Aaker, "*Building Strong Brand*." New York: The Free Press—A division of Simon & Schuster Inc. 1996.

[3] M. E. H. Creusen, and J. P. L. Schoormans, "The different roles of product appearance in consumer choice." *Journal of Product Innovation Management*, vol. 22, no. 1, pp. 63–81, 2005.

[4] R. L. Underwood, and N. M. Klein, "Packaging as brand communication: Effects of product pictures on consumer responses to the package and brand." *Journal of Marketing Theory and Practice*, vol. 10, no. 4, pp. 58–68, 2002.

[5] M. Lindstrom, "Broad sensory branding." *Journal of Product & Brand Management*, vol. 14, no. 2, pp. 84–87, 2005.

[6] B. J. Phillips, E. F. McQuarrie, and W. G. Griffin, "How visual brand identity shapes consumer response." *Psychology & Marketing*, vol. 31, no. 3, pp. 225–236, 2014.

[7] R. Beise-Zee, "Brand equity retention after rebranding: a resource-based perspective." *Journal of Brand Management*, vol. 29, no. 2, pp. 208–224, 2022.

[8] A. Jakšić-Stojanović, and N. Šerić, "Brand Identity of Montenegro through Verbal and Visual Elements of its Logo." *Journal of Marketing Development and Competitiveness*, vol. 12, no. 4, pp. 134–144, 2018.

[9] V. U. Vinitha, D. S. Kumar, and K. Purani, "Biomorphic visual identity of a brand and its effects: a holistic perspective." *Journal of Brand Management*, vol. 28, no. 3, 272–290, 2021.

[10] N. T. Van Trang, T. L. Nghiem, and T. M. Do, "Improving the competitiveness for enterprises in brand recognition based on machine learning approach." *Global Changes and Sustainable Development in Asian Emerging Market Economies*, vol. 1, pp. 359–373, 2022.

[11] E. Martarello, "*Exploring CNNs and Attention Mechanisms for Brand Identification in Fashion Runway Shows*." Master's thesis, Ca' Foscari University of Venice, 2023.

[12] R. R. Selvaraju, M. Cogswell, A. Das, Vedantam, R., Parikh, D., and Batra, D. "Grad-CAM: Visual explanations from deep networks via gradient-based localization." *International Journal of Computer Vision*, vol. 128, pp. 336–359, 2017.

[13] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

[14] J. An, and I. Joe, "Attention map-guided visual explanations for deep neural networks." *Applied Sciences*, vol. 12, no. 8, 3846, 2022.

[15] Jean-Noël. Kapferer, "Why are we seduced by luxury brands?." *Journal of Brand Management*, vol. 6, no. 1, pp. 44–49, 1998.

[16] GazeRecorder. Online Eye Tracking Software. [Online]. Available from: https://gazerecorder.com/ 2024.05.06

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010, 2017.

[18] I. Borg and P. J. F. Groenen, "*Modern Multidimensional Scaling: Theory and Applications*," Springer Series in Statistics, Springer-Verlag New York, 2005.