Journal of Organizational and End User Computing
Volume 36 • Issue 1

A Novel Deep Learning-Based Visual Search Engine in Digital Marketing for Tourism E-Commerce Platforms

Yingli Wu, School of Business Management, Jiaxing Nanhu University, China
Qiuyan Liu, School of Economics and Management, Zhejiang University of Water Resources and Electric Power, China*

ABSTRACT

Visual search technology, because of its convenience and high efficiency, is widely used by major tourism e-commerce platforms in product search functions. This study introduces an innovative visual search engine model, namely CLIP-ItP, aiming to thoroughly explore the application potential of visual search in tourism e-commerce. The model is an extension of the CLIP (contrastive language-image pre-training) framework and is developed through three pivotal stages. Firstly, by training an image feature extractor and a linear model, the visual search engine labels images, establishing an experimental visual search engine. Secondly, CLIP-ItP jointly trains multiple text and image encoders, facilitating the integration of multimodal data, including product image labels, categories, names, and attributes. Finally, leveraging user-uploaded images and jointly selected product attributes, CLIP-ItP provides personalized top-k product recommendations.

KEYWORDS

CLIP, CLIP-ItP, Multimodal Data, Top-K Product Recommendation, Tourism E-Commerce, Visual Search Engine

The application of visual search engines (Wolfe, 2021) in tourism e-commerce (Anand et al., 2021) holds significant importance. Firstly, visual search engines employing image recognition technology (Li, 2022) empower users to conduct product searches through images rather than text. This intuitive search method enhances user experience, reduces search time, and makes shopping more convenient and enjoyable. Subsequently, leveraging advanced deep learning models (Chen et al., 2022a), visual search engines can comprehend and interpret image content more accurately, facilitating precise product retrieval. This aids in eliminating semantic ambiguities, increasing the relevance of search results, and presenting users with products more aligned with their needs.

Moreover, visual search engines not only enable image-based searches for similar products uploaded by users but also integrate various information such as product categories and attributes. This integration results in providing users with more personalized and tailored recommendations. This, in turn, helps tourism e-commerce platforms better understand user preferences, offer personalized shopping suggestions, and enhance user engagement and satisfaction. Finally, by delivering a more intuitive, accurate, and personalized search and recommendation experience, visual search engines contribute to stimulating user purchasing desires, increasing conversion rates, and thereby fostering

DOI: 10.4018/JOEUC.340386 *Corresponding Author

sales growth. Users can swiftly locate products of interest, leading to a higher likelihood of completing purchase transactions (He, 2021). In conclusion, through visual search, users can effortlessly discover new products and brands, providing tourism e-commerce products with opportunities to expand their market and increase brand exposure. Therefore, for emerging brands or unique products, visual search serves as a potent promotional tool.

Visual search engines have witnessed numerous successful applications in real-life scenarios, and a prominent example is Google Lens. Google Lens empowers users to capture and recognize objects, text, or images through the camera of their mobile devices, providing pertinent information and operational suggestions. Users can employ Google Lens for object recognition by capturing various surroundings, such as plants, animals, and buildings. The application endeavors to identify and furnish relevant information based on the captured images. Additionally, the software facilitates text translation, allowing users to capture text and translate it into other languages, thereby enhancing understanding and communication. Moreover, the software finds utility in store and product recognition, enabling users to capture images of products in stores. The application identifies the products and provides information on prices, reviews, and purchase options. Furthermore, Google Lens extends its functionality to landmark recognition, allowing users to capture images of buildings or landmarks (Chen et al., 2022b). Google Lens recognizes and provides pertinent historical and contextual information in such cases.

From the current research perspective, the application of deep learning methods in visual search for tourism e-commerce (Hou et al., 2021) holds profound significance (Dagan et al., 2023; Du et al., 2022; Foping et al., 2022; Lepage et al., 2023; Monjur et al., 2023; Singh et al., 2023; Waqas et al., 2023; Yang Guang, 2021). Firstly, deep learning models can enhance the understanding and classification capabilities of product images by training on large-scale datasets, learning richer and more abstract feature representations. This enables tourism e-commerce platforms to match user search queries more accurately, providing more relevant product recommendations, thereby improving the quality of search results. Secondly, the substantial progress achieved by deep learning methods in the field of image recognition empowers models to comprehend complex information within images, including the appearance, color, and shape features of products. This advanced image understanding capability allows tourism e-commerce platforms to better showcase products, attracting user attention and increasing product click-through rates and conversion rates. Additionally, deep learning methods support the processing of multi-modal data, such as simultaneously considering image and text information. By synthesizing different types of data, models can comprehensively understand products, offering users more accurate search results and personalized recommendations. This contributes to enhancing user experience and strengthening user engagement with tourism e-commerce platforms. On the other hand, deep learning methods exhibit strong generalization capabilities, adapting to various types and styles of product images. This flexibility enables tourism e-commerce platforms to effectively respond to the diversity of products and evolving market trends, maintaining competitiveness in the fiercely competitive tourism e-commerce landscape.

In summary, deep learning methods bring higher precision, comprehensive image understanding, and robust generalization capabilities to visual search in tourism e-commerce. This plays a crucial role in improving user experience, increasing sales effectiveness, and sustaining competitiveness in the market. The application of this advanced technology will drive continuous innovation in the tourism e-commerce domain, providing users and businesses with more intelligent and convenient interactive experiences. The current common models of deep learning for visual search are:

1. Convolutional Neural Network (CNN) (Nayeem et al., 2022): Translation: CNN is a deep learning model specifically designed for processing image data. It automatically learns features in images through components such as convolutional layers, pooling layers, and fully connected layers, thereby achieving abstract representations of image content. In visual search, CNN is widely applied for image feature extraction. Through pretrained CNN models, images can be

transformed into feature vectors with rich semantic information, thus enhancing the accuracy of image search.

2. Recurrent Neural Network (RNN) (Gaafar et al., 2022): Translation: RNN is a deep learning model suitable for sequential data, equipped with memory units that retain previous information and consider contextual relationships when processing sequences. In visual search, RNN can be employed for handling text information related to images, such as product descriptions. By considering the correlation between images and text, RNN contributes to improving the relevance of search results.
3. Generative Adversarial Network (GAN) (Kang et al., year needed): Translation: GAN consists of a generator and a discriminator, employing adversarial training to enable the generator to learn to produce realistic images while the discriminator learns to distinguish between real and generated images. In visual search, GAN can be used to generate additional training data, especially for scarce or rare products. By generating diverse images, the model's generalization capabilities for products are enhanced.
4. Adversarial Autoencoder (AAE) (Abubakr et al., 2022): Translation: AAE combines the principles of autoencoders and GANs. It uses an autoencoder to learn a low-dimensional representation of data and employs adversarial training to enhance the realism of generated images. In visual search, AAE can be applied for image reconstruction and generation, contributing to improving image quality and diversity, thereby enhancing result diversity and personalization.
5. Contrastive Language-Image Pretraining (CLIP) Model (Radford et al., 2021): Translation: CLIP, through joint training of image and text encoders, aligns images and text in an embedding space, enabling the model to understand the semantic relationship between images and text. In visual search within tourism e-commerce, the CLIP model can directly handle multi-modal data, improving the depth of understanding of products and the accuracy of searches. The advantage of the CLIP model lies in its unified representation of different modal data, exhibiting excellent performance in recommendation and search tasks.

To explore the application value of visual search in tourism e-commerce, this study proposes a multi-modal online product retrieval engine model, CLIP-ItP, based on the extension of the CLIP model. The model applies the CLIP model to the task of product retrieval on a tourism e-commerce website, constructing an experimental visual search engine. The model is divided into three stages. In the first stage, this visual search engine trains an image feature extractor and a linear model for labeling images. In the second stage, CLIP-ItP further jointly trains multiple text and image encoders, enabling the retrieval of various information about products on the tourism e-commerce website, including correct pairings of product image labels, product categories, product names, and product attributes. In the third stage, this visual search engine utilizes the user-uploaded image and jointly selected product attributes, employing CLIP-ItP to provide recommended Top-k products and their pictures.

The proposed multi-modal online product retrieval engine model, CLIP-ItP, has made three innovative contributions to visual search and tourism e-commerce:

1. Integration and Expansion of Multi-Modal Data: The CLIP-ItP model extends the CLIP model to integrate multi-modal data, co-training image and text information, thereby achieving a more comprehensive understanding of products on tourism e-commerce websites. This multi-modal integration encompasses not only product image tags but also includes various information such as product categories, names, and attributes. By co-training image and text encoders, the model enhances abstract representations of product information on multiple levels, strengthening its understanding of semantic relationships within products.
2. Recommendation System Based on User-Uploaded Images and Product Attributes: This study has implemented a recommendation system based on user-uploaded images and jointly selected product attributes. By combining user-uploaded images with selected product attributes, the

model utilizes CLIP-ItP to provide personalized Top-k product recommendations. This innovative recommendation system integrates users' visual inputs and selection behaviors, offering more accurate and personalized product recommendations, thereby enhancing the user interaction experience on tourism e-commerce platforms.

3. Comprehensive Experiments and Performance Analysis: This research conducted extensive experiments and performance analyses of the CLIP-ItP model on multiple public datasets and two private datasets. The results demonstrate the accuracy, robustness, and generalization capabilities of the model for product recommendation tasks. This comprehensive experimental evaluation contributes to validating the model's effectiveness and suitability for practical applications in tourism e-commerce.

In the rest of this paper, we will introduce the recently related work in section 2, including visual search technology in tourism e-commerce, visual retrieval based on multi-modal data, and zero-shot learning using the CLIP framework. Section 3 presents the proposed methods: overview, data acquisition and feature learning pipeline, and zero-shot learning using the CLIP framework. Section 4 introduces the experimental part, including practical details, comparative experiments, and an ablation study. Section 5 includes a conclusion and an outlook.

RELATED WORK

Visual Search Technology in Tourism E-Commerce

In tourism e-commerce, visual search technology has emerged as a transformative and impactful tool, fundamentally altering the way consumers explore and discover products online. This technology utilizes advanced computer vision and machine learning algorithms to analyze and comprehend the visual content of images. By extracting key features, patterns, and characteristics from product images, visual search systems can accurately identify and match items, providing users with relevant and visually similar results. This capability is particularly beneficial in scenarios where users find it challenging to articulate their search intent through text or encounter language barriers.

However, the challenges associated with this task remain substantial. Firstly, employing deep learning methods for visual search typically demands substantial computational resources, potentially posing challenges for some small to medium-sized tourism e-commerce platforms. Secondly, the processing of image data may involve user privacy concerns, necessitating effective measures on the part of tourism e-commerce platforms to safeguard the privacy and security of user data. Overall, the application of visual search technology in tourism e-commerce continues to evolve, offering users and merchants a more intelligent and convenient interactive experience, with the potential to become a key driving force in the future development of the tourism e-commerce industry.

Visual Retrieval Based on Multi-Modal Data

The visual retrieval method based on multi-modal data enriches the information representation of the visual retrieval system by simultaneously considering various data modalities such as images, text, and audio, thereby providing users with more comprehensive and accurate retrieval results. Common multi-modal visual retrieval methods achieve correlated learning between different data modalities by simultaneously training multiple modal feature extractors. This approach, through information sharing, captures the inter-modality correlations more effectively, enhancing the performance of the retrieval system (Ye & Zhao, 2023; Ye et al., 2023; Yuan et al., 2022).

In terms of application effectiveness, multi-modal visual retrieval based on diverse data modalities integrates information from images, text, audio, and more, offering a more comprehensive visual understanding that assists in meeting users' retrieval needs more accurately. Furthermore, the consideration of multi-modal data aids in modeling semantic relationships between different

data modalities, improving the retrieval system's understanding of complex semantic information and thereby enhancing the relevance of retrieval results. Lastly, multi-modal learning contributes to improving the system's generalization ability to new data, especially when dealing with limited samples or scarce modalities, enabling better adaptation to diverse data distributions.

However, the use of deep learning and joint learning methods may involve higher computational complexity, posing challenges in environments with limited computing resources. An additional challenge lies in the laborious and time-consuming process of annotating multi-modal data, particularly when dealing with intricate semantic relationships between modalities, which further increases the difficulty of annotation.

Zero-Shot Learning Using CLIP Framework

Developed from few-shot learning (Tian et al., 2024), zero-shot learning (ZSL) represents a crucial paradigm in the field of machine learning (Ning et al., 2024), emphasizing the ability to recognize and classify objects or concepts without direct exposure to training samples.

The integration of the CLIP (Contrastive Language-Image Pretraining) framework into the realm of zero-shot learning has garnered widespread attention, marking a significant advancement in this domain. The CLIP model is a multi-modal learning model developed by OpenAI. This model engages in contrastive learning between images and text, constructing a shared embedding space through joint pretraining on large-scale image and text data. This enables images and text to be represented in the same space. The design objective of CLIP is to empower the model to comprehend the semantic relationships between images and text, thereby achieving robust cross-modal understanding and generalization capabilities. The innovation of the CLIP model lies in its departure from traditional supervised learning methods. Instead, it adopts a contrastive learning approach, training by maximizing the similarity between positive sample pairs and minimizing the dissimilarity between positive and negative sample pairs. This methodology allows CLIP to excel in handling multi-modal data and zero-shot learning, providing a robust performance foundation for tasks such as visual search and image retrieval. The key principles of employing the CLIP framework for zero-shot learning revolve around utilizing pre-trained models that amalgamate image and text understanding (Feng & Chen, 2022).

Through joint training of image and text encoders, CLIP aligns images and their textual descriptions in a shared embedding space. This alignment empowers the model to comprehend the semantic relationships between images and their corresponding textual information, laying the foundation for achieving zero-shot learning. The zero-shot learning process involves presenting the model with classes or concepts not encountered during training. Leveraging the existing semantic understanding encoded in CLIP, the model can generalize and make predictions for these unseen classes, exploiting the model's ability to bridge the semantic gap between textual descriptions and visual content.

From existing research, the integration of CLIP provides a multi-modal understanding of images and text, enabling the model to generalize across different data modalities—a critical aspect of zero-shot learning. Furthermore, the semantic alignment achieved by CLIP aids in the transfer of knowledge from seen to unseen classes, enhancing the model's adaptability to new concepts. Lastly, employing CLIP for zero-shot learning reduces the reliance on extensive labeled data for unseen classes, particularly valuable in scenarios where data annotation is challenging or costly. However, using CLIP for zero-shot learning may pose challenges when dealing with highly specific or nuanced concepts, especially when the model has not been explicitly trained for such scenarios. Achieving fine-grained recognition in a zero-shot setting, particularly for subtle visual distinctions, remains an ongoing challenge. Additionally, the model's performance may be impacted by significant shifts in the data distribution between unseen and seen classes.

In conclusion, the integration of CLIP into zero-shot learning showcases promising avenues for advancing machine learning capabilities. While it brings forth several advantages, addressing challenges related to specific or nuanced concepts and maintaining performance in the face of data

distribution shifts remains crucial. As research progresses, the synergies between zero-shot learning and the CLIP framework are expected to play a pivotal role in pushing the boundaries of machine learning capabilities (Zhang et al., 2022).

METHOD

Overview of the Visual Search Engine and CLIP-ItP Model

This visual search engine comprises the following key components:

1. Product Image and Search Clue Input Interface: This component allows users to input images through methods such as image upload or providing image URLs. The images provided by users serve as the input data for the search.
2. Product Image Processing Module: The product image processing module is responsible for preprocessing the images provided by users. This may include operations such as resizing, normalization, noise reduction, etc., to ensure that the input images comply with the processing requirements of the search engine.
3. Feature Extractor: The feature extractor is a deep learning model. It is responsible for extracting crucial features from the input images, capturing advanced representations of the images for subsequent comparison and retrieval.
4. Index Database: Extracted image features and text information related to the images are stored in the index database for efficient similarity matching. This study adopts a vector-based database to support rapid vector similarity searches.
5. Search Algorithm: The search algorithm utilizes the user-provided image features and product search clue information to compare them with the stored product image features and attributes in the index database. In this study, a similarity calculation function is chosen to compute the similarity score between the input image and the platform's products, determining the products in the database most like the user-provided image.
6. Result Presentation: The retrieved similar images are presented to the user through a user interface or an API. This may include image arrangements, product information, purchase links, etc.

This architecture of the visual search engine encompasses the interfaces for product image and search clue input, the product image processing module, the feature extractor, the index database, the search algorithm, and the result presentation. These components collaboratively enable users to perform intuitive and intelligent searches and retrievals based on product images. The efficient search model, CLIP-ItP, proposed in this study, is pivotal for the successful operation of this engine. Figure 1 shows the architecture of this visual search engine.

In the search algorithm module of this visual search engine, a CLIP-ItP model is designed in this study. The principle of this model is shown in Figure 2.

As shown in Figure 2, in the first phase, the CLIP-ItP model trains an image feature extractor and a linear model for labeling images. In the second phase, CLIP-ItP further co-trains multiple text and image encoders to be able to retrieve various product information on tourism e-commerce websites, including the correct pairing of product image labels, product categories, product names, and product attributes. In the third phase, the CLIP-ItP model recommends Top-k products and their images using user-uploaded images and jointly selected product attributes.

Definition of the Problem

The input dataset for the algorithm is in the form of category-product information pairs

Figure 1. The Architecture of This Visual Search Engine

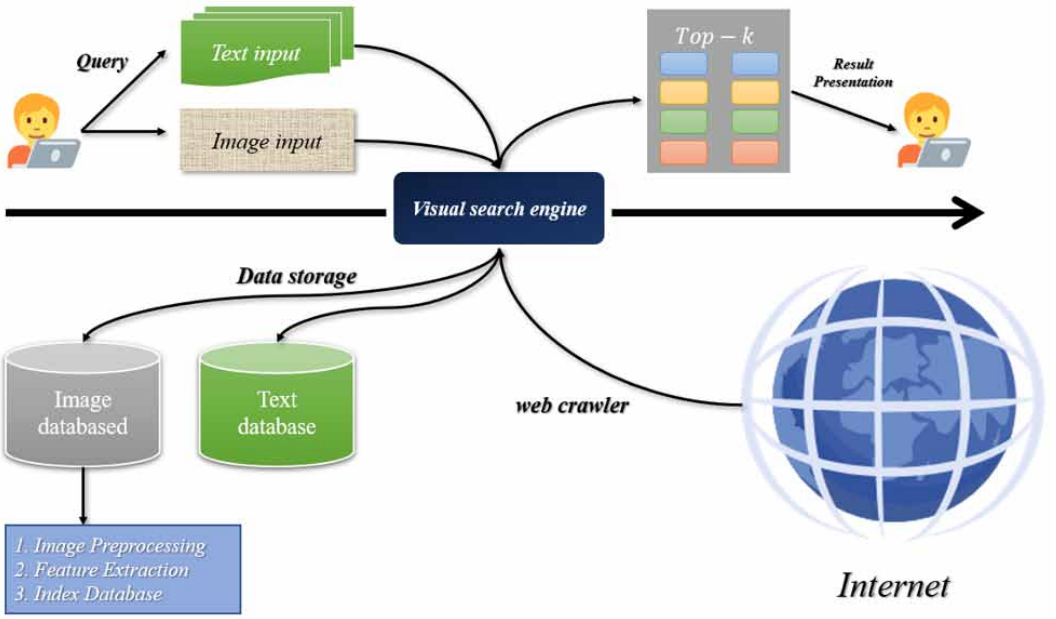
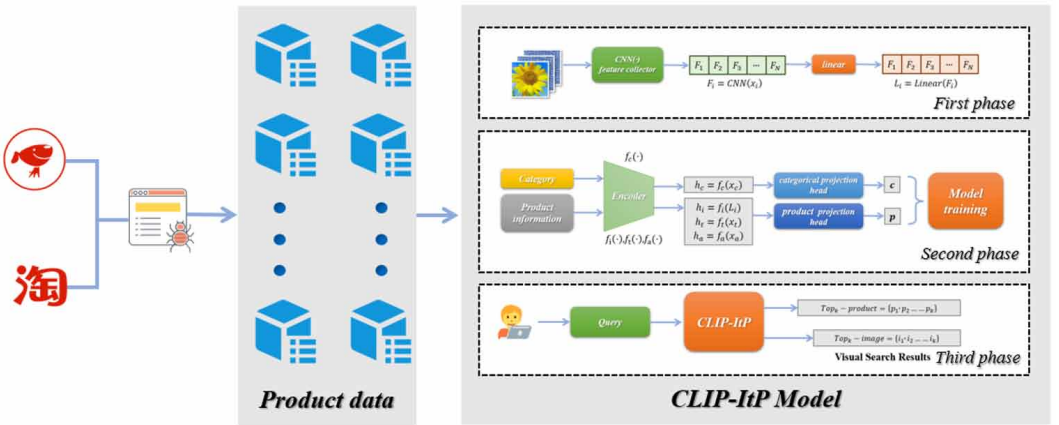


Figure 2. The Principle of the CLIP-ItP Model



$$D = \left\{ \left(x_c^{(1)}, x_p^{(1)} \right), \left(x_c^{(2)}, x_p^{(2)} \right) \ldots\ldots \left(x_c^{(N)}, x_p^{(N)} \right) \right\} \tag{1}$$

where x_c represents product categories and x_p represents product information belonging to category x_c. The product category x_c is taken from the category tree T and is denoted by the name of this subcategory.

The product information x_p consists of product title x_t, product image x_i, and product attributes x_a, i.e., $x_p = \left\{ x_i, x_t, x_a \right\}$. In the tourism e-commerce platform retrieval task, we use the name of the

target product category x_c as the query, and the model returns a sorted list of the $top-k$ images and products belonging to this product category x_c :

$$Top_k - product = \left\{\left(x_c^1, x_p^1\right), \left(x_c^2, x_p^2\right) \ldots\ldots \left(x_c^k, x_p^k\right)\right\} \tag{2}$$

and

$$Top_k - image = \left\{x_i^1, x_i^2 \ldots\ldots x_i^k\right\} \tag{3}$$

Data Acquisition and Feature Learning Pipeline

In this study, a distributed crawler was developed using Python Scrapy to crawl a large amount of product information from two websites, Jingdong (JD.com) and Taobao (Taobao.com), successively.

For the collected data, the two-step preprocessing pipeline is described below. The pipeline is shown in Figure 3.

Step 1: One is the resizing operation, i.e., the end side of the image is matched with the CLIP input size.
Step 2: Two is the center cropping operation, the result of which is input size × input size output = a square result image.

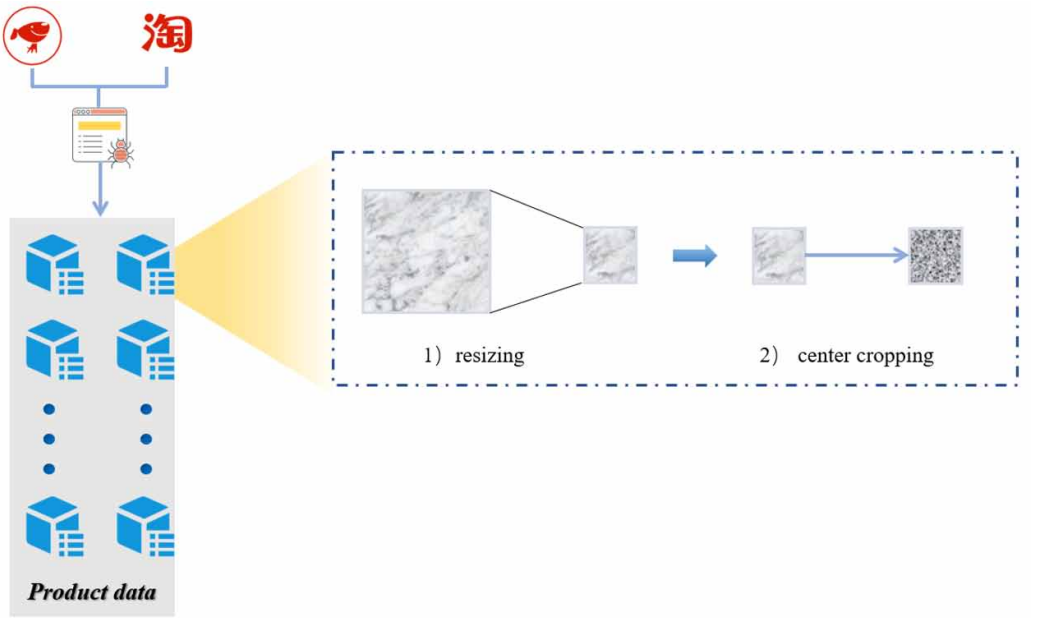
CLIP-ItP Framework for Product Search

The steps of the CLIP-ItP model are described below.

First phase:

Step 1: Images labeling. For each image in the dataset, an image feature collector, $CNN\left(\cdot\right)$, is run to collect features F_i from each image x_i in the dataset.

Figure 3. The Data Acquisition and Feature Learning Pipeline



$$F_i = CNN\left(x_i\right) \tag{4}$$

Next, using a linear classifier $Linear\left(\cdot\right)$, input a feature vector for each image F_i and use the output of the linear classifier L_i to label each image feature.

$$L_i = Linear\left(F_i\right) \tag{5}$$

Second phase:

Step 2: Category encoding. A text encoder f_c is used to transform the input product categorization information x_c into its representation h_c.

$$h_c = f_c\left(x_c\right) \tag{6}$$

where f_c is realized by a pre-trained MPNet model. Subsequently, we use a categorical projection head g_c to project h_c into a d -dimensional multi-modal space.

$$\boldsymbol{c} = g_c\left(h_c\right) \text{ where } \boldsymbol{c} \in \mathbb{R}^d \tag{7}$$

Step 3: Product encoding. Using a similar approach, the product information $x_p = \left\{L_i, x_t, x_a\right\}$ is output as h_i, h_t, and h_a, through multiple encoders.

$$h_i = f_i\left(L_i\right) \tag{8}$$

$$h_t = f_t\left(x_t\right) \tag{9}$$

$$h_a = f_a\left(x_a\right) = \frac{1}{M}\sum_{i=1}^{M} x_{ai} \tag{10}$$

Similarly, a product projection head g_p to project h_i, h_t, and h_a into a d -dimensional multi-modal space.

$$\boldsymbol{p} = g_c\left(concat\left(h_i, h_t, h_a\right)\right) \text{where } \boldsymbol{p} \in \mathbb{R}^d \tag{11}$$

Step 4: CLIP-ItP Model training. It is worth mentioning that the loss function used for model training is bidirectional contrastive loss:

$$\mathcal{L}oss = \frac{1}{\beta}\sum_{j=1}^{\beta}\left(\lambda \ell_j^{(p\to c)} + \left(1-\lambda\right)\ell_j^{(c\to p)}\right)$$

where β is the batch size, $\lambda \in \left[0,1\right]$ is a scalar weight, and

$$\ell_j^{(c\to p)} = -\log\frac{\exp\left(f_{sim}\left(\mathbf{c}_j,\mathbf{p}_j\right)/\tau\right)}{\sum_{k=1}^{\beta}\exp\left(f_{sim}\left(\mathbf{c}_j,\mathbf{p}_k\right)/\tau\right)}$$

and

$$\ell_j^{(p\to c)} = -\log\frac{\exp\left(f_{sim}\left(\mathbf{p}_j,\mathbf{c}_j\right)/\tau\right)}{\sum_{k=1}^{\beta}\exp\left(f_{sim}\left(\mathbf{p}_j,\mathbf{c}_k\right)/\tau\right)}$$

where $f_{sim}\left(\cdot\right)$ is cosine similarity, $\tau \in \mathbb{R}^+$ is a temperature parameter.

Third phase:

Step 5: Top-k product recommended. In the online application scenario of a tourism e-commerce website, the user inputs parameters to the model, including q_c and q_p, where $q_p = \left\{q_i, q_t, q_a\right\}$. After the calculation of CLIP-ItP model, Top-k recommended products are obtained:

$$Top_k - product = \left\{p_1, p_2 \ldots\ldots p_k\right\}$$
$$Top_k - image = \left\{i_1, i_2 \ldots\ldots i_k\right\}$$

The pipeline of CLIP-ItP model is shown in Figure 4.

EXPERIMENTS

Experimental Design

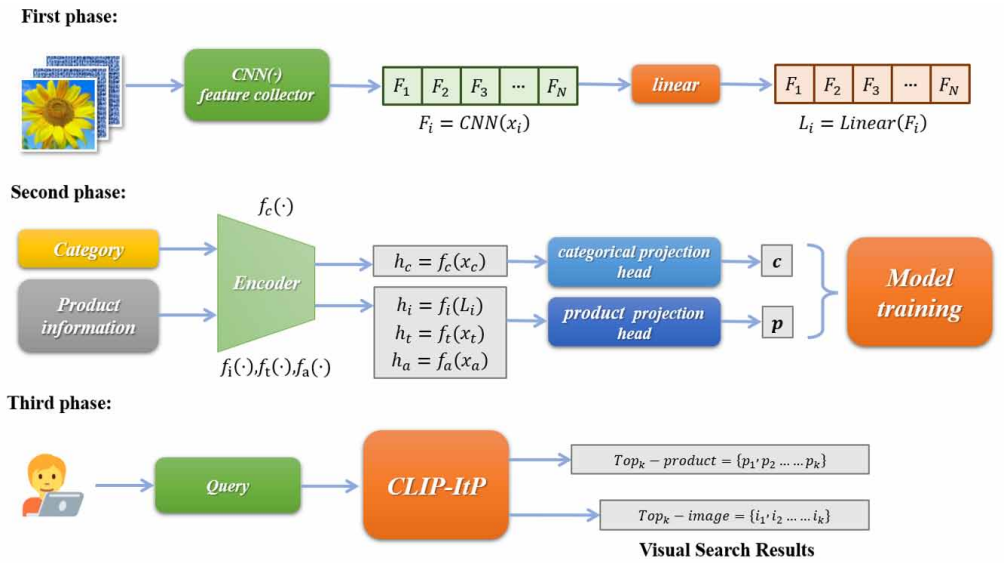
To comprehensively evaluate the performance of the CLIP-ItP model, a total of eight comparison experiments and one ablation experiment were designed in this study. Among them, experiments 1–4 were conducted on the public dataset to validate the accuracy difference between the CLIP-ItP model and the six baseline models under multiple input data types, respectively. Experiments 5–6 were conducted on the collected Taobao private dataset, and experiments 7–8 were conducted on the collected Jingdong private dataset, and the four experiments compare the accuracy of top-k recommendation results of the CLIP-ItP model with the six baseline models, respectively. Subsequent ablation experiments remove each module of the CLIP-ItP model separately, thus verifying the functional validity of each module.

The experimental setup is as follows:

Hardware Environment

1. Server: Dell PowerEdge R940 server is employed, equipped with 4 NVIDIA Tesla V100 GPUs.

Figure 4. The Pipeline of the CLIP-ItP Model



2. GPU: NVIDIA Tesla V100 Tensor Core GPUs are utilized, featuring 32GB HBM2 memory per card.
3. Storage: A 1TB NVMe SSD is configured for high-speed storage to ensure rapid data read and write operations.

Software Environment

1. Operating System: Ubuntu Server 20.04 LTS is employed as the operating system.
2. Deep Learning Frameworks: TensorFlow 2.5.0 and PyTorch 1.9.0 are chosen as the deep learning frameworks.
3. Database: Faiss is used as a vector-based database system to store image features and related textual information.
4. Web Application Interface: A web application interface is designed and implemented based on the Django framework. The system provides an interface for image uploading and interaction with users.

Additionally, to evaluate model performance, we use Precision@K($P@K$) where $K = \{1,5,10\}$, and $mAP@K$ where $K = \{5,10\}$, and R -precision.

CLIP Model Parameters Setting

1. The version of the CLIP model utilized in this study has a size of 24 million parameters.
2. The model was trained on a dataset comprising four hundred million images, and the training process was completed over a span of two weeks using 256 V100 GPUs.

Pre-Treatment Before Experimentation

1. Image Data Preprocessing:
   (1) Size Adjustment: All images are resized to the same dimensions to ensure consistency in model input.

(2) Normalization: Normalization is applied to scale image pixel values to the range of 0 to 1, aiding in improved training stability.
(3) Noise Removal: Optional noise removal operations are considered to reduce unnecessary visual disturbances.

2. Text Data Preprocessing:
   (1) Tokenization: Textual information such as product names and attributes are tokenized into sequences of words or subwords to facilitate model understanding and processing.
   (2) Embedding: Pre-trained word embedding models (e.g., Word2Vec, GloVe) are utilized to embed text into high-dimensional space, capturing the semantic information of words.
3. Data Augmentation:
   (1) Random Rotation and Flipping: Random rotation and flipping are applied to training images to increase data diversity.
   (2) Color Transformation: Random adjustments to image brightness, contrast, and hue are made to enhance the model's adaptability to different lighting and color conditions.
4. Dataset Splitting:
   (1) Training Set, Validation Set, and Test Set: The dataset is partitioned into training, validation, and test sets in a ratio of 7:2:1.

Datasets and Baseline Models

The data in this article comes from CIFAR-10, Birdsnap, Oxford-IIITPets, OxfordFlowers102, and FGVCAircraft. Below is a basic description of the five datasets.

CIFAR-10 (Canadian Institute for Advanced Research - 10) (Abouelnaga et al., 2016): CIFAR-10 is a widely used dataset in the field of computer vision and machine learning. It consists of 60,000 32x32 color images in 10 different classes, with 6,000 images per class. The dataset is divided into 50,000 training images and 10,000 testing images. Each class represents a specific object or animal, making it suitable for image classification tasks. Classes of CIFAR-10:1. Airplane 2. Automobile 3. Bird 4. Cat 5. Deer 6. Dog 7. Frog 8. Horse 9. Ship 10. Truck. The CIFAR-10 dataset is used in one of the experiments in this study, possibly to evaluate and compare the performance of the CLIP-ItP model in a general image classification task. As a widely used benchmark dataset for image classification, CIFAR-10 provides diverse images covering different categories of objects, aiding in assessing the model's generalization capability (Shen et al., 2019).

Birdsnap (Berg et al., 2014): Birdsnap is a dataset specifically designed for fine-grained bird species recognition. It contains images of 500 bird species, with a total of 50,000 images. The dataset aims to challenge algorithms to distinguish between visually similar bird species, making it suitable for fine-grained classification tasks. Classes of Birdsnap: The dataset consists of 500 bird species, each representing a class. Examples include various types of sparrows, finches, eagles, and more. The Birdsnap dataset is likely employed to validate the CLIP-ItP model's performance in bird image classification tasks. By using Birdsnap, researchers can assess the model's ability in fine-grained classification of birds, which is crucial for accuracy and granularity in specific products or goods on tourism e-commerce platforms.

Oxford-IIIT Pets (Parkhi et al., 2012): The Oxford-IIIT Pets dataset is created for fine-grained pet image classification. It contains images of 37 different breeds of dogs and cats, with a total of over 7,000 images. The dataset is well-suited for tasks that require distinguishing between different breeds of pets. Classes of Oxford-IIIT Pets: The dataset includes 12 cat breeds and 25 dog breeds, each forming a separate class. Examples of classes include Persian, Maine Coon, Beagle, and Golden Retriever. This dataset is probably used to test the CLIP-ItP model in the task of pet image classification. It offers a variety of images of pets, helping researchers better understand the model's capability in handling images related to pets.

Oxford Flowers 102 (Mete & Ensari, 2019): Oxford Flowers 102 is a dataset designed for fine-grained flower classification. It contains images of 102 different flower categories, with each category consisting of between 40 and 258 images. The dataset is suitable for tasks that involve recognizing and classifying diverse flower species. Classes of Oxford Flowers 102: There are 102 flower categories, with examples such as roses, tulips, daisies, sunflowers, and orchids, among others. The Oxford Flowers 102 dataset may be used to evaluate the CLIP-ItP model's performance in flower image classification tasks. This dataset provides images of flowers, allowing the testing of the model's ability to recognize different flower species, which is important for flower product searches on tourism e-commerce platforms.

FGVCAircraft (Maji et al., 2013): FGVCAircraft is a dataset for aircraft classification, specifically focusing on fine-grained recognition of aircraft types. It consists of images from 100 different aircraft models, with a total of approximately 10,000 images. The dataset challenges models to distinguish between visually similar aircraft classes. Classes of FGVCAircraft: The dataset includes 100 aircraft models, representing classes such as Boeing 747, Airbus A320, Cessna 172, and more. The FGVC Aircraft dataset is likely utilized to examine the CLIP-ItP model in aircraft image classification tasks. As a fine-grained classification dataset, it assists in evaluating the model's accuracy when dealing with similar but different objects belonging to distinct categories.

The six baseline models selected for this study are listed below:

Model 1: Arumugam and Subramani (2022) proposed to utilize direct as well as indirect relevance score computation for the images using text-based and context-based information to improve the performance of the search engine concerning the web image search results.

Model 2: Eswaran and Varshini (2022) proposed an image search model specific to the domain of garments in the fashion industry combined with CNN, ResNet50, and the Annoy Indexing algorithm developed by Spotify.

Model 3: Mustafic et al. (2019) proposed a method of training a distance function between two images using the deep neural network VGG19 and transfer learning.

Model 4: Nainani et al. (2023) presented a solution that utilizes zero-shot learning to create image queries with only user-provided text descriptions. This model uses OWL to check for the presence of bounding boxes and sorts images based on cosine similarity scores.

Model 5: Varish (2022) proposed a simple feature fusion scheme for content-based image retrieval (CBIR) using color, texture, and shape feature moments. In this algorithm, the fused single feature descriptors are computed by proficient fusion of color, texture, and shape feature moments.

Model 6: Yu and Liu (2022) proposed an image content retrieval method for a K-means clustering algorithm (KCA). The K-means clustering algorithm in this study is used to classify the color, pattern, shape, and content of images.

In summary, Models 1 and 4 used the image retrieval methods based on relevant textual cues, while Models 2, 3, and 5 used the computer vision-based approaches to improve the accuracy of retrieval results by deeply analyzing the features of the images. Model 6 employs a machine learning approach for image retrieval.

Comparison Study

Comparative Experiments on Public Datasets

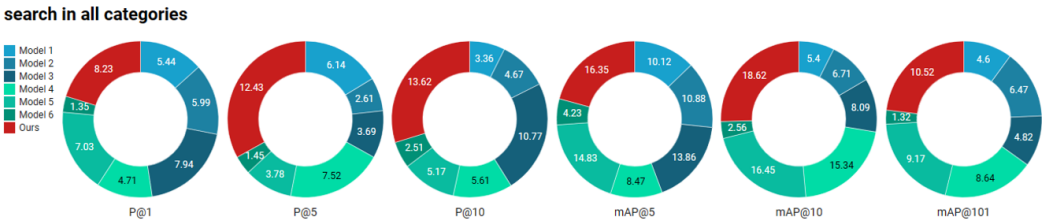
Experiment 1. *Experiment 1* uses an image-to-image retrieval method to compare the accuracy of retrieval results of the proposed model and the baseline model on six datasets, respectively. The results of the experiment are shown in Table 1.

The results of the experiment are presented graphically in Figure 5.

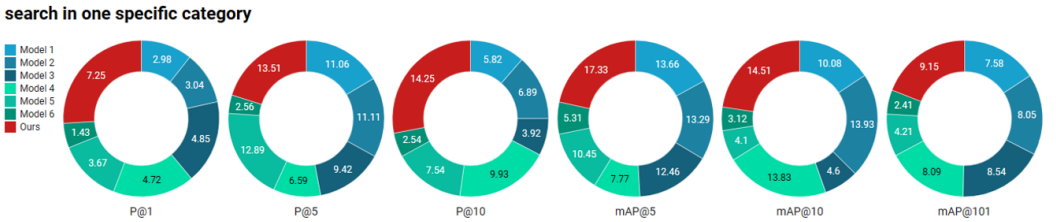
Table 1. Results of Experiment 1

Model	P@1	P@5	P@10	mAP@5	mAP@10	mAP@10
	Search in All Categories					
Model 1	5.44	6.14	3.36	10.12	5.4	4.6
Model 2	5.99	2.61	4.67	10.88	6.71	6.47
Model 3	7.94	3.69	10.77	13.86	8.09	4.82
Model 4	4.71	7.52	5.61	8.47	15.34	8.64
Model 5	7.03	3.78	5.17	14.83	16.45	9.17
Model 6	1.35	1.45	2.51	4.23	2.56	1.32
Ours	8.23	12.43	13.62	16.35	18.62	10.52
Model	**P@1**	**P@5**	**P@10**	**mAP@5**	**mAP@10**	**mAP@10**
	Search in One Specific Category					
Model 1	2.98	11.06	5.82	13.66	10.08	7.58
Model 2	3.04	11.11	6.89	13.29	13.93	8.05
Model 3	4.85	9.42	3.92	12.46	4.6	8.54
Model 4	4.72	6.59	9.93	7.77	13.83	8.09
Model 5	3.67	12.89	7.54	10.45	4.1	4.21
Model 6	1.43	2.56	2.54	5.31	3.12	2.41
Ours	7.25	13.51	14.25	17.33	14.51	9.15

Figure 5. Results of Experiment 1



(a) results of the search experiment in all categories



(b) results of the search experiment in one specific category

Analyzing the results from *experiment 1*, the CLIP-ItP model employs a similarity scoring function to determine the similarity score between the input image and platform products, identifying the product in the database most similar to the user-provided image. This calculation function is meticulously designed, leveraging the characteristics of the CLIP framework, making the model more

efficient in similarity matching. Other models may adopt different strategies for similarity calculation, potentially lacking the efficiency demonstrated by the CLIP-ItP model.

Experiment 2. *Experiment 2* uses only product categorization as the cue information for image retrieval and compares the accuracy of retrieval results between the proposed model and the baseline model on six datasets, respectively. The results of the experiment are shown in Table 2.

The results of the experiment are presented graphically in Figure 6.

Analyzing the results from *experiment 2*, the CLIP-ItP model proposed in this paper adopts the concept of zero-shot learning. By learning joint representations of images and text during the pretraining phase, the model is capable of predicting new categories without direct exposure to training samples. This enhances the search engine's versatility, enabling it to adapt to various types of images uploaded by users without requiring additional labeled data for accommodating new categories, a requirement that other models might necessitate.

Experiment 3. *Experiment 3* employs product classification and product name as the cue information for image retrieval and compares the accuracy of retrieval results between the proposed model and the baseline model on six datasets, respectively. The results of the experiment are shown in Table 3.

The results of the experiment are presented graphically in Figure 7.

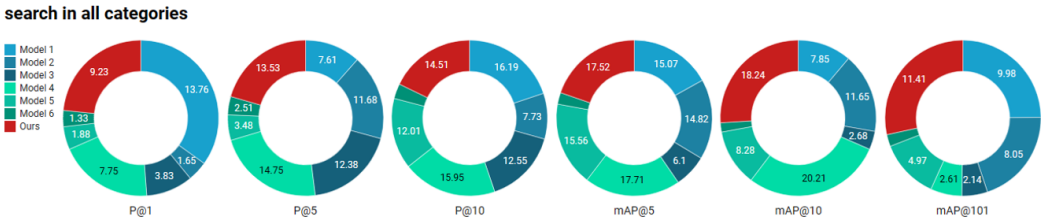
Analyzing the results from *experiment 3*, by employing deep learning and zero-shot learning techniques, CLIP-ItP may possess enhanced data generalization capabilities, allowing it to better adapt to various types and styles of product images. This enables the search engine to maintain high performance when confronted with evolving market trends and product diversity, while other models may exhibit more limited performance in terms of data generalization.

Experiment 4. *Experiment 4* simultaneously used product classification, product name, and product attributes as cue information for image retrieval to compare the accuracy of retrieval results

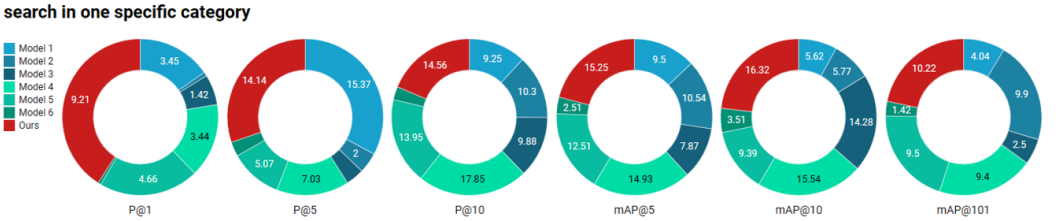
Table 2. Results of Experiment 2

Model	P@1	P@5	P@10	mAP@5	mAP@10	mAP@10
	Search in All Categories					
Model 1	13.76	7.61	16.19	15.07	7.85	9.98
Model 2	1.65	11.68	7.73	14.82	11.65	8.05
Model 3	3.83	12.38	12.55	6.10	2.68	2.14
Model 4	7.75	14.75	15.95	17.71	20.21	2.61
Model 5	1.88	3.48	12.01	15.56	8.28	4.97
Model 6	1.33	2.51	2.54	2.14	1.35	0.99
Ours	9.23	13.53	14.51	17.52	18.24	11.41
Model	**P@1**	**P@5**	**P@10**	**mAP@5**	**mAP@10**	**mAP@10**
	Search in One Specific Category					
Model 1	3.45	15.37	9.25	9.50	5.62	4.04
Model 2	0.18	2.00	10.30	10.54	5.77	9.90
Model 3	1.42	1.77	9.88	7.87	14.28	2.50
Model 4	3.44	7.03	17.85	14.93	15.54	9.40
Model 5	4.66	5.07	13.95	12.51	9.39	9.5
Model 6	0.14	1.41	2.12	2.51	3.51	1.42
Ours	9.21	14.14	14.56	15.25	16.32	10.22

Figure 6. Results of Experiment 2



(a) results of the search experiment in all categories

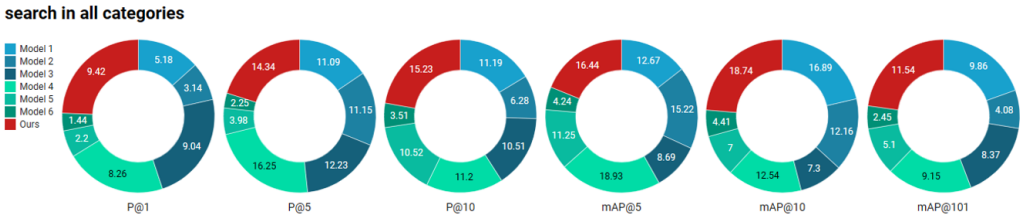


(b) results of the search experiment in one specific category

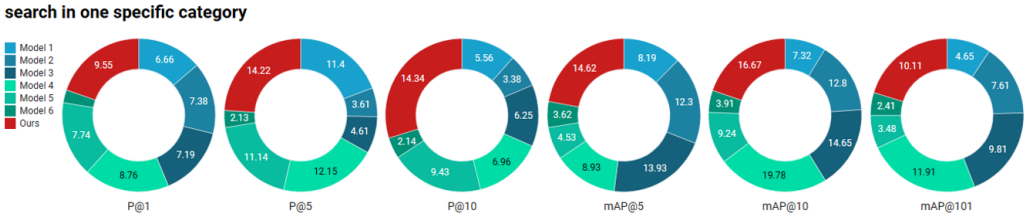
Table 3. Results of Experiment 3

Model	P@1	P@5	P@10	mAP@5	mAP@10	mAP@10
	Search in All Categories					
Model 1	5.18	11.09	11.19	12.67	16.89	9.86
Model 2	3.14	11.15	6.28	15.22	12.16	4.08
Model 3	9.04	12.23	10.51	8.69	7.30	8.37
Model 4	8.26	16.25	11.20	18.93	12.54	9.15
Model 5	2.2	3.98	10.52	11.25	7	5.1
Model 6	1.44	2.25	3.51	4.24	4.41	2.45
Ours	9.42	14.34	15.23	16.44	18.74	11.54
Model	**P@1**	**P@5**	**P@10**	**mAP@5**	**mAP@10**	**mAP@10**
	Search in One Specific Category					
Model 1	6.66	11.40	5.56	8.19	7.32	4.65
Model 2	7.38	3.61	3.38	12.30	12.80	7.61
Model 3	7.19	4.61	6.25	13.93	14.65	9.81
Model 4	8.76	12.15	6.96	8.93	19.78	11.91
Model 5	7.74	11.14	9.43	4.53	9.24	3.48
Model 6	1.32	2.13	2.14	3.62	3.91	2.41
Ours	9.55	14.22	14.34	14.62	16.67	10.11

Figure 7. Results of Experiment 3



(a) results of the search experiment in all categories



(b) results of the search experiment in one specific category

between the proposed model and the baseline model on six datasets, respectively. The results of the experiment are shown in Table 4.

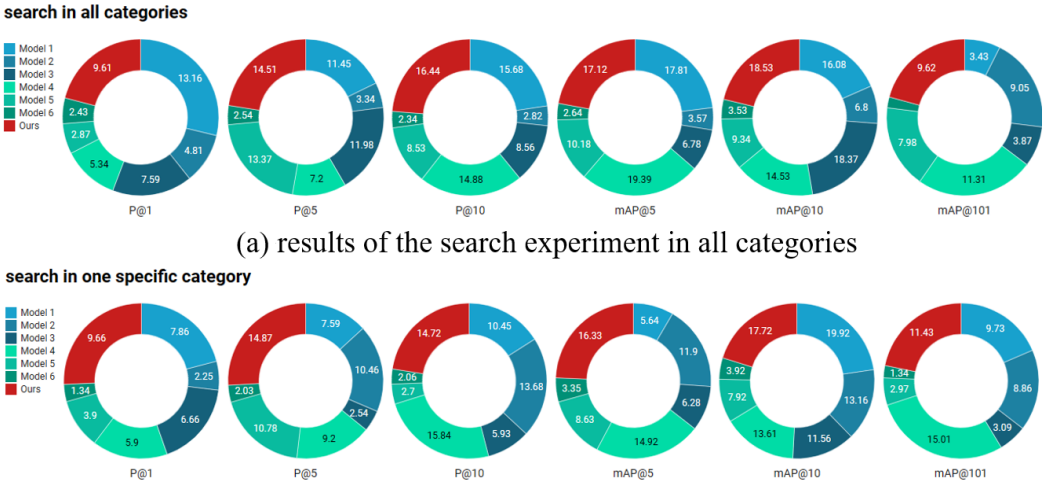
The results of the experiment are presented graphically in Figure 8.

Analyzing the results from *experiment 4*, the CLIP framework employs a semantic alignment approach, aligning images and text in the embedding space, enabling the model to comprehend the

Table 4. Results of Experiment 4

Model	P@1	P@5	P@10	mAP@5	mAP@10	mAP@10
	Search in All Categories					
Model 1	13.16	11.45	15.68	17.81	16.08	3.43
Model 2	4.81	3.34	2.82	3.57	6.80	9.05
Model 3	7.59	11.98	8.56	6.78	18.37	3.87
Model 4	5.34	7.20	14.88	19.39	14.53	11.31
Model 5	2.87	13.37	8.53	10.18	9.34	7.98
Model 6	2.43	2.54	2.34	2.64	3.53	1.03
Ours	9.61	14.51	16.44	17.12	18.53	9.62
Model	**P@1**	**P@5**	**P@10**	**mAP@5**	**mAP@10**	**mAP@10**
	Search in One Specific Category					
Model 1	7.86	7.59	10.45	5.64	19.92	9.73
Model 2	2.25	10.46	13.68	11.90	13.16	8.86
Model 3	6.66	2.54	5.93	6.28	11.56	3.09
Model 4	5.90	9.20	15.84	14.92	13.61	15.01
Model 5	3.9	10.78	2.7	8.63	7.92	2.97
Model 6	1.34	2.03	2.06	3.35	3.92	1.34
Ours	9.66	14.87	14.72	16.33	17.72	11.43

Figure 8. Results of Experiment 4



(a) results of the search experiment in all categories

(b) results of the search experiment in one specific category

semantic relationships between images and text. In the context of search engine tasks, this semantic alignment capability allows CLIP-ItP to better understand the images provided by users and their associated textual clues, thereby enhancing the relevance of search results.

In order to evaluate the runtime overhead of the models, the runtime overhead of several models of this experiment is compared, and the comparison experiment uses a subset of 1,000 experimental images, and the results obtained are: model 1 consumes 10.12s, model 2 consumes 11.53s, model 3 consumes 13.14s, model 4 consumes 16.25s, model 5 consumes 9.53s, Model 6 consumes 21.34s, and the model in this paper consumes 12.82s.

Additionally, in a case where the model made an incorrect recommendation, a user uploaded an image featuring a bottle of shampoo. However, the CLIP-ItP model erroneously identified this shampoo bottle as a bottle of shower gel and provided product recommendations related to shower gel. Upon analysis, potential reasons for this error include:

(1) Similar Appearance: Shampoo and shower gel may have a similar appearance with comparable shapes and colors, making it challenging for the model to accurately distinguish between them visually.
(2) Keywords in Text Descriptions: If the user-uploaded image lacks sufficient visual cues and the model heavily relies on textual information, it may be influenced by the keywords provided by the user, leading to an incorrect classification.
(3) Training Data Limitations: If the distribution of images for shampoo and shower gel in the training data is uneven or if there are images sharing similar features, the model may experience confusion during the learning process for these products.

Comparative Experiments on Real Datasets

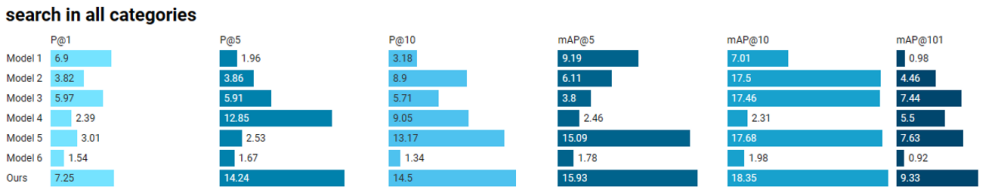
Experiment 5. *Experiment 5* compares the proposed model with six baseline models on the collected Taobao product dataset. In this, only the product categorization is used as the cue information for image retrieval. The results of the experiment are shown in Table 5.

The results of the experiment are presented graphically in Figure 9.

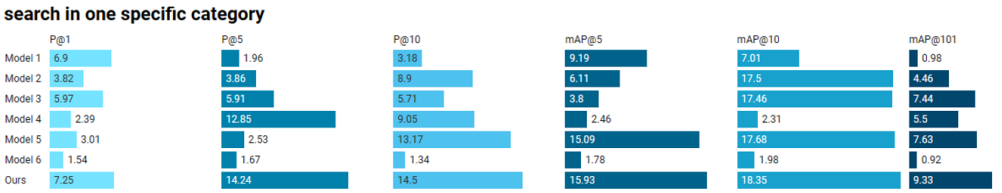
Table 5. Results of Experiment 5

Model	P@1	P@5	P@10	mAP@5	mAP@10	mAP@10
	Search in All Categories					
Model 1	6.90	1.96	3.18	9.19	7.01	0.98
Model 2	3.82	3.86	8.90	6.11	17.50	4.46
Model 3	5.97	5.91	5.71	3.80	17.46	7.44
Model 4	2.39	12.85	9.05	2.46	2.31	5.50
Model 5	3.01	2.53	13.17	15.09	17.68	7.63
Model 6	1.54	1.67	1.34	1.78	1.98	0.92
Ours	7.25	14.24	14.5	15.93	18.35	9.33
Model	**P@1**	**P@5**	**P@10**	**mAP@5**	**mAP@10**	**mAP@10**
	Search in One Specific Category					
Model 1	5.18	4.73	9.30	2.32	2.90	1.12
Model 2	4.78	2.45	14.50	13.90	4.30	5.43
Model 3	1.74	3.89	5.52	5.52	2.99	6.53
Model 4	1.27	11.92	2.01	12.18	10.43	5.80
Model 5	5.55	9.07	6.08	9.98	9.59	6.83
Model 6	0.91	0.99	1.03	1.42	1.6	1.03
Ours	6.78	16.45	16.26	18.36	14.73	9.13

Figure 9. Results of Experiment 5



(a) results of the search experiment in all categories



(b) results of the search experiment in one specific category

Analyzing the results from *experiment 5*, due to the design of zero-shot learning and multi-modal fusion in CLIP-ItP, it possesses greater generality and adaptability, enabling it to accommodate various types of images uploaded by users. This flexibility makes the search engine not only capable of handling common products but also adept at dealing with novel, rare, or items from different domains, thereby enhancing the practicality of the search engine.

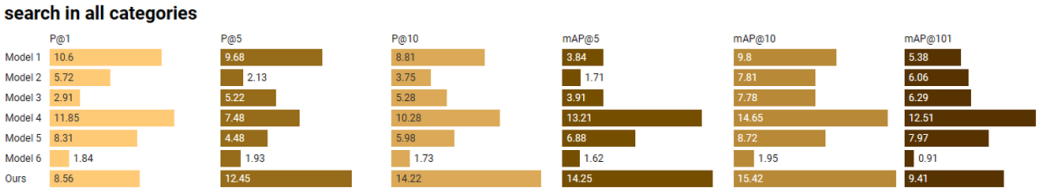
Experiment 6. *Experiment 6* compares the proposed model with six baseline models on the collected Taobao product dataset. In this case, product classification, product name, and product attributes are also used as cue information for image retrieval. The results of the experiment are shown in Table 6.

The results of the experiment are presented graphically in Figure 10.

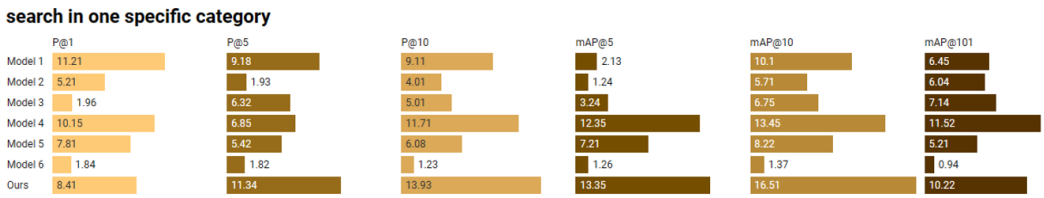
Table 6. Results of Experiment 6

Model	P@1	P@5	P@10	mAP@5	mAP@10	mAP@10
	Search in All Categories					
Model 1	10.60	9.68	8.81	3.84	9.80	5.38
Model 2	5.72	2.13	3.75	1.71	7.81	6.06
Model 3	2.91	5.22	5.28	3.91	7.78	6.29
Model 4	11.85	7.48	10.28	13.21	14.65	12.51
Model 5	8.31	4.48	5.98	6.88	8.72	7.97
Model 6	1.84	1.93	1.73	1.62	1.95	0.91
Ours	8.56	12.45	14.22	14.25	15.42	9.41
Model	**P@1**	**P@5**	**P@10**	**mAP@5**	**mAP@10**	**mAP@10**
	Search in One Specific Category					
Model 1	11.21	9.18	9.11	2.13	10.1	6.45
Model 2	5.21	1.93	4.01	1.24	5.71	6.04
Model 3	1.96	6.32	5.01	3.24	6.75	7.14
Model 4	10.15	6.85	11.71	12.35	13.45	11.52
Model 5	7.81	5.42	6.08	7.21	8.22	5.21
Model 6	1.84	1.82	1.23	1.26	1.37	0.94
Ours	8.41	11.34	13.93	13.35	16.51	10.22

Figure 10. Results of Experiment 6



(a) results of the search experiment in all categories



(b) results of the search experiment in one specific category

Analyzing the results from *experiment 6*, the use of CLIP for zero-shot learning reduces the reliance on a large amount of labeled data for unseen categories. This makes CLIP-ItP particularly valuable in scenarios where data annotation is challenging or expensive. Other models may not perform as well as CLIP-ItP when dealing with situations that lack labeled data.

Experiment 7. *Experiment 7* compares the proposed model with six baseline models on the collected Jingdong product dataset. In this, only the product categorization is used as the cue information for image retrieval. The results of the experiment are shown in Table 7.

Analyzing the results from *experiment 7*, CLIP-ItP adopts a pretraining approach, learning the joint representation of images and text during the pretraining phase. This allows the model to predict new categories without direct exposure to training samples. The pretraining paradigm gives CLIP-ItP a distinct advantage, especially when adapting to novel, unseen product categories. Other models may require more labeled data to adapt to new categories, which can be a significant challenge in practical scenarios.

Experiment 8. *Experiment 8* compares the proposed model with six baseline models on the collected Jingdong product dataset. In this case, product categorization, product name, and product attributes are also used as cue information for image retrieval. The results of the experiment are shown in Table 8.

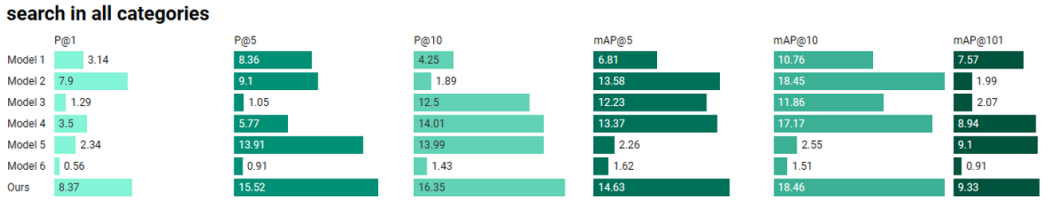
The results of the experiment are presented graphically in Figure 12.

Analyzing the results from *experiment 8*, on tourism e-commerce platforms, the regular introduction of new products is a common occurrence. CLIP-ItP demonstrates greater flexibility in adapting to these changes, while other models may require additional annotated data to accommodate new categories. This poses a challenging issue in practical scenarios.

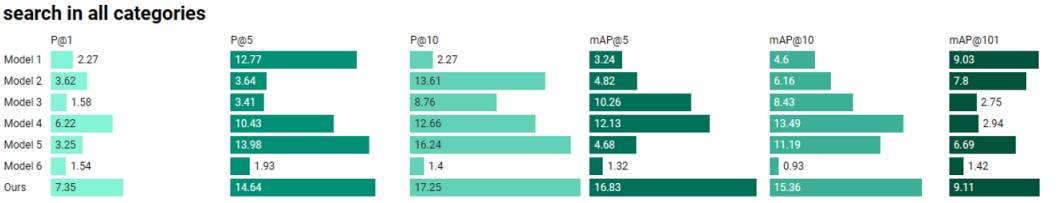
Table 7. Results of Experiment 7

Model	P@1	P@5	P@10	mAP@5	mAP@10	mAP@10
	Search in All Categories					
Model 1	3.14	8.36	4.25	6.81	10.76	7.57
Model 2	7.90	9.10	1.89	13.58	18.45	1.99
Model 3	1.29	1.05	12.50	12.23	11.86	2.07
Model 4	3.50	5.77	14.01	13.37	17.17	8.94
Model 5	2.34	13.91	13.99	2.26	2.55	9.1
Model 6	0.56	0.91	1.43	1.62	1.51	0.91
Ours	8.37	15.52	16.35	14.63	18.46	9.33
Model	**P@1**	**P@5**	**P@10**	**mAP@5**	**mAP@10**	**mAP@10**
	Search in One Specific Category					
Model 1	2.27	12.77	2.27	3.24	4.60	9.03
Model 2	3.62	3.64	13.61	4.82	6.16	7.80
Model 3	1.58	3.41	8.76	10.26	8.43	2.75
Model 4	6.22	10.43	12.66	12.13	13.49	2.94
Model 5	3.25	13.98	16.24	4.68	11.19	6.69
Model 6	1.54	1.93	1.4	1.32	0.93	1.42
Ours	7.35	14.64	17.25	16.83	15.36	9.11

Figure 11. Results of Experiment 7



(a) results of the search experiment in all categories

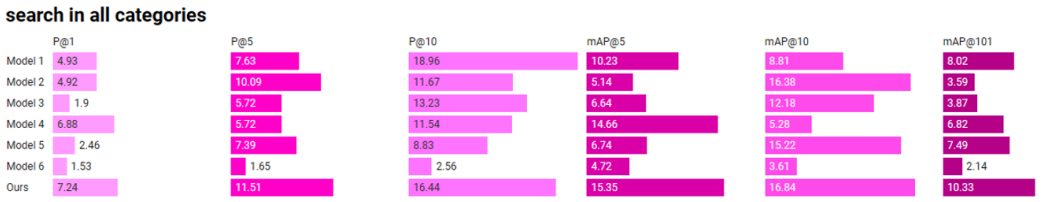


(b) results of the search experiment in one specific category

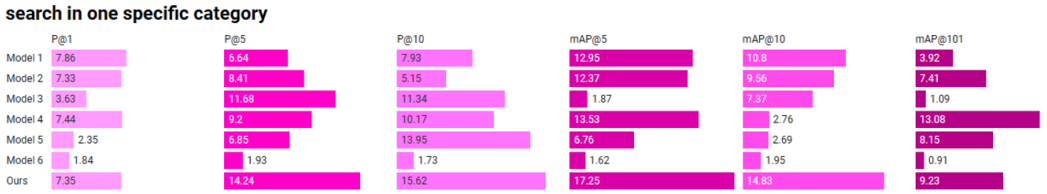
Table 8. Results of Experiment 8

Model	P@1	P@5	P@10	mAP@5	mAP@10	mAP@10
	Search in All Categories					
Model 1	4.93	7.63	18.96	10.23	8.81	8.02
Model 2	4.92	10.09	11.67	5.14	16.38	3.59
Model 3	1.90	5.72	13.23	6.64	12.18	3.87
Model 4	6.88	5.72	11.54	14.66	5.28	6.82
Model 5	2.46	7.39	8.83	6.74	15.22	7.49
Model 6	1.53	1.65	2.56	4.72	3.61	2.14
Ours	7.24	11.51	16.44	15.35	16.84	10.33
Model	**P@1**	**P@5**	**P@10**	**mAP@5**	**mAP@10**	**mAP@10**
	Search in One Specific Category					
Model 1	7.86	6.64	7.93	12.95	10.80	3.92
Model 2	7.33	8.41	5.15	12.37	9.56	7.41
Model 3	3.63	11.68	11.34	1.87	7.37	1.09
Model 4	7.44	9.20	10.17	13.53	2.76	13.08
Model 5	2.35	6.85	13.95	6.76	2.69	8.15
Model 6	1.84	1.93	1.73	1.62	1.95	0.91
Ours	7.35	14.24	15.62	17.25	14.83	9.23

Figure 12. Results of Experiment 8



(a) results of the search experiment in all categories



(b) results of the search experiment in one specific category

Ablation Study Results and Analysis

To evaluate the functionality of each module in the CLIP-ItP model, an ablation experiment was conducted to eliminate each module in the CLIP-ItP model separately and compare the differences in the results of the model runs. The experimental results are shown in Table 9.

The results of the ablation experiment are presented graphically in Figure 13.

Then we compare the results of the three model running results of CLIP without image labeling, CLIP with image labeling, and CLIP-ItP. Firstly, both CLIP with image labeling and CLIP without image labeling exhibit limitations in integrating information compared to CLIP-ItP, resulting in a less profound understanding of user-uploaded images. Secondly, the design of CLIP-ItP provides certain advantages in terms of model performance, including higher data generalization capability and stronger semantic alignment, contributing to its superior performance in the ablation experiment.

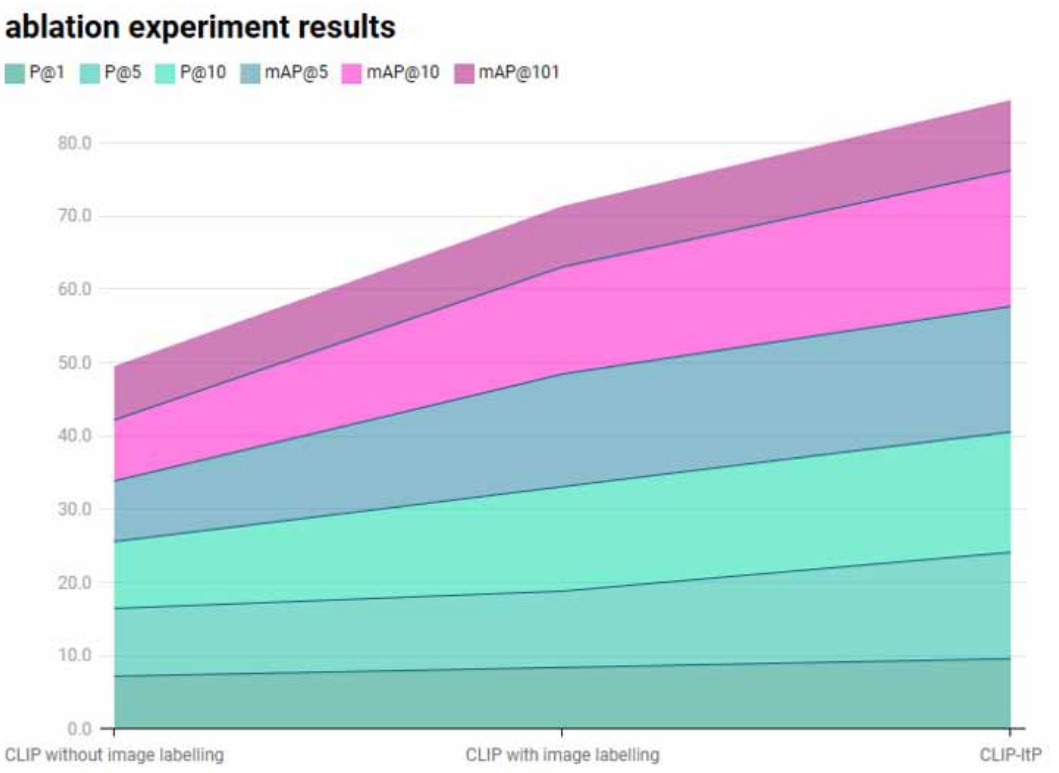
CONCLUSION

This study introduces an innovative visual search engine model, namely CLIP-ItP, which integrates multi-modal data by extending the CLIP framework, incorporating zero-shot learning to enhance adaptability to new products. Through comparative experiments, CLIP-ItP demonstrates outstanding performance on tourism e-commerce platforms, particularly excelling in handling new product listings, data generalization, and user-uploaded images compared to other baseline models. The research validates the accuracy, robustness, and generalization capabilities of CLIP-ItP in the domain of visual

Table 9. Results of Ablation Experiment

Models	P@1	P@5	P@10	mAP@5	mAP@10	mAP@10
CLIP without image labeling	7.23	9.25	9.11	8.25	8.35	7.35
CLIP with image labeling	8.42	10.42	14.24	15.35	14.62	8.34
CLIP-ItP	9.61	14.51	16.44	17.12	18.53	9.62

Figure 13. Results of Ablation Experiment



search tasks in tourism e-commerce. The model proves to be a powerful tool for enhancing search engine efficiency and user experience.

OUTLOOK

While CLIP-ItP exhibits outstanding performance in handling common and newly introduced products, it may face challenges when dealing with certain specialized domains or unique items. To enhance the model's performance in these specific domains, future research could consider incorporating domain-specific expert knowledge or domain-specific pretraining to improve the model's understanding of complex concepts and attributes within the domain. Further customization and domain adaptation studies could enable CLIP-ItP to achieve superior performance across a broader spectrum of product domains.

Another potential constraint is CLIP-ItP's adaptability to scenarios involving few samples or scarce modalities. In future work, performance in situations with limited samples and modalities can be enhanced by introducing more sophisticated data augmentation techniques, exploring meta-learning approaches, or employing technologies such as Generative Adversarial Networks (GANs). Additionally, considering data privacy concerns, research efforts could focus on refining the model's handling of privacy-sensitive data to provide efficient search services while safeguarding user privacy. These improvements will contribute to expanding the applicability and practicality of CLIP-ItP.

Additionally, a limitation of this study lies in its pronounced performance on known categories; however, it may not perform as well when dealing with unknown or extremely rare categories. To address this, future research could consider incorporating an active learning

framework, allowing the model to actively select samples for annotation, thereby enhancing performance on unknown categories.

Finally, this study may face challenges in handling some highly specific or subtle concepts, especially when the model has not been explicitly trained for them. To tackle this issue, future research could explore the introduction of more sophisticated fine-grained feature extractors or investigate the use of techniques such as Generative Adversarial Networks (GANs) to enhance the model's ability to recognize subtle visual differences.

The CLIP-ItP model proposed in this study has achieved significant contributions to visual search in tourism e-commerce. Through the extension and optimization of the CLIP framework, CLIP-ItP demonstrates innovation in areas such as multi-modal data integration and zero-shot learning, exhibiting commendable performance. Experimental results validate its notable advantages in handling new product listings, data generalization, and user-uploaded images. This research holds important reference value for enhancing search engine efficiency, adaptability, and driving technological advancements in tourism e-commerce platforms, providing more intelligent and flexible solutions for the future of tourism e-commerce.

COMPETING INTERESTS

All Authors declare that they have no conflict of interest.

FUNDINGS

This work was supported by The Ministry of Education's Industry University Cooperation Collaborative Education Project(Grant No. 230821141307061), Research and Creation Project of Zhejiang Provincial Department of Culture and Tourism (Grant No. 2023KYY039), and General Research Project of Zhejiang Provincial Department of Education (Grant No. Y202352193).

REFERENCES

Abouelnaga, Y., Ali, O. S., Rady, H., & Moustafa, M. (2016). Cifar-10: Knn-based ensemble of classifiers. *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*.

Abubakr, A. G., Jovančević, I., Mokhtari, N. I., Abdallah, H. B., & Orteu, J.-J. (2022). Learning deep domain-agnostic features from synthetic renders for industrial visual inspection. *Journal of Electronic Imaging*, *31*(5), 051604–051604. doi:10.1117/1.JEI.31.5.051604

Anand, G., Wang, S., & Ni, K. (2021). Large-scale visual search and similarity for e-commerce. *Applications of Machine Learning*.

Arumugam, S., & Subramani, S. B. (2022). Enhancing the web image search results through direct and indirect relevance model. *AIP Conference Proceedings*.

Berg, T., Liu, J., Woo Lee, S., Alexander, M. L., Jacobs, D. W., & Belhumeur, P. N. (2014). Birdsnap: Large-scale fine-grained visual categorization of birds. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Chen, C., Lu, M. Y., Williamson, D. F., Chen, T. Y., Schaumberg, A. J., & Mahmood, F. (2022a). Fast and scalable search of whole-slide images via self-supervised deep learning. *Nature Biomedical Engineering*, *6*(12), 1420–1434. doi:10.1038/s41551-022-00929-8 PMID:36217022

Chen, M., Liu, Q., Huang, S., & Dang, C. (2022b). Environmental cost control system of manufacturing enterprises using artificial intelligence based on value chain of circular economy. *Enterprise Information Systems*, *16*(8-9), 1268–1287. doi:10.1080/17517575.2020.1856422

Dagan, A., Guy, I., & Novgorodov, S. (2023). Shop by image: Characterizing visual search in e-commerce. *Information Retrieval*, *26*(1), 2. doi:10.1007/s10791-023-09418-1

Du, M., Ramisa, A., KC, A. K., Chanda, S., Wang, M., Rajesh, N., Li, S., Hu, Y., Zhou, T., & Lakshminarayana, N. (2022). Amazon Shop the Look: A visual search system for fashion and home. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Eswaran, A., & Varshini, E. (2022). Reverse image search engine for garment industry. *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*.

Feng, Z., & Chen, M. (2022). Platformance-based cross-border import retail e-commerce service quality evaluation using an artificial neural network analysis. *Journal of Global Information Management*, *30*(11), 1–17. Advance online publication. doi:10.4018/JGIM.306271

Foping, F., Tchagna, A., & Mih, T. (2022). A visual search system powered by locality sensitive hashing. *Journal of Computer Vision and Pattern Recognition*, *1*(1), 1–10.

Gaafar, A. S., Dahr, J. M., & Hamoud, A. K. (2022). Comparative analysis of performance of deep learning classification approach based on LSTM-RNN for textual and image datasets. *Informatica (Vilnius)*, *46*(5).

Guang, , J. Y., Xiang, C., & Xu, S. (2021). Stack overflow question post classification method based on deep learning. *Journal of Jilin University*, *59*(4).

He, G. (2021). Enterprise e-commerce marketing system based on big data methods of maintaining social relations in the process of e-commerce environmental commodity. *Journal of Organizational and End User Computing*, *33*(6), 1–16. doi:10.4018/JOEUC.20211101.oa16

Hou, J., Li, Q., Liu, Y., & Zhang, S. (2021). An enhanced cascading model for e-commerce consumer credit default prediction. *Journal of Organizational and End User Computing*, *33*(6), 1–18. doi:10.4018/JOEUC.20211101.oa13

KangW.AhnS.HongJ.-H. (n.d.). Investigating the potential of gan-based interactive image editing for novel image search. *Available at* SSRN 4062469.

Li, Y. (2022). Research and application of deep learning in image recognition. *2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA)*.

Mete, B. R., & Ensari, T. (2019). Flower classification with deep CNN and machine learning algorithms. *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*.

Monjur, M. E. I., Rifat, A. H., Islam, M. R., & Bhuiyan, M. R. (2023). The impact of artificial intelligence on international trade: Evidence from B2C giant e-commerce (Amazon, Alibaba, Shopify, eBay). *Open Journal of Business and Management*, *11*(5), 2389–2401. doi:10.4236/ojbm.2023.115132

Mustafic, F., Prazina, I., & Ljubovic, V. (2019). A new method for improving content-based image retrieval using deep learning. *2019 XXVII international conference on information, communication and automation technologies (ICAT)*.

Nayeem, T. A., Motaharuzzaman, S., Hoque, A. T., & Rahman, M. H. (2022). Computer vision based object detection and recognition system for image searching. *2022 12th International Conference on Electrical and Computer Engineering (ICECE)*.

Ning, X., Yu, Z., Li, L., Li, W., & Tiwari, P. (2024). DILF: Differentiable rendering-based multi-view image–language fusion for zero-shot 3D shape understanding. *Information Fusion*, *102*, 102033. doi:10.1016/j.inffus.2023.102033

Parkhi, O. M., Vedaldi, A., Zisserman, A., & Jawahar, C. (2012). Cats and dogs. *2012 IEEE conference on computer vision and pattern recognition*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., & Clark, J. (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning*.

Shen, C.-W., Min, C., & Wang, C.-C. (2019). Analyzing the trend of O2O commerce by bilingual text mining on social media. *Computers in Human Behavior*, *101*, 474–483. doi:10.1016/j.chb.2018.09.031

Singh, D., Mathew, J., Agarwal, M., & Govind, M. (2023). DLIRIR: Deep learning based improved reverse image retrieval. *Engineering Applications of Artificial Intelligence*, *126*, 106833. doi:10.1016/j.engappai.2023.106833

Tian, S., Li, L., Li, W., Ran, H., Ning, X., & Tiwari, P. (2024). A survey on few-shot class-incremental learning. *Neural Networks*, *169*, 307–324. doi:10.1016/j.neunet.2023.10.039 PMID:37922714

Varish, N. (2022). A modified similarity measurement for image retrieval scheme using fusion of color, texture and shape moments. *Multimedia Tools and Applications*, *81*(15), 20373–20405. doi:10.1007/s11042-022-12289-1

Waqas, U., Visser, J. W., Choe, H., & Lee, D. (2023). Multi-modal fused deep learning networks for domain specific image similarity search. *Computers, Materials & Continua*, *75*(1), 243–258. doi:10.32604/cmc.2023.035716

Wolfe, J. M. (2021). Guided search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*, *28*(4), 1060–1092. doi:10.3758/s13423-020-01859-9 PMID:33547630

Ye, S., Yao, K., & Xue, J. (2023). Leveraging empowering leadership to improve employees' improvisational behavior: The role of promotion focus and willingness to take risks. *Psychological Reports*. doi:10.1177/00332941231172707 PMID:37092876

Ye, S., & Zhao, T. (2023). Team knowledge management: How leaders' expertise recognition influences expertise utilization. *Management Decision*, *61*(1), 77–96. doi:10.1108/MD-09-2021-1166

Yu, M., & Liu, X. (2022). Computer image content retrieval considering K-means clustering algorithm. *Mathematical Problems in Engineering*.

Yuan, L., Li, H., Fu, S., & Zhang, Z. (2022). Learning behavior evaluation model and teaching strategy innovation by social media network following learning psychology. *Frontiers in Psychology*, *13*, 843428. Advance online publication. doi:10.3389/fpsyg.2022.843428 PMID:35936300

Zhang, H., Fan, L., Chen, M., & Qiu, C. (2022). The impact of SIPOC on process reengineering and sustainability of enterprise procurement management in e-commerce environments using deep learning. *Journal of Organizational and End User Computing*, *34*(8), 1–17. Advance online publication. doi:10.4018/JOEUC.306270