INVESTIGATING AND GENERATING THE EVALUATION METHOD BETWEEN HUMAN FASHION AESTHETIC AND GENERATIVE AI

HSI YEH WANG

MASTER OF SCIENCE
IN
DIGITAL TRANSFORMATION TECHNOLOGY

SCHOOL OF INFORMATION TECHNOLOGY
MAE FAH LUANG UNIVERSITY

2023

INVESTIGATING AND GENERATING THE EVALUATION METHOD BETWEEN HUMAN FASHION AESTHETIC AND GENERATIVE AI

HSI YEH WANG

THIS THESIS IS A PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
DIGITAL TRANSFORMATION TECHNOLOGY

SCHOOL OF INFORMATION TECHNOLOGY
MAE FAH LUANG UNIVERSITY
2023

INVESTIGATING AND GENERATING THE EVALUATION METHOD BETWEEN HUMAN FASHION AESTHETIC AND GENERATIVE AI

HSI YEH WANG

THIS THESIS HAS BEEN APPROVED
TO BE A PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE
IN
DIGITAL TRANSFORMATION TECHNOLOGY
2023

EXAMINATION COMMITTEE

.....CHAIRPERSON
(Asst. Prof. Santichai Wicha, Ph. D.)

.....ADVISOR
(Asst. Prof. Worasak Rueangsirarak, Ph. D.)

.....CO-ADVISOR
(Surapong Uttama, Ph. D.)

Teerawat Kamnardsiri.
.....EXTERNAL EXAMINER
(Asst. Prof. Teerawat Kamnardsiri, Ph. D.)

ACKNOWLEDGEMENT

I extend my sincere appreciation to all individuals who played a role, whether direct or indirect, in the accomplishment of this thesis. Grateful acknowledgments are owed to Mae Fah Luang University, School of Information Technology, for granting permission for the development of this thesis and for their invaluable ideas and suggestions. My heartfelt gratitude goes to Dr. Surapong Uttama for administrative guidance and for providing valuable insights throughout the preparation of this thesis.

I would like to express my thanks to Asst. Prof. Dr. Santichai Wicha for his lectures and unwavering support, Asst. Prof. Dr. Worasak Rueangsirarak for guiding me through various topics in Digital Research Methodology, and the esteemed dean of the School of Information Technology, Mae Fah Luang University, Asst. Prof. Dr. Teeravisit, for fostering an inspiring academic environment.

Special appreciation is reserved for my parents for their kindness and assistance. I extend my respect to all my teachers for their valuable recommendations and to my seniors, friends, and the dedicated staff members of Mae Fah Luang University for their collaborative efforts in bringing this thesis to completion.

Hsi Yeh Wang

Thesis Title	Investigating and Generating the Evaluation Method Between Human Fashion Aesthetic and Generative AI
Author	Hsi Yeh Wang
Degree	Master of Science (Digital Transformation Technology)
Advisor	Asst. Prof. Worasak Rueangsirarak, Ph. D.
Co-Advisor	Surapong Uttama, Ph. D.

ABSTRACT

AI drawing tools set off a revolutionary trend in the field of image creation. However, no clear and appropriate evaluation standard to rank AI graphics in fashion exists. In addition, most fashion industry insiders have never used AI tools before. This research aims to evaluate whether AI-generated images could satisfy fashion designer's needs by comparing automatic and human evaluations, and see if there is a need to create a new evaluation method. Therefore, in the first part, AI-generated fashion datasets with 25 images using Leonardo AI were created, and a survey was conducted to check how the experts ranked the AI images. Automatic evaluation methods, such as FID and Clip scores of each picture were measured to observe the correlation with human evaluation. The result showed that the correlation coefficient between expert and FID scores is only 0.30, while the correlation coefficient between expert and CLIP scores is 0.05. In other words, human evaluation and automatic evaluation are not so related and both have insufficiencies. Automatic evaluation is unable to provide judgments on fashion and aesthetics. The evaluations of different experts vary greatly due to the subjective consciousness and cannot provide fair and objective standards. Thus, it is necessary to create a new evaluation method that can evaluate the generated image in both fashion and AI aspects.

In the second part, based on the conclusion of the first part, it is necessary to create a new evaluation with the general public aspect. Therefore, the research addresses these issues by (1) creating a dataset of AI-generated fashion images labeled with customer rankings and (2) establishing a new evaluation method from the perspectives of the general public and the market. Two CNN models were trained: one for regression predicting continuous aesthetic scores and another for classification categorizing images into discrete rating intervals. Performance evaluation revealed that the regression model constrained by clip techniques maintained the original distribution of data, while the classification model provided a benchmark for design indicators. The study concludes that addressing data imbalance and applying augmentation techniques yielded significant results. Specifically, the regression model achieved an RMSE of 0.866 while the classification model attained an accuracy of 93%.

Keywords: Generative AI, Fashion Design, AI Evaluation Method, Convolutional Neural Network, Classification, Regression

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Background and Importance of the Research Problem

As times have progressed, fashion designers have encountered numerous significant challenges. One major challenge is the pressure from fast fashion, which requires designers to release multiple collections within short periods, potentially compromising the quality and creativity of their designs. Statistical data from fashion websites indicates that clothing production doubled in the first fifteen years of the 21st century. Additionally, since 2000, European fashion brands have increased their new collections from just two per year to as many as 24 (Fashion Discounts, 2023). This suggests that designers are struggling to produce enough creative ideas to meet market demands. The advent of innovative production and distribution methods has also compressed fashion cycles, reducing the design period from months to mere weeks. The number of fashion seasons each year has surged from two to potentially 50-100 micro-seasons, as illustrated in Figure 1.1 (Drew & Yehounme, 2017). These developments imply that designers are grappling not only with an increased volume of designs but also with intense time constraints.

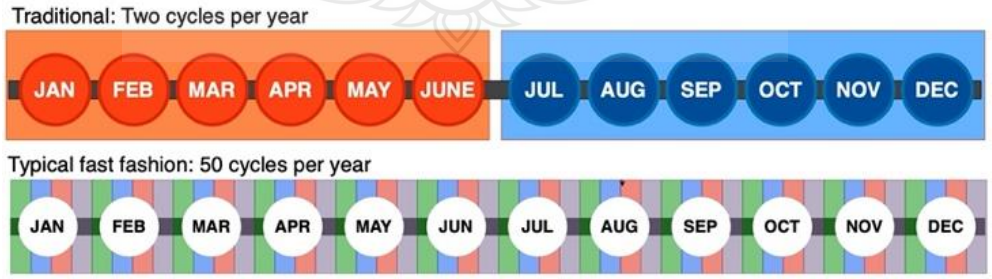


Figure 1.1 The decreasing fashion design cycle

Figure 1.2 The comparison of real images and AI-generated image

Fortunately, the rapid advancement of artificial intelligence has led to significant progress in image-generative AI. Integrating technology with traditional fashion appears to be the clear solution to this challenge. According to Market.us, the market size of generative AI in fashion is expected to grow from 69 million USD in 2022 to 1481 million USD in 2023 (Maket.us, 2024). Currently, extensive research is focused on leveraging deep learning and generative models for image synthesis, providing a robust tool for creating fashion images (Yan et al., 2022). The most notable advancement in generative AI is the development of Generative Adversarial Networks (GANs), which excel in tasks such as image generation and semantic segmentation. Figure 1.2 shows examples of images generated by the GAN model.

There are many examples of using generative AI in fashion design. For the Spring/Summer 2024 collections, fashion houses Heliot Emil and Collina Strada drew inspiration from this area. Both input pictures of earlier iterations into a generative AI tool to generate fresh designs that could be improved upon (BoF, 2023).

However, despite the increasing use of generative AI in the fashion industry, current evaluation methods have not yet been proven sufficient for assessing the performance of AI-generated images. Furthermore, without an appropriate evaluation method, it is challenging to determine if an image-generative AI tool can effectively assist designers in completing fashion designs. Therefore, it is crucial to identify any shortcomings in existing evaluation methods.

Typically, several methods are commonly used to evaluate the performance of generated images. However, these evaluation methods for generative AI often focus on assessing whether the images appear realistic or unique, or if the prompt description matches the image using machine learning. Consequently, it is difficult to determine if the results are valuable from a fashion perspective. Additionally, human evaluation also involves certain uncertainties, making it unclear whether human judgment can provide an objective standard.

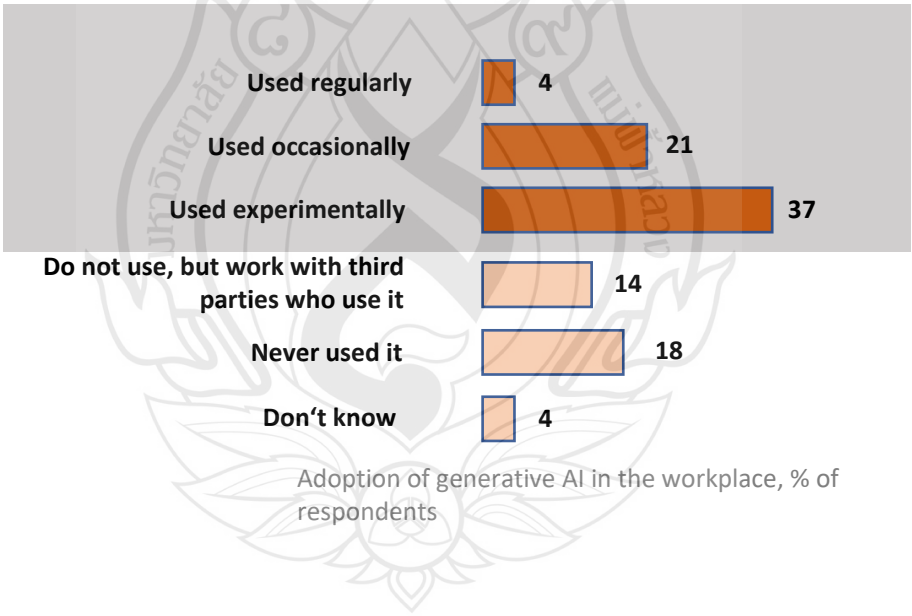


Figure 1.3 Adoption of generative AI in the workplace

On the other hand, generative AI in fashion design is still in the experimental stage in most related enterprises. Figure 1.3 shows that most of the companies still use generative AI experimentally. Due to the instability of AI generation or the lack of algorithms, the output design is unpredictable and may not meet the standards. The

target audience of fashion design is ultimately humans rather than AI, so it is difficult to know if AI-generated fashion designs meet human aesthetic standards and market demand. After the AI fashion design is generated, there is still a lack of an evaluation method to confirm whether the design has market value and human aesthetic from the consumer's point of view (BoF, 2023). Therefore, it is necessary to validate whether the existing evaluation methods (human evaluation and AI evaluation), are suitable for evaluating AI fashion pictures. Moreover, if the existing evaluation methods are unsuitable, we will try to establish a new evaluation method for the performance of AI-generated fashion design.

1.2 Research Objective

1.2.1 To create the dataset in the AI fashion domain

1.2.2 To compare the existing evaluation methods with the human evaluation based on an established dataset.

1.2.3 To establish a new evaluation method for fashion AI images with general public aspect and market value

1.3 The Importance of Research

This study will greatly help the development of fashion AI and fashion product market assessment in the future. Existing AI image evaluation methods and human judgment have not proven effective in evaluating AI fashion, so finding out the correlation between AI and human evaluation is crucial to understand if they can both evaluate with the same evaluation direction and provide a comprehensive and persuasive judgment. If the existing methods cannot fully meet the required factors for evaluating fashion AI, creating new evaluation methods that combine customer aspects and human aesthetics can provide a marketing reference value, and also assist designers and design companies in the early prediction and evaluation of the adaptability of AI-generated products in the market.

1.4 Research Question

This study aims to answer these research questions as follows:

1.4.1 Is there a correlation between AI evaluation and human evaluation? What are the Fashion designers' and industry insiders' views on using AI?

1.4.2 If the existing evaluation cannot evaluate AI fashion images, how to establish a new evaluation method for fashion AI images?

1.5 Scopes of Research

This research investigates the application of generative AI in fashion design by surveying five fashion experts in the industry and using existing AI image evaluation on 25 AI fashion images.

To create a new evaluation method, approximately 1200 images and customer scores collected from fashion websites will be used to train the two CNN models. The classification model will be separated into five classes.

1.6 Expected Result

The expected results aim to contribute to the advancement of fashion AI by addressing evaluation challenges, providing valuable insights into the correlation between AI and human assessments, and offering a novel hybrid evaluation approach that aligns with both AI capabilities and human aesthetic considerations. The expected results will include dataset creation, comparative evaluation (correlation analysis results), industry perspectives, and the development of a new evaluation method. Below are the anticipated outcomes and measurement criteria.

1.6.1 Establishment of Fashion AI-Generated Image Dataset

Anticipated Outcome: Successfully use the GAN-based tool, Leonardo AI to create an AI-generated image dataset in the fashion domain. The dataset for comparing existing evaluation methods was created requiring the trending key work. The dataset

for training the new evaluation method was created based on the real-time design of a fashion website annotated with customer ranking.

Measurement Criteria: Dataset size, diversity, and relevance to fashion aesthetics.

1.6.2 Comparative Evaluation of Existing Methods

Anticipated Outcome: A thorough comparison between currently available evaluation methods (AI and human) to assess the results of the established dataset.

Measurement Criteria: Correlation analysis between AI and human evaluations, identifying strengths and limitations of each method.

1.6.3 Development of a New Evaluation Method

Anticipated Outcome: Establish a new evaluation framework that integrates AI capabilities and human aesthetic considerations to ensure AI-generated fashion designs meet market and aesthetic standards.

Measurement Criteria: Design and implement the evaluation system with trustworthy accuracy, ensuring a balanced consideration of human aesthetics and AI-based assessments.

CHAPTER 2

LITERATURE REVIEW

2.1 Theoretical Reviews

The study is a discussion on cross-curricular learning that combines the traditional fashion industry and artificial intelligence, therefore, Basic theoretical backgrounds in both fields can provide a useful reference. The literature review comprehensively covers the framework of all the fundamental theories of this study and provides relevant arguments and supporting evidence, such as data collection, model use, experimental evaluation methods, etc. The directions of previous related research can be mainly divided into several areas: Application of AI-generated images in fashion and evaluation method of AI images, CNN model, and aesthetic evaluation using classification and regression.

2.1.1 GAN's Rise and Development

The field of artificial intelligence has been consistently redefined by pioneering advancements., particularly the introduction of Generative Adversarial Networks (GANs), which have captivated researchers and implementers alike. Introduced by Ian Goodfellow and his colleagues in 2014, GANs represent a significant advancement in generative modeling. A GAN model consists of two neural network sub-models: the generator, which creates new examples, and the discriminator, which attempts to distinguish between real and fake examples. The adversarial interplay between these components has led to a robust technique for producing highly convincing and contextually coherent data. GANs are particularly remarkable for their capacity to generate images, audio, video, and text that closely mimic human-created content, exceeding traditional expectations for machine-generated output ("Generative Adversarial Network", 2023).

Through continued research, more versatile and powerful GAN models have been developed by optimizing and adjusting parameters. Examples include CycleGAN and StyleGAN, among other specialized GANs designed for various purposes (Das, 2023).

Beyond their technical sophistication, GANs have had a transformative impact on numerous industries and creative fields. In the realm of art, Mario Klingemann is a trailblazer who uses GANs to create artwork. His work often features "style transfer," which merges the style of one piece of art with the content of another, resulting in innovative and unique designs. Klingemann has used GANs to produce not only images but also animations, interactive installations, and generative music. Furthermore, GANs are extensively employed in synthetic media to generate new images, videos, and audio. For example, NVIDIA has used GANs to create realistic images of human characters, animals, and landscapes for use in video games, movies, and other digital media (Panopticon, 2023).

2.1.2 AI Drawing Tools

Amid the global surge in AI technology, AI-driven drawing has become a prominent topic. New AI tools leveraging GANs are frequently introduced, including DALL-E, Midjourney, and Stable Diffusion. Figure 2.1 illustrates the structure of a GAN-based tool. Despite their computational and application capabilities, many of these tools lack user-friendly control interfaces, making them challenging for beginners. Leonardo.AI stands out among AI drawing tools for its user-friendliness, being based on the open-source Stable Diffusion. It offers a clear web user interface (UI) that allows users to bypass complex coding and utilize functions directly from the webpage. The most crucial aspect of AI drawing is crafting a precise prompt for text-to-image functionality. Leonardo.AI also allows for the use of pre-existing models to generate images or perform calculations by uploading reference images (Image-to-Image). A prompt generator is available to assist those unfamiliar with prompt writing. These features make it easier to produce images that meet user expectations (Leonardo.ai., 2023; Azza, 2023). Given these advantages, Leonardo.AI will be employed as the tool for the experiment.

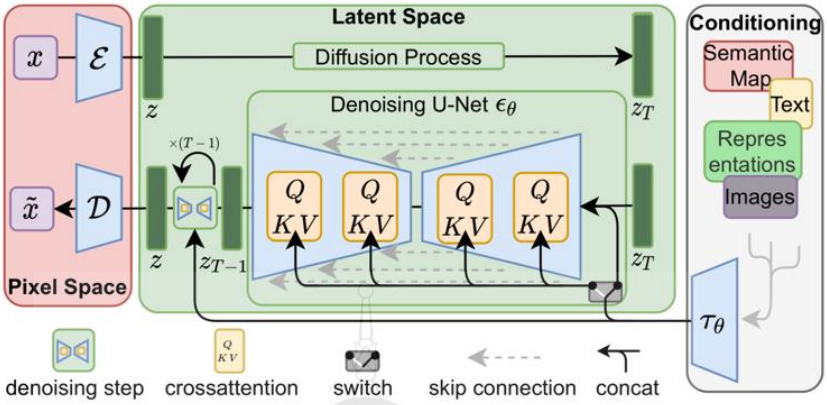


Figure 2.1 Overview of stable diffusion

2.1.3 Fashion Design Process

Generally speaking, it must go through several stages before the adoption of a fashion design is finalized, requiring an understanding of the design process model. As illustrated in Figure 2.2, this process covers every step from conceptualization to the completion of production. We have organized a flowchart with a more generalized design process to explain how the entire industry operates (Smith, 2022).

Figure 2.2 Workflow of the design process

1. The first step is for designers to obtain inspiration through a series of trends and market research, and compile all the information into emotion boards to facilitate organized ideas and inspirations.

2. Then, according to the emotional board completed in step 1, the designer will start to outline your clothing creativity and narrow the scope of the final design by using sketches.

3. After the designer has a rough idea of the design, the next step will be creating a technical sketch that reflects the correct construction of the garment by using CAD (computer-aided design).

4. Once the technical sketch is complete, it will be used to create a Tech pack for the garment. the silhouette and design in the technical sketch will also be determined as construction details and specs and sent to factories and raw material suppliers.

5. After the factories and raw material suppliers know your needs, the designer can source the different elements of the design such as manufacturing

(where the product is assembled), trims (supplier), fabrics (supplier), and even labels and packaging.

6. After confirming the materials and manufacturing factory, the sample stage can begin. The factory will conduct sampling based on your technical package specifications, fabric, and decoration, and the purpose of sampling is to inspect whether there are any problems or parts that need to be adjusted in the design after actual production.

7. Generally speaking, there will be many rounds of sampling, the designer needs to review the sample and adjust the tech pack until it reaches the standard of mass- production.

8. To reach the final step: mass production, you will need to ask the supervisor to examine all the tech packs to make sure every information and detail is correct, and the design is being approved by the top development team.

9. After completing all the above steps, the factory and suppliers can start to proceed with the mass production of goods.

The AI generative tool aims to take is assume to improve and simplify the first three steps, including the inspiration, fashion illustration and sketches. But in the research, what we want to discuss is the AI fashion evaluation method, which more focus on optimizing the review and design approval.

2.1.4 Evaluate Method for AI-generated Image

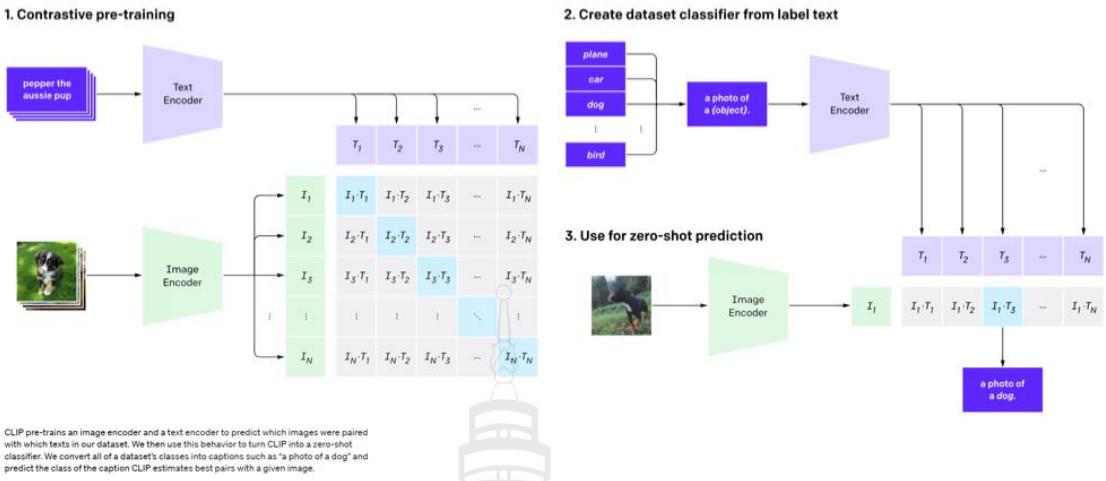
The two main criteria for checking the performance of AI-generated images encompass image quality (reflecting photorealism or fidelity), and text-image alignment (how accurately the generated images correspond to text semantics). Inception Score (IS), Frechet Inception Distance (FID), and CLIP Score serve as key metrics in evaluating these aspects.

Inception Score (IS) serves as an automated metric for assessing the performance of AI-generated models. This score is devised considering two crucial factors: image quality and diversity. A higher Inception Score indicates that the generated images exhibit both high quality and diversity in their features (Heusel et al., 2017).

Frechet Inception Distance (FID) is another widely used metric to evaluate image quality. FID calculates the distance between feature vectors between real and AI-generated images. It provides a quantitative measure of how similar the statistics of computer vision features are between real and AI-generated images. Lower FID scores indicate greater similarity, with a perfect score of 0.0 signifying identical statistics between the two groups of images. Below is the formula for the FID score (Shena et al., 2022).

$$FID = \|\mu_r - \mu_g\|^2 + T_r(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}) \quad (1)$$

Addressing text-image alignment, the CLIP Score stands out as a prevalent text-to-image similarity metric. The CLIP Score evaluates the correlation between a generated caption for an image and the textual content that describes the image. A higher CLIP Score indicates a more accurate alignment between the generated captions and the actual content of the image. To derive a CLIP score, the similarity between an image and a corresponding text description is assessed within a shared embedding space. Cosine similarity, a metric used to quantify the cosine of the angle between two vectors in a multidimensional space, is employed to gauge this similarity. The cosine similarity scale ranges from -1 to 1: a score of +1 signifies identical vectors, 0 indicates orthogonality, and -1 implies opposition between the vectors (Uni Matrix Zero, 2023).



Source Radford et al. (2021)

Figure 2.3 The structure of clip score

Some studies suggest that existing evaluation metrics may not adequately capture human perception, particularly in assessing the performance of cutting-edge generation models like FID and Clip Score. Therefore, the study will not only examine the automated evaluation metrics but also incorporate manual evaluations to enhance the reliability of conclusions (Zhang et al., 2023; Otani et al., 2023; CLIP score, n.d.).

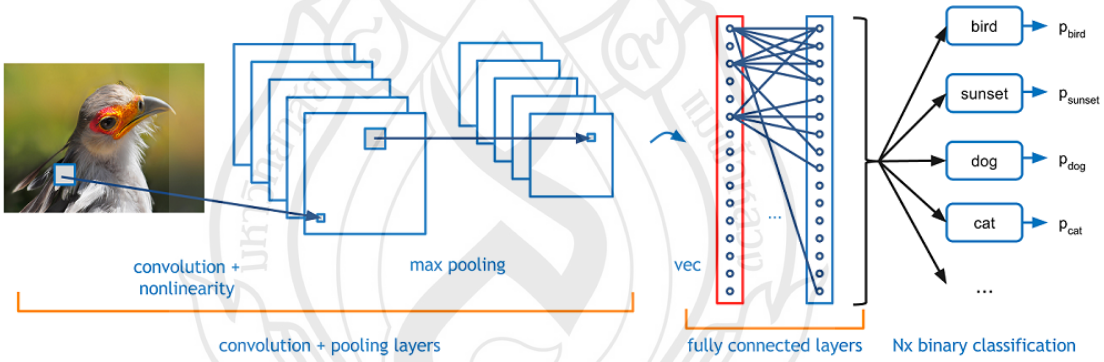
2.1.5 Supervised Learning for Image Classification

Supervised learning-based classification is a prominent machine learning approach employed for predicting future outcomes. This methodology involves learning from various features or variables by establishing a function that maps independent variables to dependent variables. The labeled data, or the model's output, can be categorical, addressing classification problems, or numeric, tackling regression problems. In the context of this study, which centers around fashion styles a categorical data domain, the focus is on classification models. Specifically, two classification models have been implemented, including ensemble models: a choice between a random forest model or a gradient boosting model, and a convolutional neural network (CNN) model. The utilization of these models aims to effectively capture and

comprehend intricate patterns within the categorical fashion style data, thereby enhancing predictive accuracy (Woottisart et al., 2022).

2.1.6 Convolutional Neural Network (CNN)

A Convolutional Neural Network (CNN) is a type of artificial neural network specifically designed for processing and analyzing structured grid data, such as images. CNNs have proven to be highly effective in tasks related to computer vision, including image classification, object detection, and image recognition. The key feature of CNNs is the use of convolutional layers, which are responsible for capturing local patterns and spatial hierarchies in the input data. CNNs excel in capturing hierarchical features and are well-suited for tasks involving grid-structured data like images, making them a fundamental technology in computer vision applications. Figure 2.4 shows the structure of CNN (Dshahid380, 2019).



Source Dshahid380 (2019)

Figure 2.4 Example of CNN architecture

2.2 Related Studies

2.2.1 Human Evaluation and AI Evaluation

Several studies are exploring the correlation between the automated evaluation of AI-generated images and human evaluation, Otani et al. (2023) mention that the effectiveness of validating text-to-image generation models depends on key elements of human evaluation, and considering the complexity of this cognitive process, a profound understanding of text and visual content is required. However, the recent 37 papers have revealed a common trend that many works either rely entirely on automated indicators such as FID or conduct poorly defined human evaluations, lacking reliability and repeatability. In light of the considerations, researchers have proposed a comprehensive and precisely outlined human assessment plan aimed at enhancing the verifiability and reproducibility of human assessments in future research. The experimental results emphasize that the current automatic evaluation methods have shortcomings in maintaining consistency with human perceptual judgments when evaluating the effectiveness of text-to-image generation results.

Moreover, Marin et al. (2020) also provide an overview of the most commonly used qualitative and quantitative evaluation measures for assessing the quality of generated images and the learned representations of adversarial networks. Through empirical comparisons in the context of human face image synthesis, the study demonstrates that the evaluation scores of two widely accepted quantitative metrics, Inception Score (IS) and Frechet Inception Distance (FID), are not correlated. IS score is deemed inappropriate as an evaluation metric for specific problems, while FID exhibits good performance that aligns well with visual inspections of generated samples. Qualitative evaluation serves as a complement to quantitative assessment, offering deeper insights into the learned data representations and facilitating the detection of potential overfitting phenomena. Therefore, this study chooses FID as the evaluation standard of AI picture quality.

2.2.2 CNN Classification in Fashion

Convolutional Neural Networks (CNNs) are widely utilized in image processing tasks, including recognition, detection, and classification, due to their exceptional performance. They excel in learning hierarchical features like edges, textures, and shapes, making them adept at recognizing objects within images. Consequently, CNNs are frequently employed in artificial intelligence applications, particularly in image classification tasks. In image classification, CNNs analyze input images and assign them to specific labels or categories based on learned patterns from labeled data. Their ability to automatically extract relevant spatial features from images makes them highly effective for this purpose (Sharma, 2024; "Example of a CNN for image classification", n.d.).

Given the emphasis of this study on fashion design, the following are additional studies on the use of CNN classification models in related fields. Reference (Gh, 2019) discusses the application of deep learning, particularly CNNs, in the fashion industry for apparel image classification to solve the difficulties due to various apparel categories and the lack of labeled image data for each category. Their methodology involves pre-training the GoogLeNet architecture on the ImageNet dataset and fine-tuning it on a fine-grained fashion dataset (Seo & Shin, 2018).

Xuan et al. (2021) involves designing eight CNN models based on transfer learning and convolutional neural networks, training them on the Street-FashionData dataset, and grouping them into three categories with specific design variations to improve clothing image classification accuracy. Transfer learning is used to apply knowledge gained from one problem to another related problem.

2.2.3 Classification in Aesthetic Evaluation

However, the above research only focuses on image classification of clothing types by category and mainly focuses on object detection, which cannot meet the research objectives. Therefore, it is crucial to use the CNN classification model and image corresponding labels to enable the model to learn human aesthetic scores for images. Many studies have been related to this, the vast majority of cases focus on the aesthetic of artistic works or photography. Although the application fields are different, their research methods still have certain reference values (Areeb et al., 2021).

In further exploration, we get the example from Zhang et al. (2023) who put forward the synergetic assessment of image quality and aesthetics to better understand human subjective preferences for digital images. The research proposed a two-stream learning network to simultaneously evaluate both the quality and aesthetic aspects of images. This network adopts a top-down perception mechanism, learning from finely-grained details and holistic image layout at the same time. Several researchers apply the CNN model for image aesthetics prediction. In other words, by training a Convolutional Neural Network (CNN) model, obtain human-evaluated results without any human intervention.

Areeb et al. (2021) discuss the use of AI techniques for assessing visual aesthetics in digital art posters. They highlight the challenges of categorizing images based on aesthetic appeal and the limitations of AI models compared to human artists. Their methodology involved training a Convolutional Neural Network (CNN) on a self-assembled dataset of digital art posters to categorize them as having low or high aesthetic value, achieving 89% accuracy. This demonstrates that CNNs can effectively classify posters in binary classification scenarios based on learned labels and extracted features. However, it remains unproven whether CNNs can maintain accuracy with multi-class classification or continuous labels, presenting a research gap to explore.

2.2.4 Regression in Aesthetic Evaluation

Another research focuses on using Convolutional Neural Networks (CNNs) for image aesthetics prediction (Kao et al., 2015) interpret aesthetic quality assessment as a regression problem, and they apply the convolutional network to learn the features. Subsequently, train a regression model using these aesthetic features. Due to the extremely unbalanced distribution of the aesthetic scores on this dataset, the prediction capability of existing methods is limited. Jin et al. (2016) overcome the limitation by introducing the use of weighted CNNs for image aesthetics prediction, present regression, and histogram prediction models to improve accuracy and estimate assessment difficulty, and demonstrate an image enhancement application. The research claims that they approach aesthetic quality assessment as a regression problem for two main reasons. First, regression models closely resemble how the human visual system processes aesthetic quality. A classification model can only predict aesthetic

class (high or low), whereas a regression model can quantify the degree of aesthetic quality, similar to the human visual system. Second, the distribution of the features learned by convolutional networks may enhance the solvability of the regression problem.

Previous studies have predominantly applied classification models under the assumption that they can only handle binary classes. This research, however, aims to explore the use of multiple classes for aesthetic assessment. Additionally, prior research has not focused on evaluating AI-generated fashion design images. Therefore, this study attempts to train a model using AI-generated images to develop an assessment method, thereby providing new insights into this field.

CHAPTER 3

METHODOLOGY

3.1 Overall Methodology

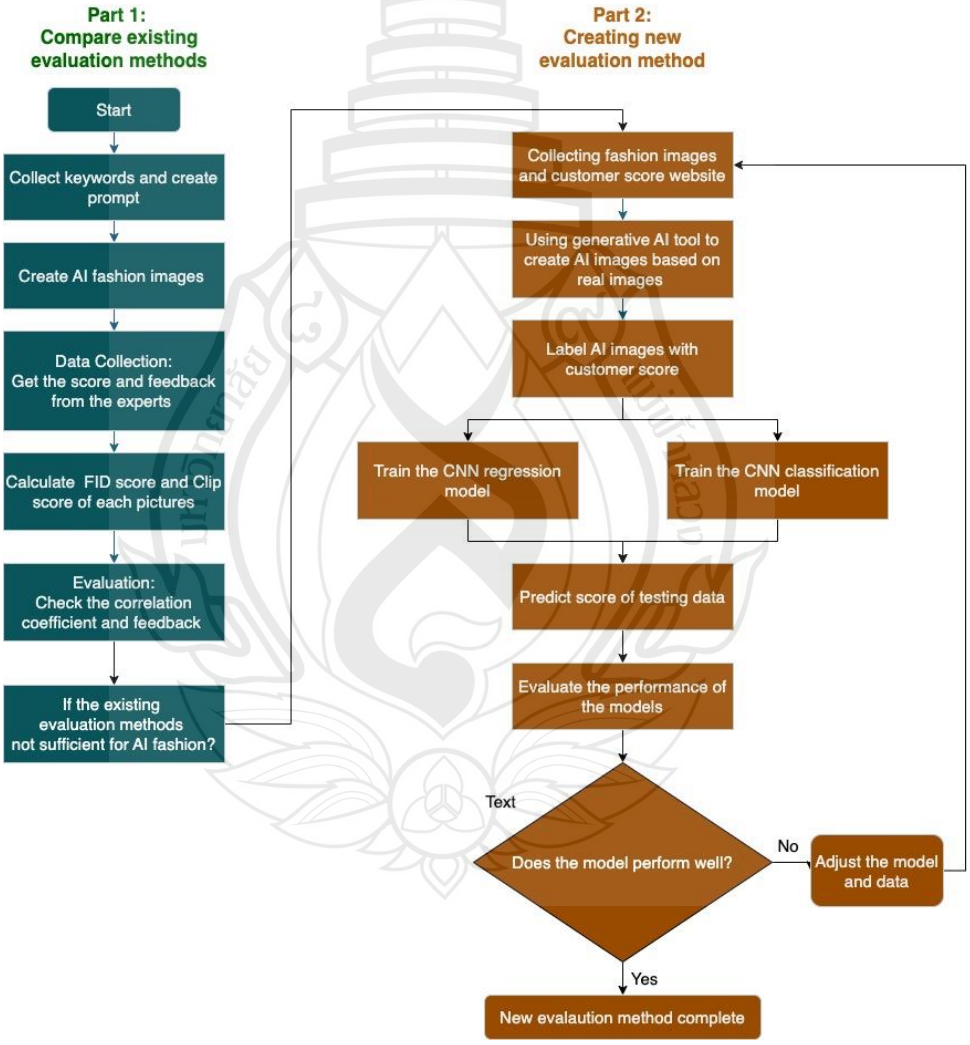


Figure 3.1 Methodological framework

This study is divided into two parts according to the objective. The first part is a comparison between human experts' evaluation and automatic evaluation. The scores from the experts represent the human evaluation while the FID score and Clip score represent the existing AI evaluation. This part aims to verify if both methods possess human and automated evaluation perspectives by confirming the correlation between the two to understand whether these two evaluation methods, one focused on fashion and the other on AI images, can effectively assess AI-generated fashion designs. However, if the validation results show a low correlation between the two, a second part is needed, which is to provide a more effective evaluation method specifically for AI-generated fashion. In the second part, we will apply CNN classification and regression model to create a new evaluation method.

3.2 Methodology for Comparing the Existing Evaluation Method

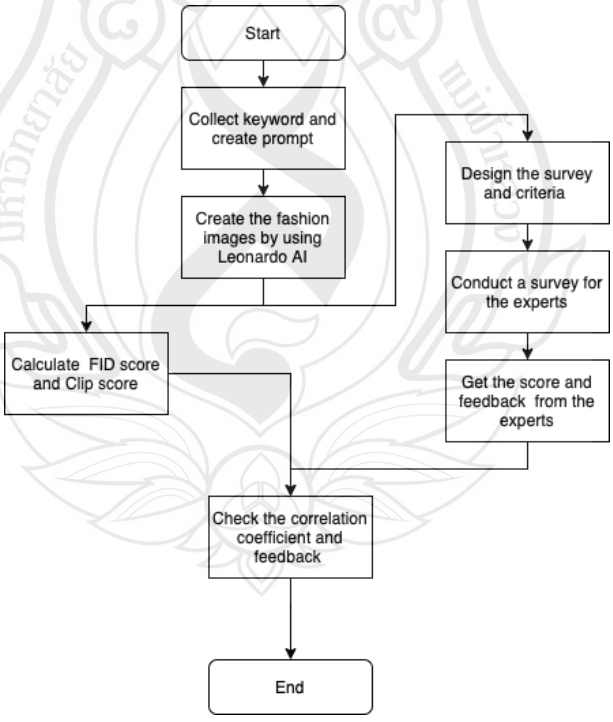


Figure 3.2 Methodology of comparing the existing evaluation method

3.2.1 Image Generation

The AI image-generated tool used in this research is Leonardo. AI. This platform offers an exceptionally intuitive interface, allowing users to input prompts (text descriptions) for desired outputs, upload reference images, and adjust the balance between textual and visual inputs. As a result, it can generate diverse and adaptable image outputs.

The prompts used to generate images are derived from four key factors within the fashion design process: contemporary fashion trends, emerging future fashion trends, material and detailing, and garment patterns. Keywords for future fashion trends are sourced from fashion trend forecasts for 2024 across multiple fashion websites.

Establishing clear regulations for prompt creation ensures that the generated outputs are comprehensive and prevents the result from being affected by intuition or subjective biases.

Table 3.1 shows how the keywords become prompts to generate images, and examples of generated images and the corresponding prompts by using Leonardo AI are shown in Figure 3.3

		
A beige pleated suit trousers with a small plaid pattern, crafted from luxurious silk material, cropped to a slim fit with a full leg photo.	**a skirt with a graffiti style rose pattern and silk texture**	**A zebra patterned printed mini dress with a delicate lace trim and a ruffled hemline. silk texture.**

Figure 3.3 Generated image and prompt

Table 3.1 Example of prompt

Example of generating prompts		
	Keywords	**Using prompt generator**
fashion forecast	ex. Y2K style, pink, vintage	ex.
Current fashion	ex., wide trouser, glittering, outdoor	1. A navy t-shirt with a pocket on the right side, made of heavy-weight cotton fabric, with a relaxed fit.
pattern	ex. loose fit, cropped, long sleeve	2. A light pink crop-top long-sleeve shirt, rendered in a Y2K style and made of linen material.
material	ex. linen, silk, functional material, water repellent	
vibe	ex. Sexy, Sporty, Cute	

3.2.2 Survey Design

This study obtained the attributes of each product image by referring to the relevant fashion categories on Style.com. However, due to the website's overly detailed classification, eight broader fashion-related categories were adopted to ensure the diversity and independence of population data (Farfetch, n.d.). In addition, three images were selected for each category to ensure an even distribution of the survey, provide sufficient data for the evaluation of each category, and get more balanced results.

This survey targets experts who work in the fashion industry at least five years, which including designer and product planners, aiming to analyze the evaluations of AI-generated images by industry professionals. Prompts and the corresponding design images will be shown in the survey. Respondents will rank various image attributes based on seven criteria derived from keywords extracted from searches on “criteria for judging the quality of fashion design,” as outlined below, and the score will range from 1 to 5. Before the participants start the questionnaire, the complete process of using AI tools to generate images and the detailed definition of each criterion will be introduced at the beginning of the questionnaire to clearly define the evaluation criteria and avoid confusion or bias among the participants.

1. Accuracy: Do the generated- images match the description of the prompt?
2. Creativity: How innovative and unique is the design? Does it include novel elements, unusual combinations, and eye-catching visual effects?
3. Popularity: Does the design align with current or future fashion trends, including popular colors, shapes, materials, and styles?
4. Detail and Pattern: How well are the details and patterns represented in the generated image, such as stitching, accessories, and patterns?
5. Material Choice: How would you evaluate the quality of material generation and rendering effect?
6. Practicality: Can the generated pictures assist in subsequent designs?
7. Production Feasibility: How feasible is it that the design in the picture could be produced in reality?

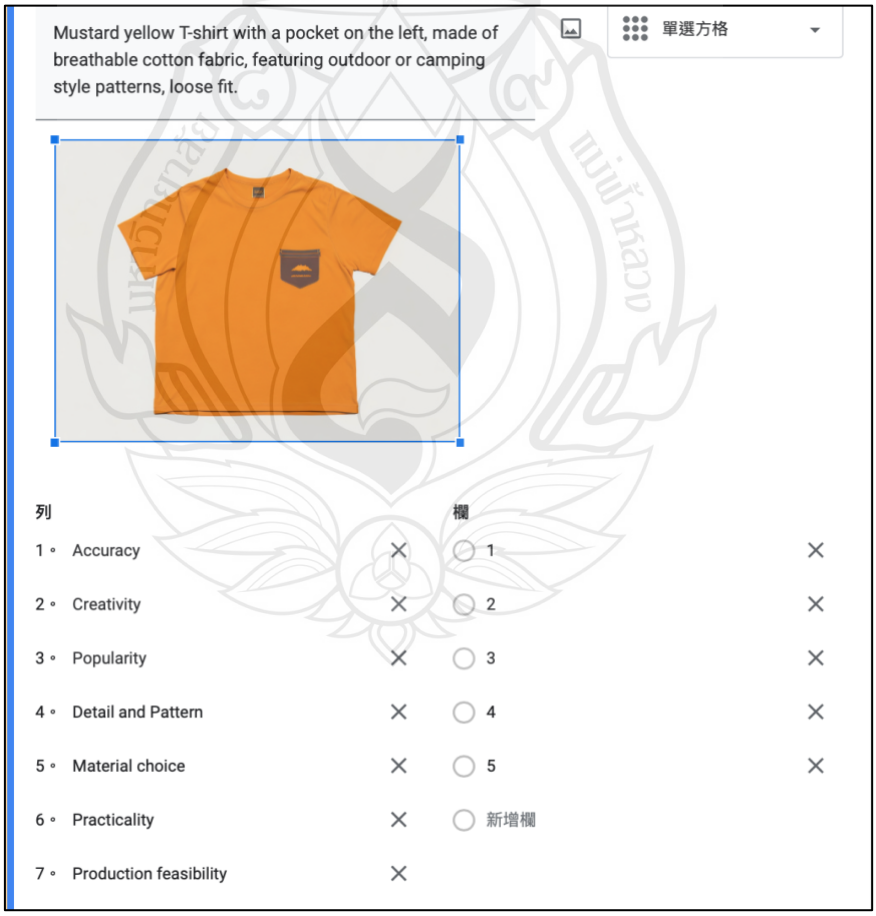


Figure 3.4 Example of the survey

The inclusion of accuracy and practicality in the standards is mainly to verify the efficiency of applying AI tools in real work environments, ensuring the effectiveness of the evaluation method for assessing AI performance.

Despite scoring images, the survey includes two open-ended questions at the bottom to gather the perspectives of fashion professionals on the prospects and views of using AI in their work.

3.2.3 AI-generated Evaluation

Clip and FID scores will be computed as AI ranking using PyTorch to compare with human rankings of AI-generated images. Clip scores will be derived by preparing generated images and corresponding prompt text to obtain output scores. On the other hand, the DeepFashion dataset (Liu et al., 2016), a real-world fashion dataset, will be used to calculate the distance of feature vectors with AI-generated images to obtain the FID scores. Both AI-generated and real images will undergo preprocessing to align with the Inception V3 model, crucial for FID feature categorization.

3.2.4 Result Analysis

After the data of the FID score, Clip score, and expert score are collected, We will examine the correlation coefficient between each evaluation method to analyze whether they have mutual influence and whether there are correlation behind the numbers., which is a statistical measure of the strength of a linear relationship between two variables (Fernando, n.d.). Additionally, responses to open-ended questions are summarized and concluded as current designers' perceptions and perspectives toward using AI.

3.3 Methodology of for Creating the New Evaluation Method

Figure 3.5 Methodology of creating a new evaluation method

3.3.1 Data Collection

Three types of data are needed for the study: product images, product descriptions, and scores. The images and product descriptions serve as reference images and prompts used to create similar fashion AI images. The scores represent the labels of the pictures

Three types of data were essential for this study: product images, product descriptions, and scores. These components served as references and prompts for generating similar fashion AI images, with scores representing picture labels. Despite extensive searches, existing datasets did not combine fashion images with customer ratings, and can’t meet the research requirements. Consequently, we resorted to web scraping fashion e-commerce websites for data collection. Shein was selected due to

several reasons: (1) Most websites lacked customer ranking scores. (2) Even when available, insufficient customer comments and rankings compromised score credibility. (3) Some websites prohibited web scraping. Given these constraints, Shein emerged as the optimal choice, offering a wide product range and necessary ranking scores.

Web Scraper, a Google Chrome extension tool was used to collect the data. Approximately 1400 product data points were collected. Figure 3.6 illustrates an example of data extracted from Shein, including product images, product descriptions, and scores. Figure 3.7 shows the fashion images that were downloaded from the collected data.

web-scraper-order	web-scraper-start-url	product_link	product_link-href	name	picture-src	score
1709106741-3	https://us.shein.com/	SHEIN LUNE Women V	https://us.shein.com/SHE			5.00
1709106753-6	https://us.shein.com/	DAZY Solid Color Shor	https://us.shein.com/DAZ			4.94
1709106761-9	https://us.shein.com/	SHEIN LUNE Family M	https://us.shein.com/SHE			5.00
1709106773-12	https://us.shein.com/	SHEIN Slayr Valentine's	https://us.shein.com/SHE			4.35
1709106785-15	https://us.shein.com/	SHEIN EZwear Women	https://us.shein.com/SHE			5.00
1709106814-18	https://us.shein.com/	SHEIN LUNE 1pc Wor	https://us.shein.com/SHE			5.00
1709106823-21	https://us.shein.com/	SHEIN LUNE Valentine	https://us.shein.com/SHE			5.00
1709106833-24	https://us.shein.com/	DAZY Solid Zip Up Wi	https://us.shein.com/DAZ			4.93
1709106844-27	https://us.shein.com/	DAZY High Collar Soli	https://us.shein.com/DAZ			4.95

Figure 3.6 Data collected from SHIEN

Figure 3.7 Fashion images collected from SHIEN

However, upon initial model training, a significant data imbalance was discovered, leading to inaccurate results. This imbalance primarily stemmed from the fact that many fashion websites predominantly showcase products with higher ratings, while concealing those with lower ratings. Consequently, within the score range of 0-5, the collected data was heavily skewed towards ratings of 4 to 5. To address this challenge, additional customer ratings were necessary from the range of 0 to 3, which could not be obtained via web scraping. Consequently, manual data collection was undertaken to extract lower ratings from customer reviews, supplementing the dataset with ratings ranging from 0 to 3 instead of relying only on average ratings. Ultimately, a total of 1286 data points were collected for the research.

3.3.2 AI Images Generation

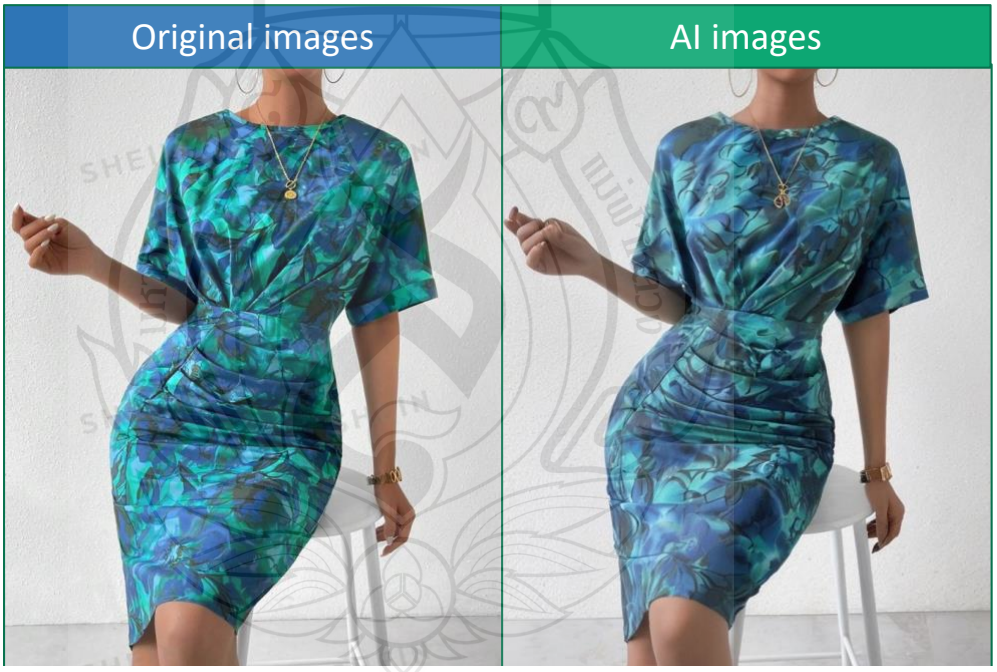


Figure 3.8 Real image and similar AI image

Due to the focus of this study on evaluating AI images, the product names obtained from the data will serve as prompts, while the product images will serve as reference images to generate AI images with high similarity.

After trying several different AI generation tools such as DALL E and Midjourney, Leonardo AI was chosen to generate fashion images because of its fine-tuning parameters and the similarity of the generated results. Figure 3.8 is an example of a comparison of original images and similar images generated by Leonardo AI.

3.3.3 CNN Classification Model

As in the related work mentioned, they approach aesthetic quality assessment as a regression problem for two main reasons. First, a regression model better mimics how the human visual system evaluates aesthetic quality. Unlike a classification model, which can only predict aesthetic class (high or low). Second, the features learned by the convolutional network can make the regression task more tractable.

Nevertheless, this study aims to establish an evaluation method to examine whether fashion designs generated by AI should be adopted or if they have market potential. Moreover, the classification model does not merely possess binary classification capabilities; it can classify multiple categories, thus providing a benchmark for indicators. For these reasons, the classification model still holds research necessity and reference value. Therefore, we have resolved to employ both CNN regression and classification methods for evaluation.

Figure 3.9 The label of two different models

Since customer ratings are continuous values, and the nature of classification models is to extract features based on each label, directly using these continuous values as labels would result in too many classes, making it difficult for the model to effectively learn features and consequently leading to poor training performance. To solve the problem, a common approach is to divide continuous values into several discrete intervals, so we will divide ratings from 1 to 5, with each interval corresponding to a discrete label. Figure 3.9 is an example of how images are labeled in regression and classification models.

```
Layer (type)                 Output Shape              Param #
=================================================================
sequential_3 (Sequential)    (None, 224, 224, 3)       0

rescaling_6 (Rescaling)      (None, 224, 224, 3)       0

rescaling_7 (Rescaling)      (None, 224, 224, 3)       0

conv2d_6 (Conv2D)            (None, 224, 224, 16)      448

max_pooling2d_6 (MaxPoolin   (None, 112, 112, 16)      0
g2D)

conv2d_7 (Conv2D)            (None, 112, 112, 32)      4640

max_pooling2d_7 (MaxPoolin   (None, 56, 56, 32)        0
g2D)

conv2d_8 (Conv2D)            (None, 56, 56, 64)        18496

max_pooling2d_8 (MaxPoolin   (None, 28, 28, 64)        0
g2D)

dropout (Dropout)            (None, 28, 28, 64)        0

flatten_2 (Flatten)          (None, 50176)             0

dense_4 (Dense)              (None, 128)               6422656

outputs (Dense)              (None, 6)                 774

=================================================================
Total params: 6447014 (24.59 MB)
Trainable params: 6447014 (24.59 MB)
Non-trainable params: 0 (0.00 Byte)
```

Figure 3.10 CNN classification model summary

This paper presents a convolutional neural network (CNN) classification model for image classification. Figure 3.10. shows the structure of the classification model. The model normalizes input images and extracts features using three layers of convolution and pooling. These features are flattened into a one-dimensional vector and passed through fully connected layers for higher-level representation. Finally, the

output layer produces the predicted class probabilities. The model is optimized by minimizing the cross-entropy loss to improve prediction accuracy.

$$\theta * = \alpha \text{rg}\min_{\theta} \left(-\frac{1}{\text{N}} \sum_{i=1}^{\text{N}} \sum_{c=1}^{\text{C}} y_{i,c} \log(\hat{y}_{i,c}) \right) \quad (2)$$

The optimization goal is formulated as above, where N is the number of samples, C is the number of classes, and $y_{i,c}$ and $\hat{y}_{i,c}$ are the true labels and predicted probabilities for the i^{th} sample, respectively. Through this optimization process, we can train a high-performance classification model.

3.3.4 CNN Regression Model

Layer (type)	Output Shape	Param #
sequential (Sequential)	(None, 224, 224, 3)	0
conv2d (Conv2D)	(None, 222, 222, 32)	896
max_pooling2d (MaxPooling2D)	(None, 111, 111, 32)	0
conv2d_1 (Conv2D)	(None, 109, 109, 64)	18496
max_pooling2d_1 (MaxPooling2D)	(None, 54, 54, 64)	0
conv2d_2 (Conv2D)	(None, 52, 52, 128)	73856
max_pooling2d_2 (MaxPooling2D)	(None, 26, 26, 128)	0
flatten (Flatten)	(None, 86528)	0
dense (Dense)	(None, 128)	11075712
dropout (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 1)	129

Total params: 11169089 (42.61 MB)
Trainable params: 11169089 (42.61 MB)
Non-trainable params: 0 (0.00 Byte)

Figure 3.11 CNN regression model summary

The CNN regression model is trained on original data in which the labels are continuous numbers, utilizing mean user ratings per image as targets. Its architecture comprises convolutional, pooling, and fully connected layers, culminating in a sum-of-squares layer. With single-score labels, the final fully connected layer outputs predicted aesthetic scores. The objective of the sum-of-squares layer is to minimize the squared L2 norm between predicted and actual scores. Figure 3.11 shows the structure of the regression model.

The objective of this layer can be defined as:

$$min\theta \frac{1}{n}\sum_{i=1}^{n}\left(y_i - h(X_i;\theta)\right)^2 \quad (3)$$

where $h(X;\theta)$ represents the entire mapping process from the input image X to the predicted value y, and θ includes all the model parameters. This formula describes the working principle and goal of the entire convolutional neural network regression analysis model.

In summary, the CNN regression model efficiently predicts aesthetic scores by minimizing prediction errors between actual and predicted values using a sum-of-squares layer.

3.3.5 Performance Evaluation

Upon completion of the training and testing phases for both methods, the comparison between the predicted and real scores serves as the primary indicator of model accuracy and, consequently, is the most crucial metric for evaluating model performance.

For the classification model, we will use the evaluation matrixes below, these metrics help to evaluate different aspects of the classification model's performance.

1. Average Confidence Score: This is the mean of the confidence scores the model assigns to its predictions. It reflects how certain the model is about its predictions. Higher average confidence scores indicate that the model is generally more confident in its classifications.

2. Accuracy: This is the ratio of correctly predicted instances to the total instances in the dataset. It gives an overall measure of how often the classifier is correct.

3. Precision: This is the ratio of correctly predicted positive observations to the total predicted positives. Precision indicates the quality of the positive predictions made by the model.

4. Recall: This is the ratio of correctly predicted positive observations to all observations in the actual class. Recall indicates the ability of the model to find all the relevant cases within a dataset.

5. F1 Score: This is the harmonic mean of precision and recall. It is a balance between precision and recall and is useful when the class distribution is imbalanced.

For the regression model, we will use RMSE and MAE as evaluation matrixes. Root Mean Squared Error (RMSE): RMSE measures the average magnitude of the errors between predicted values and actual values. It is the square root of the average of the squared differences between prediction and actual observation. MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It is the average of the absolute differences between predicted values and actual values. MAE is more interpretable than RMSE and does not penalize large errors as much as RMSE.

CHAPTER 4

EXPERIMENTAL RESULT

4.1 Experiment Result of Comparing Existing Evaluation Method

4.1.1 Experts Scoring Analysis

Eventually, after the survey was distributed, five different industry professionals working in the fashion industry completed the filling out. After organizing and compiling the data as shown in Table 4.1, some interesting results were discovered through analysis.

Table 4.1 Example of collected data

	Experts	CLIP score	FID
	3.17	36.60	1370.06

The scores provided by the experts were first discussed. Figure 4.1 reveals that the correlation coefficient between the scores of each expert is not as high as expected. This result indicates that although all experts have sufficient professional knowledge and market experience, the scoring criteria were also fully explained in the questionnaire, their scoring results are still significantly different. In other words, this indirectly shows the conclusion that people's subjective opinions and personal preferences affect the judgments of fashion aesthetics.

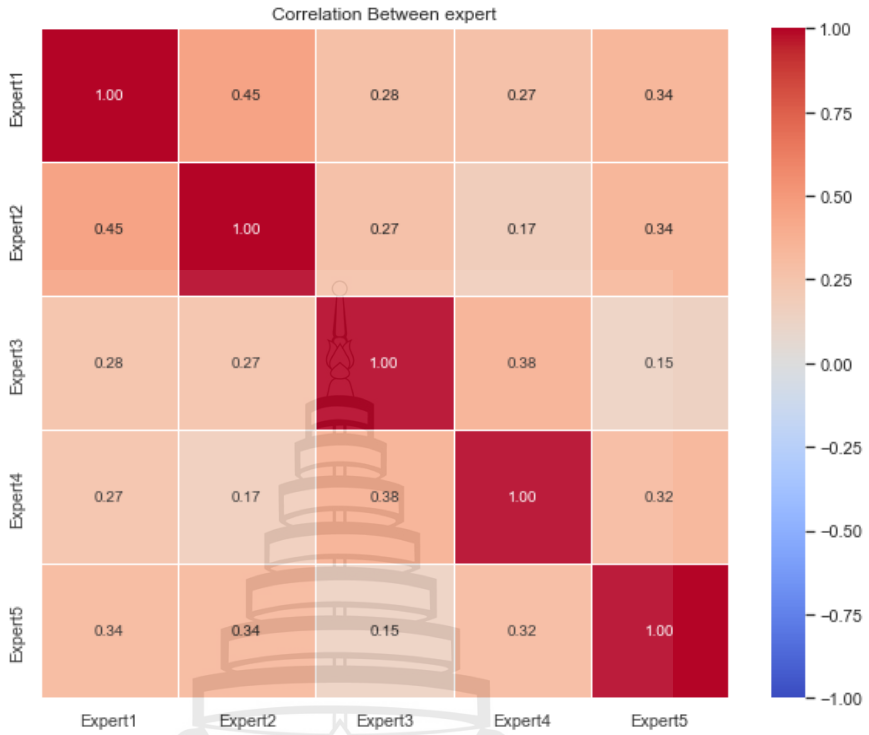


Figure 4.1 Heatmap of the correlation coefficient between experts

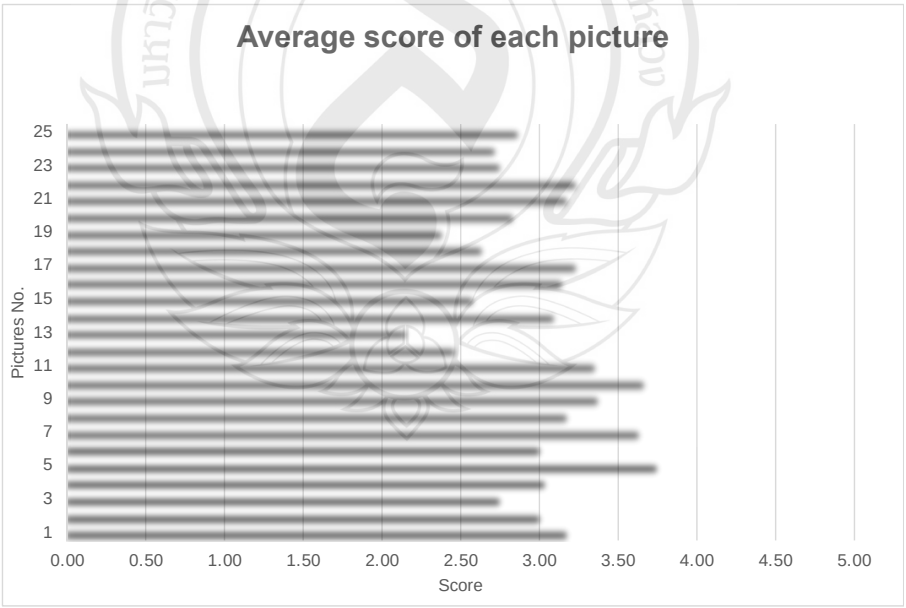


Figure 4.2 Experts' average score for each picture

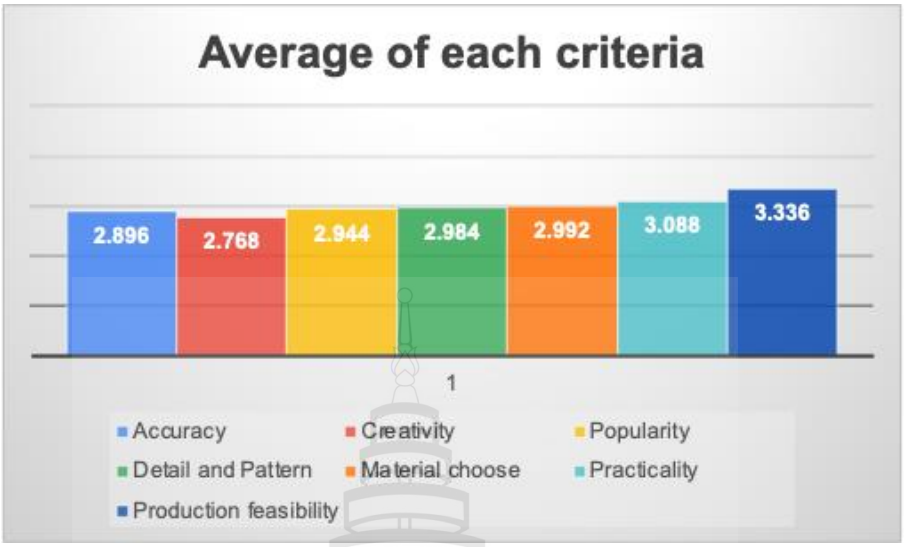


Figure 4.3 Each criterions' average score

Moreover, we also analyzed the ratings of fashion experts on each AI-generated image and the average score of each criterion. As Figure 4.2 shows, the average scores for the 25 images range between 2 and 4, indicating from the perspective of experts, that the performance of these images generated by AI is still acceptable but not too good. The average score of each criterion is shown in Figure 4.3, "Potential for production feasibility" received the best rank score of 3.3, on the other hand, "Creativity" received the lowest average score of 2.8. This data contradicts the assumption that AI tools' creativity would provide more inspiration for the designers.

In the open-ended questions, we sought the industry's views on using AI in fashion. For the question, "Advantages and innovations that may be brought by the application of AI image generation in the field of fashion and creativity," most of the respondents provided positive feedback. They noted the benefits of delivering reference value for early designs, aiding in the design outline for the purchasing team, improving the efficiency of sketching designs, and accelerating the process of proposing new designs. Additionally, beginners who are without a drawing background can also create design sketches, lowering the entry barrier and reducing the costs of the whole design process. Furthermore, compared to design sketches that start from traditional floor plans, AI imaging can now present clothing performance more realistically.

However, for the question, “Have you ever participated in or considered collaborating with AI image generation tools? If yes, please share your experience and describe the lessons and insights you have learned from it. If not, what are your expectations and concerns about this collaboration?” there were more negative opinions. None of the respondents had experience using it, and only one person considered using it in the future. Opponents emphasized several shortcomings and limitations in the current existing AI tool. Firstly, customizing the actual brand customer attributes is difficult to implement on general AI tools, also it is difficult to combine marketing analysis with prompts to generate images, making it difficult to generate products that meet market demand. Additionally, some fashion experts noted that the similarity between the text description and the output has a big gap. The text-to-image tools need to be more accurate to understand style prompts and truthfully display text details in the future. One respondent even expressed a unique aspect that the AI images generated on the platform might be taken as data by other brands. There is a risk of leaking trade secrets, and these ideas might be used by competing brands and enable them to seize business opportunities. Without any regulation of intellectual property rights to restrict other users of AI database circulation, concerns still exist.

4.1.2 Human Evaluation, FID Score and Clip Score

Calculating the FID score for a single AI image cannot get an accurate result since the FID score represents the performance of the GAN model by measuring the feature vector distance of real and AI images, therefore, Two large datasets of both types of images are needed to provide sufficient features. This ensures the calculation reflects the fidelity of the GAN model accurately. However, although the FID score for a single image cannot fully represent the performance of the GAN model, The scores still provide a reference for relative quality comparison between generated images. Therefore, it can still be used to find correlations by comparing with the expert scores and Clip scores.

As mentioned in the related work, the current existing evaluation method for AI images might not fully capture human aesthetics or evaluate advanced generation models developed recently. Thus, to understand the correlation between each criterion

of automatic and human evaluations, the correlation coefficient is calculated as shown in Figure 4.4.

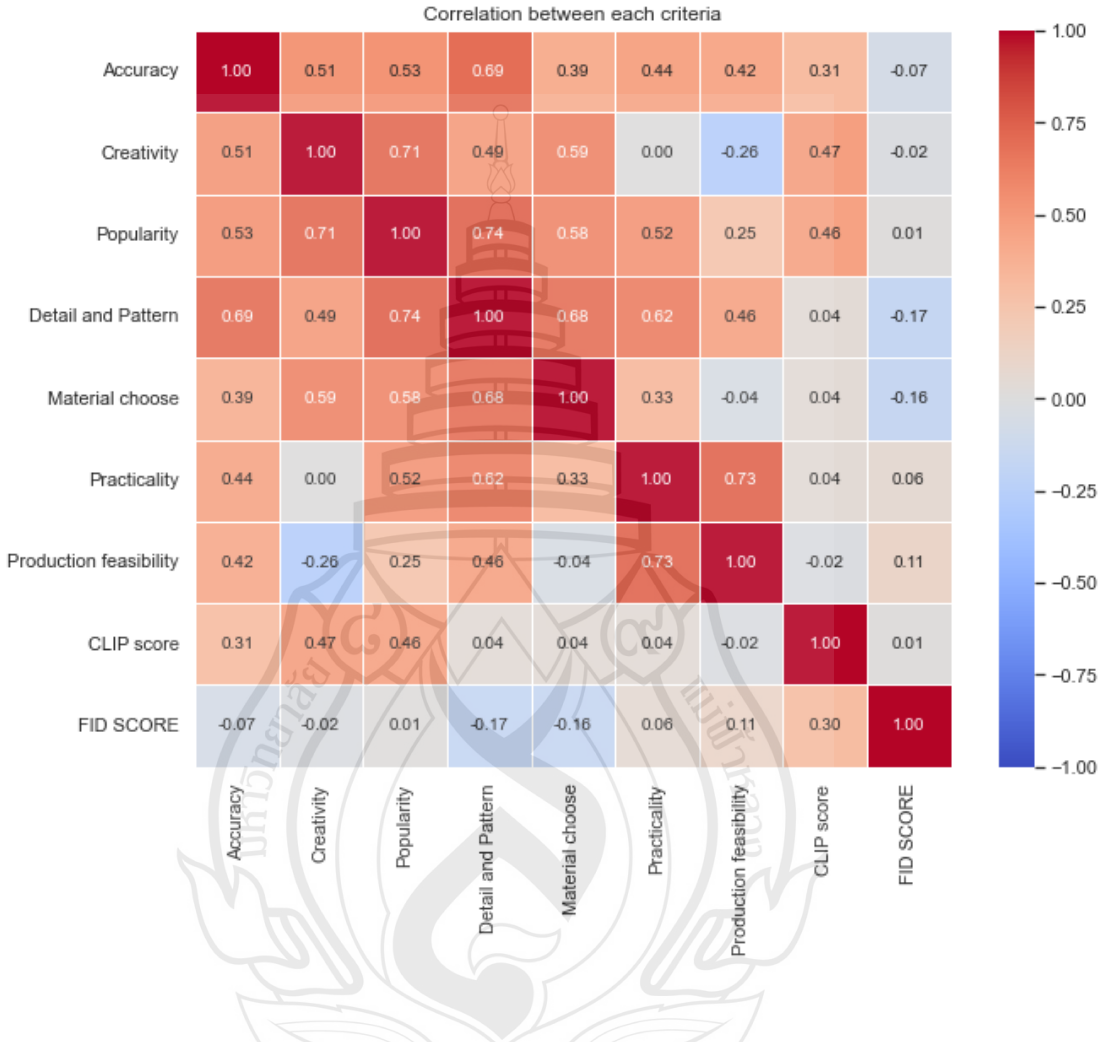


Figure 4.4 Heat map of the correlation coefficient between each criterion

The figure indicates that some criteria show high correlations. The high correlations between “Creativity” and “Detail and pattern.” and “Popularity” suggest that these criteria may overlap, indicating redundancy. In future surveys, these three scoring criteria could be integrated to avoid excessive repetition. Also, the correlation between each criterion of human evaluation and AI evaluation shows a low correlation.

The correlation coefficient between the FID score, Clip score, and the average of expert scores is shown in Figure 4.5. The weak correlation between AI evaluation

and human evaluation indicates that the two types of evaluation have significantly different standards for ranking image quality. Human evaluations are influenced by personal preferences and experience, while AI evaluations are based on objective performance metrics. Therefore, both methods provide valuable but distinct insights into exploring the performance of AI-generated fashion.

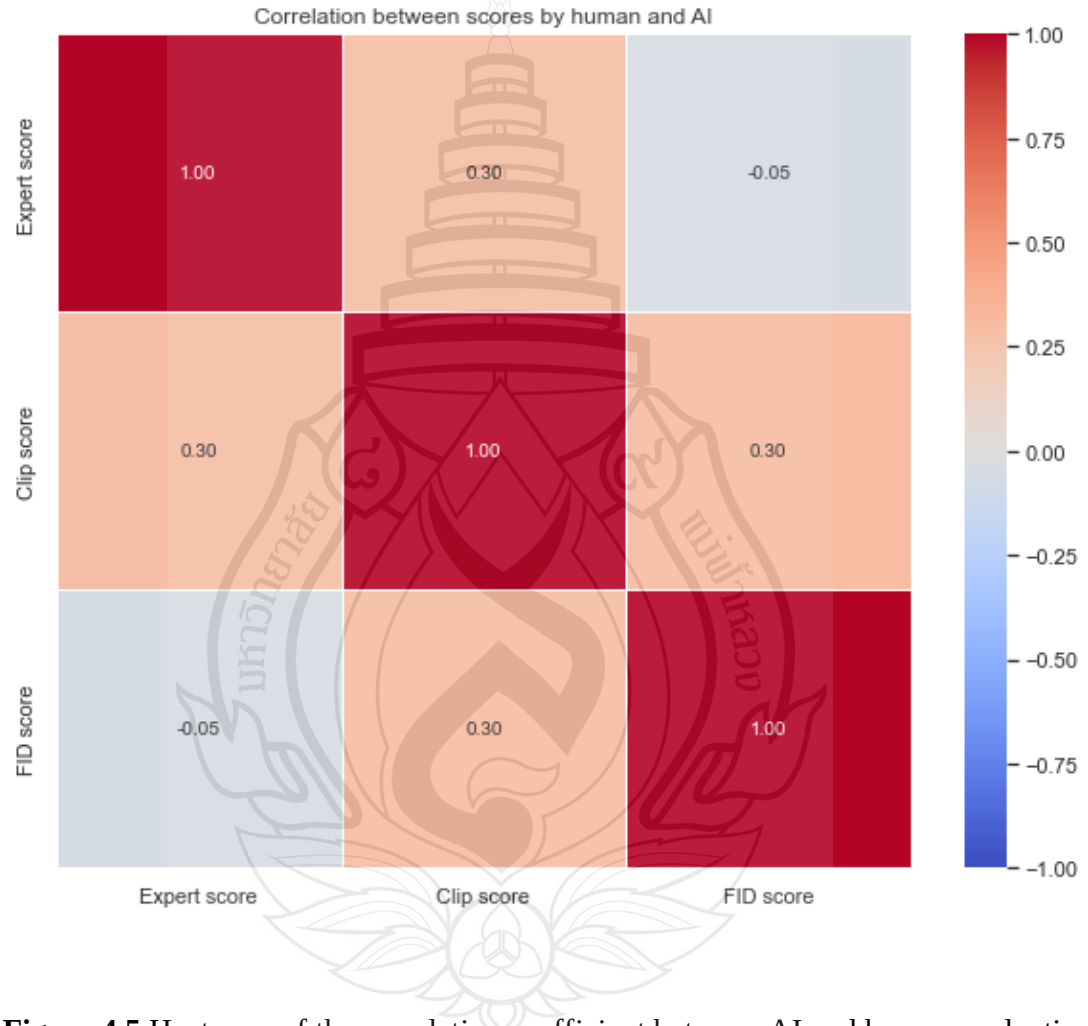


Figure 4.5 Heat map of the correlation coefficient between AI and human evaluation

4.1.3 Sufficiency for Evaluating AI Fashion Images Base on Experiment Result

From the experimental results, it can be seen that each evaluation method in the experiment can only provide reference value for their respective evaluation purposes. However, the low correlation indicates that the existing AI evaluation and human evaluation are not suitable for AI fashion images, and they cannot combine the value

of human aesthetics and AI judgment. In addition, both FID score and expert rating have database size limitations. As explained above, if only the FID score of a single AI image is calculated, it cannot effectively display the actual performance, so it is not suitable for evaluating AI fashion images. On the other hand, expert ratings provide a perspective on human aesthetics, but the sample size is not large enough, and the results show that there may be subjective aesthetic differences among experts, so there is not much difference compared to choosing public ratings. Realistically speaking, it is impossible for ordinary people to find and ask experts to help grade AI fashion design.

Based on the above experimental conclusions, to effectively apply the method in the AI fashion field, a new evaluation method needs to be established with the following characteristics:

1. Sufficiently large data size
2. Incorporation of public and market perspectives
3. AI judgment capabilities

Therefore, in the second part, we will propose two new methods based on Convolutional neural network (CNN).

4.2 Experiment Result of Creating a New Evaluation Method

4.2.1 The Solution for Data Imbalance and CNN Classification

This study encountered two main difficulties in the early stages of the experiment. The first difficulty is imbalanced data. During the data collection process, no dataset suitable for this research requirement, containing both product scores and images, was found. Therefore, product scores and images had to be collected from fashion websites through web scraping, and AI generation tools were used to convert the original images into AI images to create a dataset. However, fashion websites typically promote and display popular products with high ratings, resulting in serious data imbalance. Approximately 70% of the data has scores between 4 and 5, while scores between 3 and 4 are virtually missing.

The second challenge is the adaptability of the CNN classification model. At the beginning of this study, we hypothesized that even with many classes, the

classification model would still be able to extract features and classify them effectively. Therefore, we initially attempted to use only the CNN classification model for training without processing the dataset labels. However, the results revealed many problems that needed to be addressed. Firstly, the presence of too many classes and an insufficiently large dataset prevented the model from learning and predicting effectively, resulting in an accuracy of only about 50%. This also confirmed that when the labels are continuous numbers, the aesthetic quality assessment is more appropriately handled as a regression problem. Additionally, due to the aforementioned data imbalance, the model suffered from serious overfitting issues.

To address the first problem, we added a regression model to predict the original labels. For the classification model, we preprocess the labels by dividing ratings from 1 to 5 into intervals to reduce the number of classes, thereby increasing accuracy. We evaluate the performance of both methods in parallel. Additionally, we aim to enlarge the dataset by collecting more data from fashion websites. However, solving the issue of data imbalance is challenging, as it is difficult to collect data for products with lower scores. To mitigate this, we use the low scores from a few customer comments as labels to gather data for lower-scored products, covering the missing part of the data. On the other hand, argumentation will be applied only on the picture with the label 1-4 to solve the data imbalanced issue.

4.2.2 Result of CNN Classification Model

Figure 4.6 The application of the classification model

Overfitting and low accuracy are the two most difficult problems in the process of training CNN classification. The data structure only reduces the degree of data imbalance after collecting additional low-scoring data, but the high-scoring data still accounts for the majority, so the accuracy of training the model directly using the original data is only 72%, and the training results are also overfitting. Therefore, the problems need to be solved through argumentation and dropout. However, after applying argumentation and dropout, the accuracy is only 34%, even lower than the original result. Therefore, we have decided to try removing dropout. Dropout is a regularization technique used to prevent overfitting, but if the dropout rate is set too high, too many neurons are discarded during the training process. This can lead to insufficient information being learned by the model, affecting its convergence and overall performance. Surprisingly, after attempting to cancel dropout and only use argumentation, the experimental results achieved a considerable level of accuracy. Table 4.3 shows the performance analysis of the model after argumentation, and Figure 4.7. is the heatmap of the confusion matrix.

Training Set							
TARGET / OUTPUT	Class0	Class1	Class2	Class3	Class4	Class5	SUM
Class0	14 6.39%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	5 2.28%	19 73.68% 26.32%
Class1	0 0.00%	6 2.74%	0 0.00%	0 0.00%	1 0.46%	3 1.37%	10 60.00% 40.00%
Class2	0 0.00%	0 0.00%	12 5.48%	0 0.00%	0 0.00%	2 0.91%	14 85.71% 14.29%
Class3	0 0.00%	0 0.00%	0 0.00%	8 3.65%	0 0.00%	2 0.91%	10 80.00% 20.00%
Class4	0 0.00%	0 0.00%	0 0.00%	0 0.00%	7 3.20%	1 0.46%	8 87.50% 12.50%
Class5	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	158 72.15%	158 100.00% 0.00%
SUM	14 100.00% 0.00%	6 100.00% 0.00%	12 100.00% 0.00%	8 100.00% 0.00%	8 87.50% 12.50%	171 92.40% 7.60%	205 / 219 93.61% 6.39%

Figure 4.7 The confusion matrix

Table 4.2 Classification model performance

Average Confidence Score:	0.939
Accuracy:	0.936
Precision:	0.940
Recall:	0.936
F1 Score:	0.932

1. Average Confidence Score: The model's average confidence score is 0.939, showing that the model is highly confident in most of its predictions. This high confidence reduces the risk of misclassification.

2. Accuracy: The model achieved an accuracy of 0.936, correctly classifying 93.6% of the test samples. This indicates strong overall classification performance.

3. Precision: With a precision of 0.941, 94.1% of the samples predicted as positive were indeed positive. This means the model is effective in identifying true positives with few false positives.

4. Recall: The recall is 0.936, meaning the model correctly identified 93.6% of actual positive samples. This shows the model's ability to detect true positives with few false negatives.

5. F1 Score: The F1 score is 0.932, balancing precision and recall. This high F1 score indicates that the model effectively identifies positive samples while minimizing both false positives and false negatives.

4.2.3 Result of CNN Regression Model

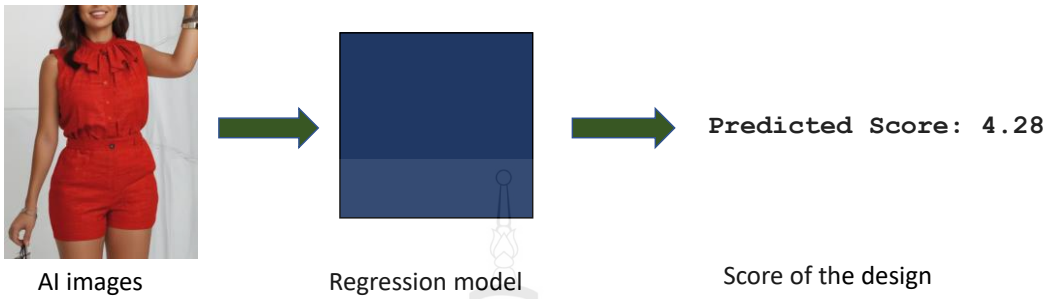


Figure 4.8 The application of the regression model

To establish a scale of 1-5, it is necessary to set the range of prediction scores accordingly. However, in the prediction of regression models, the scores often exceed this range. Therefore, the actual results need to be constrained by methods such as normalization, which scales the scores to the desired range. Nevertheless, normalization can sometimes weaken the specific characteristics of the data, thereby affecting subsequent analysis. For instance, an original prediction of 5 may be altered due to proportional scaling, thus affecting the original distribution within the range.

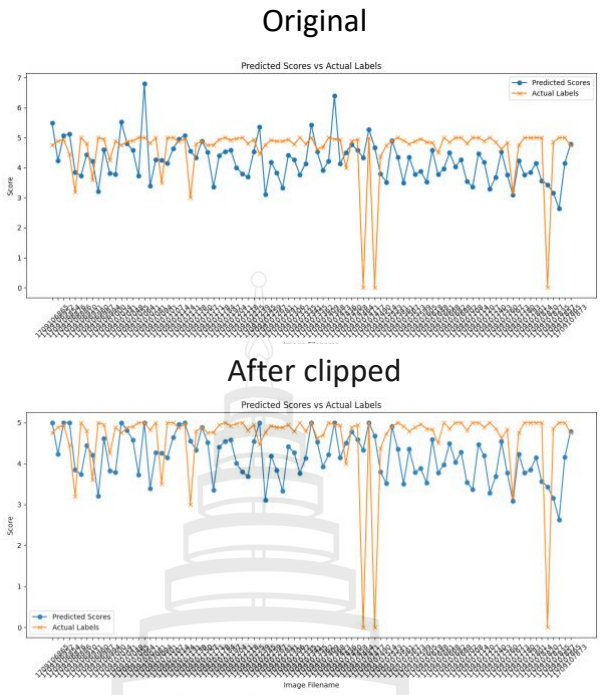


Figure 4.9 The distribution of the prediction score

In this study, we address this issue by using NumPy to clip the predicted values. This method adjusts only the values that fall outside the specified range, leaving the other data unchanged. Consequently, it avoids altering the overall distribution of the data and prevents extreme values from influencing subsequent analysis. Figure 4.9 shows the distribution comparison of the processing clip.

In the model training, we first set a random seed to ensure reproducibility of the results. Then, we define a Convolutional Neural Network (CNN) model, which includes data augmentation, multiple convolutional layers, pooling layers, fully connected layers, and an output layer. The model is compiled using the Adam optimizer with mean squared error (MSE) loss and mean absolute error (MAE) as the evaluation metric. The model is trained on the training dataset and validated on the validation dataset. Specific training parameters include epochs = 10 (indicating the number of training epochs is 10), batch size = 32 (representing the number of samples per batch is 32), and validation split = 0.2 (indicating 20% of the training data is used for validation). Finally, the

trained model is used to make predictions on the test dataset, and the prediction results are, and the prediction results are printed.

Table 4.3 Regression model performance

Evaluation	RMSE	MAE	MAPE
Average	0.866	0.682	22.16

Table 4.3 shows the RMSE and MAE with the comparison between the predicted score and the real score. Due to the slight variations in results obtained from each training session, the average Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) from ten training sessions were calculated to summarize the predictive performance of the model. This approach provides a more robust assessment of the model's performance by accounting for the variability inherent in the training process.

In the model, the RMSE is 0.866. This indicates that, on average, the squared difference between the predicted values and the actual values is 0.866. MSE is particularly sensitive to outliers, as it squares the errors, giving more weight to larger errors. A lower MSE value signifies better predictive accuracy, indicating that our model performs well in minimizing larger errors.

On the other hand, the model's MAE is 0.682, which means the average absolute difference between the predicted values and the actual values is 0.682. Unlike RMSE, MAE does not square the errors, thus it provides a linear score that gives equal weight to all errors. This makes MAE more interpretable and less sensitive to outliers compared to RMSE.

In addition, we also added MAPE to measure the percentage of error between predicted and actual values. The model's MAPE is 22.36, which means that the average relative error between predicted and actual values is 22.36%, indicating that the model has some accuracy, but there is still room for improvement. It also represents that it may be acceptable in some applications, but further improvement may be needed in applications with high precision requirements.

4.3 Experimental Environment

All the experiments in this study are run through MacBook Air 2020 with 1.1 GHz quad-core processor i5 Intel Core, Iris Plus Graphics 1536 MB. Using python with Numpy, Tensorflow and Matplotlib package on Google Colab.

CHAPTER 5

CONCLUSION

5.1 Conclusion and Discussion

Although the AI tool has gradually been used frequently and become popular in recent years, there are still many limitations in its application in the fashion industry. For most industry professionals, using AI as an auxiliary tool remains an unfamiliar working mode. Also, the low performance of both expert and CLIP scores and the responses from the experts indicate that artificial intelligence drawing software needs further development and optimize the accuracy and completeness of generated images to be used more effectively for design purposes in the future, particularly for GAN models used in text-to-image generation. These models currently lack accurate word comprehension, resulting in images that often differ from expected outcomes, significantly increasing the difficulty of use. Additionally, unreasonable image outputs are frequently encountered during the generation process, significantly reducing the tool's efficiency.

This study also attempts to compare the existing methods for evaluating the performance of AI-generated fashion images and find an effective evaluation method for AI fashion design. However, the results suggest that fashion AI images cannot be fully captured by either AI or human evaluations alone. Initially, mainstream fashion rating standards and commonly used AI-generated image rating standards were extracted to serve as the foundation for a new evaluation method. However, after a series of analyses, it is clear that both of these evaluation methods have flaws and that existing evaluation methods cannot provide accurate rankings. For instance, overlap in fashion design criteria necessitates a re-examination of whether these overlaps should be merged. Moreover, the survey targeted fashion experts to collect professional perspectives. However, finding enough experts for the survey proved challenging.

Additionally, after analysis, it was found that professionals often rank scores based on personal preferences and subjective judgment, making their evaluations not significantly different from those of the general public and resulting in an insufficient sample size.

Although the experiment indicated that existing methods do not align with the evaluation of AI fashion images, the identified shortcomings provide new directions for developing an improved evaluation method. Firstly, the choice of data sources is crucial. By utilizing customer ratings, we can effectively capture the general public's views on designs, providing a large dataset. This approach also offers additional market preference insights for brands or design companies. Secondly, improvements in practicality can be achieved by training predictive models and addressing the difficulties of human evaluations by experts or market surveys. Training a CNN predictive model is a solution that combines AI with human aesthetic perspectives. Using human evaluation scores as data, regression and classification models can be trained to predict scores, thereby endowing AI with an understanding of human fashion aesthetics.

In summary, the CNN model in this study demonstrates excellent performance in the classification task, achieving a high average confidence score, accuracy, precision, recall, and F1 score. The confusion matrix further confirms the model’s accuracy across different categories. Overall, these results suggest that the designed CNN model can effectively perform image classification tasks, exhibiting good generalization ability and classification accuracy. This provides a solid foundation and confidence for further application to larger and more complex datasets.

On the other hand, the regression model, the values of RMSE and MAE indicate that the model has good predictive accuracy. The relatively low values of these indicators prove that the model can provide convincing results for predicting aesthetic or market scores in design. However, on a scale of 0-5, There is still room for improvement. Also, after observing the data distribution results, due to the issue of data imbalance, the vast majority of prediction scores are concentrated between 4 and 5 points, for images with lower scores, the prediction gap is relatively large.

5.2 Limitation and Future Work

Both methods demonstrate their feasibility in evaluating AI-generated fashion images and provide a benchmark for the application of generative AI in the fashion industry. Compared to the two CNN models, the regression model can provide more accurate prediction scores, offering clearer indicators when making choices among multiple designs. However, the classification model, due to its limited categories, can only provide a rough estimate of the design scores. Despite this limitation, the classification model still contributes to understanding potential market preferences and aesthetic evaluations of the designs. For designers, it can still provide a benchmark for preliminary examination and screening of AI fashion design.

Nevertheless, this study encountered several difficulties during the experimental process, especially in data acquisition, firstly, there is no complete database that matches customer ratings with images. Some sources of customer ratings are neither comprehensive nor specific enough, limiting the quality and scope of data for model training and evaluation. Additionally, due to incomplete data, biases may arise during the model training process, affecting the accuracy and reliability of the prediction results. Secondly, the issue of data imbalance is significant, with most prediction scores concentrated between 4 and 5, while the prediction errors for low-scoring images are relatively large. This indicates that the model performs better with high-scoring designs than low-scoring ones, limiting its applicability for comprehensive design evaluation. Thirdly, the AI images used for training the model are generated using generative AI to resemble real images. However, since most current tools cannot generate images in bulk, each image needs to be created individually, resulting in a time-consuming process for acquiring AI images.

To optimize the regression and classification models, addressing data imbalance and expanding dataset size are crucial. Larger and more balanced datasets are expected to enhance model accuracy. Additionally, a larger dataset can enable classification divided into finer discrete intervals, offering more precise measurements. Therefore, in further research, seeking more sources to collect more comprehensive customer ratings and corresponding image data to build a more balanced and diverse dataset is necessary.

In addition, to improve the model, introducing techniques to handle data imbalance within the model, such as weighted loss functions or resampling techniques, can also be a chance to optimize the model's prediction capabilities across different score ranges.

In this study, customer ratings were used as labels to reflect consumer market evaluations and aesthetic preferences. For predicting other aspects, such as potential sales volumes, appropriate labels like sales data can be utilized. Further optimization can be achieved by incorporating underlying causal factors as model parameters, potentially providing more insightful predictive results.

REFERENCES

REFERENCES

Areeb, Q. M., Imam, R., Fatima, N., & Nadeem, M. (2021). AI art critic: Artistic classification of poster images using neural networks. In *2021 International Conference on Data Analytics for Business and Industry (ICDABI)* (pp. 37-41). IEEE.

Azza, N. (2023). *What is Leonardo AI: Everything you need to know*. https://www.digitbin.com/what-is-leonardo-ai/

BoF. (2023). *The year ahead: How gen AI is reshaping fashion's creativity*. https://www.businessoffashion.com/articles/technology/the-state-of-fashion-2024-report-generative-ai-artificial-intelligence-technology-creativity/

CLIP score. (n.d.). CLIP Score - PyTorch-Metrics 1.1.0 documentation. Retrieved February 6, 2024, from https://torchmetrics.readthedocs.io/en/stable/multimodal/clip_score.html

Das, S. (2023). *6 gan architectures you really should know, neptune.ai*. https://neptune.ai/blog/6-gan-architectures

Drew, D., & Yehounme, G. (2017). *The apparel industry's environmental impact in 6 graphics*. https://www.wri.org/insights/apparel-industrys-environmental-impact-6-graphics

Dshahid380. (2019). *Convolutional neural network, medium*. https://towardsdatascience.com/covolutional-neural-network-cb0883dd6529

Example of a CNN for image classification. (n.d.). Retrieved February 6, 2024, from https://www.researchgate.net/figure/Example-of-a-CNN-for-image-classification_fig1_332284670

Farfetch. (n.d.). *FARFETCH - The global destination for modern luxury*. Retrieved 6 February 2024, from https://www.farfetch.com/th/shopping/men/items.aspx

Fashion Discounts. (2023). *27 revealing fast fashion statistics for 2023, fashion discounts*. https://fashiondiscounts.uk/fast-fashion-statistics/

Fernando, J. (n.d.). *The correlation coefficient: What it is, what it tells investors, Investopedia*. Retrieved February 6, 2024, from https://www.investopedia.com/terms/c/correlationcoefficient.asp

Generative Adversarial Network. (2023). https://en.wikipedia.org/wiki/Generative_adversarial_network

Gh, S. (2019). *Deconstructing the fashion design process, medium*. Retrieved August 14, 2023, from https://medium.com/@fuel4fashion/deconstructing-the-fashion-design-process-fd55dbd2259f

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). *Gans trained by a two time-scale update rule converge to a local nash equilibrium*. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf

Jin, B., Segovia, M. V. O., & Süsstrunk, S. (2016). Image aesthetic predictors based on weighted CNNs. In *2016 IEEE International Conference on Image Processing (ICIP)* (pp. 2291-2295). IEEE.

Kao, Y., Wang, C., & Huang, K. (2015). Visual aesthetic quality assessment with a regression model. In *2015 IEEE International Conference on Image Processing (ICIP)* (pp. 1583-1587). IEEE.

Leonardo.ai. (2023). *Practical drawing strategy - powerful free AI drawing solution*. https://www.techbang.com/posts/106005-leonardoai-drawing-practice-guide-powerful-free-ai-drawing

Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016). *Large-scale fashion (deepfashion) database*. The Chinese University of Hong Kong, Category and Attribute Prediction Benchmark, Xiaoou TangMultimedia Laboratory.

Maket.us. (2024). *Generative AI in fashion market*. https://market.us/report/generative-ai-in-fashion-market/

Marin, I., Gotovac, S., & Russo, M. (2020). Evaluation of generative adversarial network for human face image synthesis. In *2020 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)* (pp. 1-6). IEEE.

Otani, M., Togashi, R., Sawai, Y., Ishigami, R., Nakashima, Y., Rahtu, E., . . . Satoh, S. I. (2023). Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14277-14286). IEEE.

Panopticon. (2023). *What is generative adversarial networks (GAN) art?*. https://panopticon.am/what-is-generative-adversarial-networks-gan-art/

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., . . . Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

Seo, Y., & Shin, K. S. (2018, March). Image classification of fine-grained fashion image based on style using pre-trained convolutional neural network. In *2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA)* (pp. 387-390). IEEE.

Sharma, D. (2024). *Image classification using CNN | Step-wise tutorial. analytics Vidhya*. https://www.analyticsvidhya.com/blog/2021/01/image-classification-using-convolutional-neural-networks-a-step-by-step-guide/

Shena, D., Sheaffb, C., Chen, G., Guoa, M., Blaschc, E., & Phamd, K. (2022). *General-sum game modeling of generative adversarial networks for satellite maneuver detection*. https://amostech.com/TechnicalPapers/2022/Machine-Learning-for-SSA-Applications/Shen.pdf

Smith, N. (2022). *How is a garment made? Points of measure*. https://www.pointsofmeasure.com/tutorials-education/the-design-process-in-9-simple-steps

Uni Matrix Zero. (2023). *Evaluating AI-generated images with clip score*. https://unimatrixz.com/blog/latent-space-clip-score/

Woottisart, P., Sripian, P., & Thanasuan,K. (2022). The study of fashion style classification: Harajuku-type Kawaii and street fashion. In *16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Dijon, France, 2022* (pp. 402-408). https://doi.org/10.1109/SITIS57111.2022.00067

Xuan, X., Han, R., Ji, S., & Ding, B. (2021, March). Research on clothing image classification models based on CNN and Transfer Learning. In *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)* (vol. 5, pp. 1461-1466). IEEE.

Yan, H., Zhang, H., Liu, L., Zhou, D., Xu, X., Zhang, Z., . . . Yan, S. (2022). Toward intelligent design: An AI-based fashion designer using generative adversarial networks aided by sketch and rendering generators. *IEEE Transactions on Multimedia, 25*. https://doi.org/10.1109/TMM.2022.3146010

Zhang, C., Zhang, C., Zhang, M., & Kweon, I. S. (2023). *Text-to-image diffusion model in generative ai: A survey*. https://arxiv.org/abs/2303.07909

Zhang, K., Zhu, D., Min, X., Gao, Z., & Zhai, G. (2023). *Synergetic assessment of quality and aesthetic: Approach and comprehensive benchmark dataset*. IEEE Transactions on Circuits and Systems for Video Technology.

APPENDIX

APPENDIX

CONTENT OF THE SURVEY FOR EXPERTS TO EVALUATE FASHION AI IMAGE

Text

Survey for expert to evaluate fashion AI image in Appendix only.

Figure A1 Introduction

1. 芥末黃色T恤，左側有口袋，採用透氣棉質面料制成，帶有戶外或露營風格圖案，寬鬆版型。

	1	2	3	4	5
準確性	○	○	○	○	○
創意性	○	○	○	○	○
流行性	○	○	○	○	○
細節和版型	○	○	○	○	○
材料選擇	○	○	○	○	○
實用性	○	○	○	○	○
生產可行性	○	○	○	○	○

Figure A2 Image 1

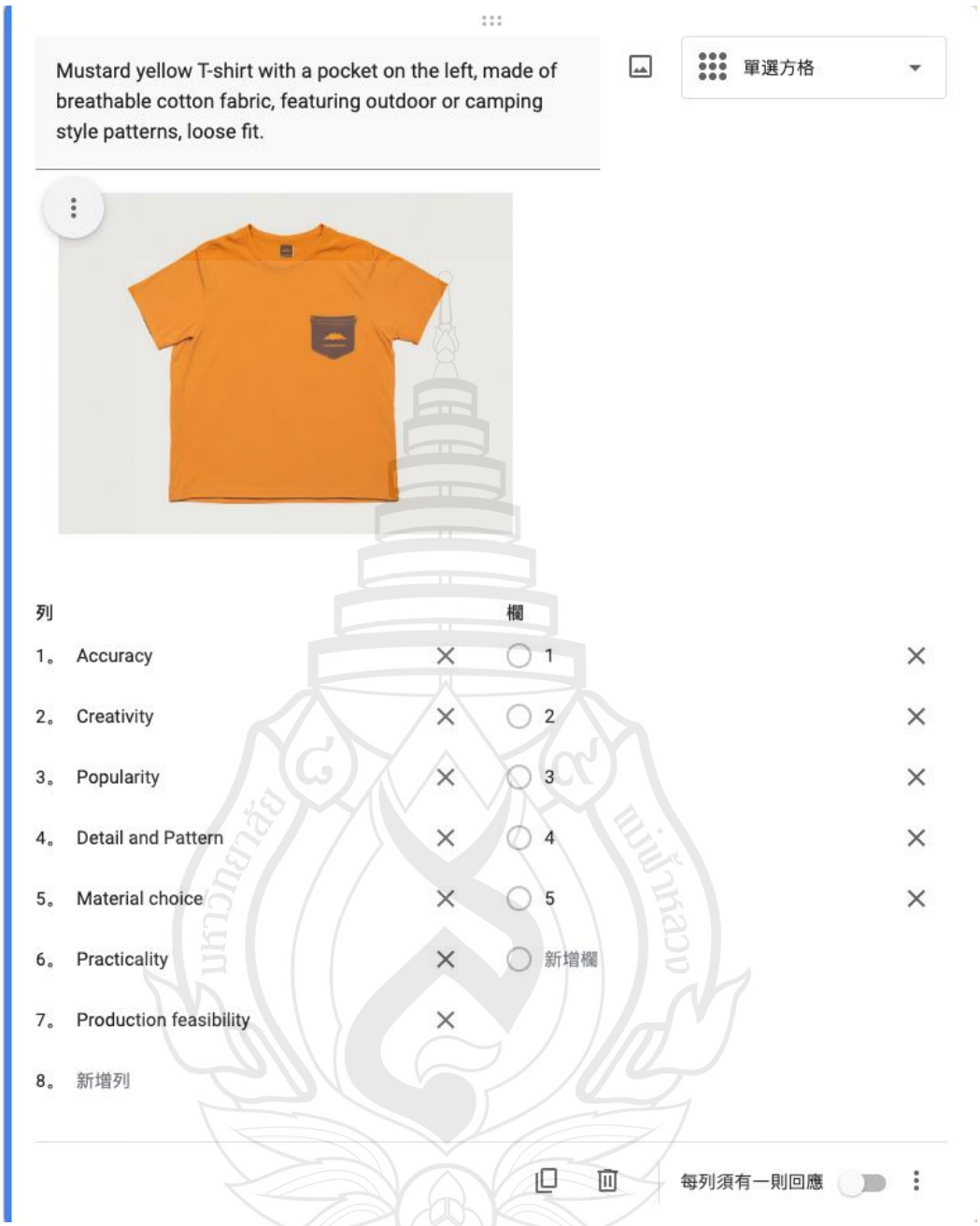


Figure A3 Image 2

3. 海軍藍亨利衫，厚棉面料，經典鬆緊袖口和短袖，日式波浪圖案，面料上波紋，寬鬆版型，圖片正面，僅有衣服

	1	2	3	4	5
準確性	○	○	○	○	○
創意性	○	○	○	○	○
流行性	○	○	○	○	○
細節和版型	○	○	○	○	○
材料選擇	○	○	○	○	○
實用性	○	○	○	○	○
生產可行性	○	○	○	○	○

Figure A4 Image 3

4. 胸口中間有卡通北極熊圖案的白色T恤，寬鬆版型。

	1	2	3	4	5
準確性	○	○	○	○	○
創意性	○	○	○	○	○
流行性	○	○	○	○	○
細節和版型	○	○	○	○	○
材料選擇	○	○	○	○	○
實用性	○	○	○	○	○
生產可行性	○	○	○	○	○

Figure A5 Image 4

5. 酒紅色連帽衫，長袖，印有復古舊報紙圖案，中間印有經典 FILA 標誌

	1	2	3	4	5
準確性	○	○	○	○	○
創意性	○	○	○	○	○
流行性	○	○	○	○	○
細節和版型	○	○	○	○	○
材料選擇	○	○	○	○	○
實用性	○	○	○	○	○
生產可行性	○	○	○	○	○

Figure A6 Image 5

6. 淺米色針織外套，帶有精致的銀色紐扣裝飾，純衣服照片

	1	2	3	4	5
準確性	○	○	○	○	○
創意性	○	○	○	○	○
流行性	○	○	○	○	○
細節和版型	○	○	○	○	○
材料選擇	○	○	○	○	○
實用性	○	○	○	○	○
生產可行性	○	○	○	○	○

Figure A7 Image 6

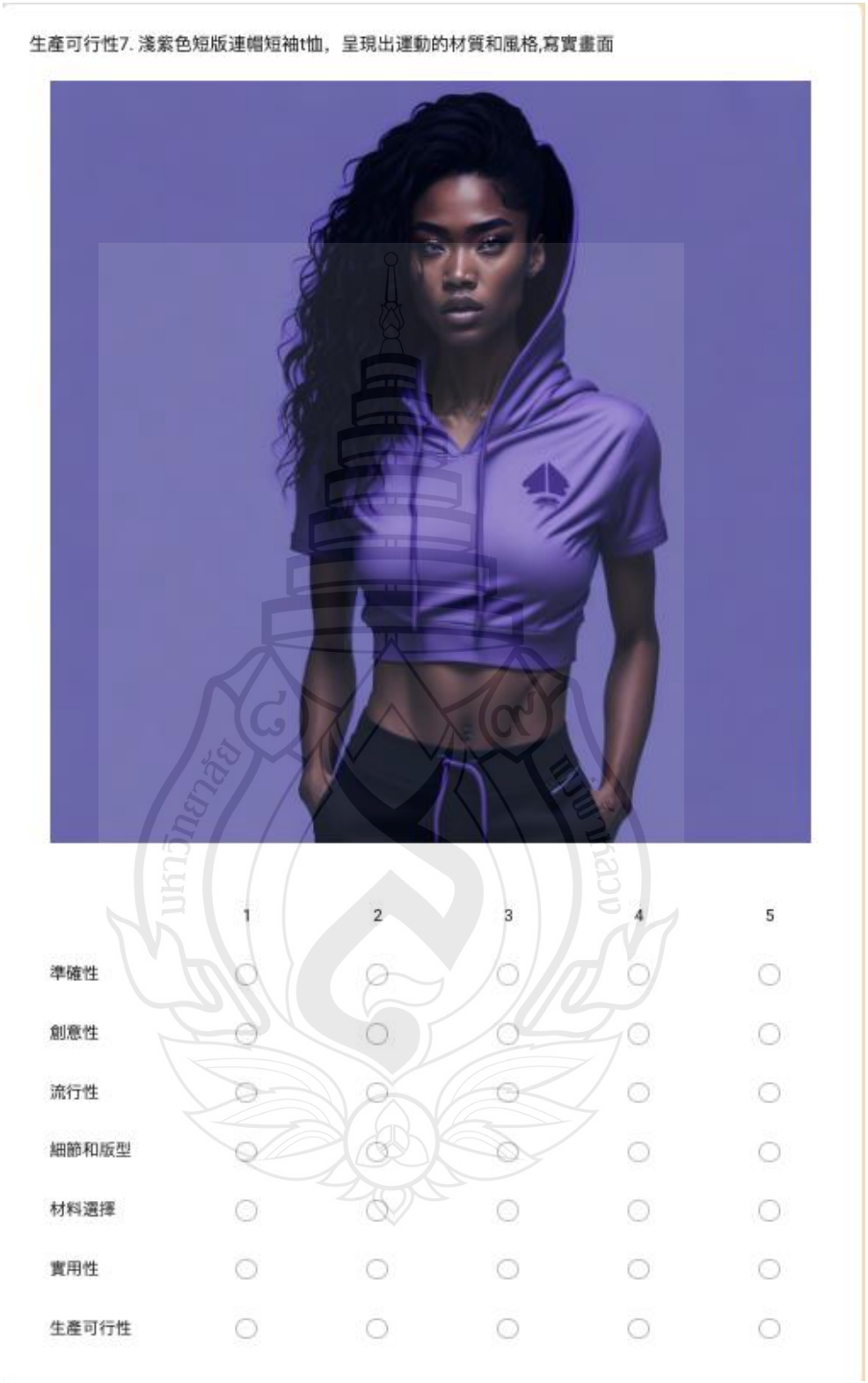


Figure A8 Image 7

Figure A9 Image 8

Figure A10 Image 9

Figure A11 Image 10

11. 米色打褶西裝褲，小格紋圖案，由奢華感的絲綢材料制成，剪裁修身，配有完整的腿部照片。

	1	2	3	4	5
準確性	○	○	○	○	○
創意性	○	○	○	○	○
流行性	○	○	○	○	○
細節和版型	○	○	○	○	○
材料選擇	○	○	○	○	○
實用性	○	○	○	○	○
生產可行性	○	○	○	○	○

Figure A12 Image 11

12. 灰色打褶西裝褲，小格紋圖案，由奢華感的絲綢材料制成，剪裁修身，配有完整的腿部照片。

	1	2	3	4	5
準確性	○	○	○	○	○
創意性	○	○	○	○	○
流行性	○	○	○	○	○
細節和版型	○	○	○	○	○
材料選擇	○	○	○	○	○
實用性	○	○	○	○	○
生產可行性	○	○	○	○	○

Figure A13 Image 12

13. 一件黑色粗花呢羊毛襯衫，寬鬆版型，以80年代復古流行文化為設計靈感

	1	2	3	4	5
準確性	○	○	○	○	○
創意性	○	○	○	○	○
流行性	○	○	○	○	○
細節和版型	○	○	○	○	○
材料選擇	○	○	○	○	○
實用性	○	○	○	○	○
生產可行性	○	○	○	○	○

Figure A14 Image 13

14. 一件黑色粗獷的羊毛襯衫，寬鬆的款式，邊緣有黃色的裝飾，陽剛風格，襯衫的全貌。

	1	2	3	4	5
準確性	○	○	○	○	○
創意性	○	○	○	○	○
流行性	○	○	○	○	○
細節和版型	○	○	○	○	○
材料選擇	○	○	○	○	○
實用性	○	○	○	○	○
生產可行性	○	○	○	○	○

Figure A15 Image 14

15. 淺黃色繡紗襯衫，寬鬆版形，符合微風徐徐的夏日氛圍。

	1	2	3	4	5
準確性	○	○	○	○	○
創意性	○	○	○	○	○
流行性	○	○	○	○	○
細節和版型	○	○	○	○	○
材料選擇	○	○	○	○	○
實用性	○	○	○	○	○
生產可行性	○	○	○	○	○

Figure A16 Image 15

16. 一款精致柔美的迷你連衣裙,細致的橙色和白色雛菊花紋。

	1	2	3	4	5
準確性	○	○	○	○	○
創意性	○	○	○	○	○
流行性	○	○	○	○	○
細節和版型	○	○	○	○	○
材料選擇	○	○	○	○	○
實用性	○	○	○	○	○
生產可行性	○	○	○	○	○

Figure A17 Image 16

17. 一款精致柔美的迷你連衣裙,細致的橙色和白色雛菊花紋。

	1	2	3	4	5
準確性	○	○	○	○	○
創意性	○	○	○	○	○
流行性	○	○	○	○	○
細節和版型	○	○	○	○	○
材料選擇	○	○	○	○	○
實用性	○	○	○	○	○
生產可行性	○	○	○	○	○

Figure A18 Image 17

Figure A19 Image 18

19. 時尚短版襯衫，飾有精致的鏤空刺繡和金色紐扣，袖口和下擺邊以藍色作點綴。

	1	2	3	4	5
準確性	○	○	○	○	○
創意性	○	○	○	○	○
流行性	○	○	○	○	○
細節和版型	○	○	○	○	○
材料選擇	○	○	○	○	○
實用性	○	○	○	○	○
生產可行性	○	○	○	○	○

Figure A20 Image 19

20. 一條燈芯絨混紡長裙，磚紅色，寬大舒適版型，適合寒冷的秋天。

	1	2	3	4	5
準確性	○	○	○	○	○
創意性	○	○	○	○	○
流行性	○	○	○	○	○
細節和版型	○	○	○	○	○
材料選擇	○	○	○	○	○
實用性	○	○	○	○	○
生產可行性	○	○	○	○	○

Figure A21 Image 20

21. Oversized 牛仔夾克，用閃閃發光的珍珠和精緻的刺繡作點綴。

	1	2	3	4	5
準確性	○	○	○	○	○
創意性	○	○	○	○	○
流行性	○	○	○	○	○
細節和版型	○	○	○	○	○
材料選擇	○	○	○	○	○
實用性	○	○	○	○	○
生產可行性	○	○	○	○	○

Figure A22 Image 21

Figure A23 Image 22

23. 材質柔軟舒的滿版印刷黃色帽T

	1	2	3	4	5
準確性	○	○	○	○	○
創意性	○	○	○	○	○
流行性	○	○	○	○	○
細節和版型	○	○	○	○	○
材料選擇	○	○	○	○	○
實用性	○	○	○	○	○
生產可行性	○	○	○	○	○

Figure A24 Image 23

24. 黑色百褶長裙，搭配深金色鬆緊梭織腰帶，運動感且舒適。

	1	2	3	4	5
準確性	○	○	○	○	○
創意性	○	○	○	○	○
流行性	○	○	○	○	○
細節和版型	○	○	○	○	○
材料選擇	○	○	○	○	○
實用性	○	○	○	○	○
生產可行性	○	○	○	○	○

Figure A25 Image 24

25. 一件粉紅色短裙，滿版塗鴉風格玫瑰圖案

	1	2	3	4	5	6
準確性	○	○	○	○	○	○
創意性	○	○	○	○	○	○
流行性	○	○	○	○	○	○
細節和版型	○	○	○	○	○	○
材料選擇	○	○	○	○	○	○
實用性	○	○	○	○	○	○
生產可行性	○	○	○	○	○	○

Figure A26 Image 25

CURRICULUM VITAE

CURRICULUM VITAE

NAME Hsi Yeh Wang

EDUCATIONAL BACKGROUND

2016 Bachelor of Social Science
Political Economy
National Sun Yat-sen University

WORK EXPERIENCE

2021-2022 Product Project Manager
FILA Taiwan

2019-2021 Sales Specialist
Tiong Liong Industrial Co., Ltd

PUBLICATION

Wang, H. Y., & Utama, S. (2023, November). Investigating the generative-ai evaluation methods and correlation with fashion designers. In *2023 7th International Conference on Information Technology (InCIT)* (pp. 508-513). IEEE.