


Article

An Image Recommendation Algorithm Based on Target Alternating Attention and User Affiliation Network

Shanshan Wan ^{1,2} , Shuyue Yang ¹, Ying Liu ¹, Jiaqi Ding ¹, Dongwei Qiu ³  and Chuyuan Wei ^{4,*}

- ¹ School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
- ² Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing 100044, China
- ³ School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
- ⁴ Network Information Center, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
- * Correspondence: weichuyuan@bucea.edu.cn

Abstract: Currently, how to exploit the deep features of images in image recommender systems to achieve image enhancement still needs further research. In addition, little research has explored the implicit and increasing preferences of users by using the affiliation generated by indirect users and virtual users of the main users, which leads to the phenomenon of information cocoon. An Image Recommendation Algorithm Based on Target Alternating Attention and User Affiliation Network (TAUA) is proposed in this paper that addresses the problems of inadequate extraction of semantic features in an image and information cocoon in image recommender systems. First, to complete the multi-dimensional description of the image, we extract the category, color, and style features of the image through a multi-channel convolutional neural network (MCNN), and we then perform migration and integration on these features. Then, to enhance the pixel-level representation ability of the image and achieve image feature enhancement, we propose target alternating attention to capture the information of surrounding pixels alternately from inside to outside. Finally, a user affiliation network, including indirect users and virtual users, is established according to the user behavior and transaction record, and the users' increasing preferences and affiliated users are mined through the implicit interaction relationship of users. Experimental results show that compared with baselines on the Amazon dataset, the results of F@10, NDCG@10, and AUC of the proposed algorithm are 4.02%, 5.00%, and 2.14% higher than those of ACF, and 5.76%, 0.86% and 1.16% higher than those of VPOI. On the Flickr dataset, our algorithm outperforms ACF by 5.74%, 5.12%, and 3.68% in F@10, NDCG@10, and AUC, respectively, and outperforms VPOI by 0.45%, 0.47%, and 0.49%. TAUA has better recommendation performance and can significantly improve the recommendation effect.

Keywords: image recommendations; multi-dimensional features; affiliation network; information cocoon; target alternating attention



Citation: Wan, S.; Yang, S.; Liu, Y.; Ding, J.; Qiu, D.; Wei, C. An Image Recommendation Algorithm Based on Target Alternating Attention and User Affiliation Network. *Appl. Sci.* **2023**, *13*, 4389. <https://doi.org/10.3390/app13074389>

Academic Editor: Vincent A. Cicirello

Received: 3 March 2023

Revised: 26 March 2023

Accepted: 27 March 2023

Published: 30 March 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Faced with the huge amount of data information on the Internet, traditional acquisition methods, such as keyword search, social media, online forums and communities, advertising, and information push, have difficulties in meeting the needs of users [1,2]. Using an image recommender system is the most effective way to solve this problem [3]. On the other hand, the rapid development of the Internet is accompanied by the problems of information overload and information scarcity [4]. According to Statista, the number of Internet users worldwide reached 4.901 billion in 2021, and the global Internet penetration rate reached 62.5%. In the global information system, the proportion of information waste is not less than 50%, and some disciplines even account for 80%. It takes a lot of time and energy to find the right information, and the cost of obtaining the information is increasing. Another problem caused by information overload in the Internet era is the information

cocoon, which confines user thinking in the information tower guided by interest, especially in the image recommender system [5]. Images have a higher complexity as well. When extracting semantic features, certain irrelevant information tends to interfere with the actual intent of users, resulting in recommendation bias. The image recommender system focuses on recommending items of interest in user historical behavior but does not have the mechanism to recommend implied preferences to users, which will lead to the narrowing of user information reception and increase the possibility of group polarization. Users and images are the two key and main information sources in the field of image recommendations, so it is particularly important to dig deeper into the content of images and users' implicit preferences for personalized recommendations.

Existing research has achieved some results in solving the problem of image features and the use of user data by merging multi-dimensional information. However, the current research ignores the features of users and items themselves and still faces some difficulties in image semantic feature extraction, how to collect image pixel contextual information, and how to deal with the information cocoon effect. The main manifestations are as follows:

1. Inadequate extraction of image semantic features

Image management first needs to obtain the dimensional information contained in the image, but e-commerce images contain complex and diverse semantic information. E-commerce images not only have the dimensions of category and color but also have other dimensions such as style. Each user also has their style preference when shopping. Style is a relatively stable characteristic, reflecting personal aesthetics and preferences expressed through items. The category and color of images describe explicit semantic information, and the style of images reflects other information.

2. The ability to capture pixel contextual information needs to be improved

Existing image processing methods, such as fully convolutional network (FCN), have made some progress, but due to their fixed geometry, they are limited to local receptive fields and short-range contextual information [6]. These limitations have a significant negative impact on FCN-based methods due to insufficient contextual information. Existing studies have proposed some methods to aggregate contextual information, but these methods have inherent shortcomings. For example, the method based on extended convolution cannot generate intensive contextual information [7,8]. The pool-based method cannot meet the requirement that different pixels require different contextual dependencies, and the criss-cross network (CCNet) is not comprehensive enough to extract critical information, and the semantic feature description is incomplete, resulting in lower accuracy [9,10].

3. Failure to pay attention to users' social adjoint relationships

A real personalized recommendation is not equal to only recommending the content that users have previously been interested in. A recommender system that conforms to the e-commerce platform should be sustainable and it can automatically update the preference information when users interact. At the same time, recommendations should be diverse. If users fall into the information cocoon, there will be problems of users' aesthetic fatigue and items' cold start. If the e-commerce platform falls into the information cocoon effect of a personalized recommendation, the thinking of users will be solidified, and the phenomenon of group polarization will occur.

Given the above problems in the image recommender system, this paper proposes an image recommendation algorithm based on target alternating attention and user affiliation network (TAUA). Specifically, the dual-channel convolutional neural network extracts the category and color features of images, while each channel in the multi-channel convolutional neural network (MCNN) corresponds to the learning tasks of different dimensions of images. The category, color, and style features of images are extracted to obtain image information for dimension migration fusion recommendations. Different from CCNet, which captures contextual information in a cross way, target alternating attention uses a recursive method to capture information in the global region of the image. When extracting

the features of a pixel, the information of surrounding pixel points is alternately captured from inside to outside according to the target method until the edge of the image.

On the other hand, different from the extraction of users' long-term and short-term preferences, user affiliation network includes affiliated users. To make full use of the implicit interaction between users, TAU integrates users' social records, long-term and short-term purchase preferences, and other information. Then, the implicit cross-fusion data among users, affiliated users, and items are found, recommending items of interest to users across-domains. The main contributions of this study are as follows:

1. A MCNN is proposed to enhance image feature description, solving the problem of inadequate semantic feature extraction to a certain extent, and to improve the ability of multi-dimensional image feature extraction.
2. We propose target alternating attention to collect pixel contextual information from images. Target alternating attention is used to enhance pixel-level representation and enrich item information, and further address the problem of information overload, enhance the integrity and reliability of the semantic feature description and obtain more accurate recommendation results.
3. A cross-domain user affiliation network is established to enrich the implicit user relationships, enhance users' knowledge, alleviate the information cocoon problem in the recommender system.

In order to mine the key implicit preferences of users, style feature is defined to potentially obtain more information about the item and user preferences. At the same time, to improve the representation ability of images, the convolutional neural network is modified to extract the multi-dimensional features of the image. To address the problem of incomplete extraction of image information, we propose target alternating attention to enhance item information. Moreover, to improve the diversity of recommendations and alleviate the problem of information cocoon, we focus on the affiliated user to mine the affiliated relationships of users and build the affiliation network.

The rest of the paper is organized as follows: Section 2 introduces the current research status of convolutional neural networks, cross-domain recommendations, and attention. Section 3 describes the details of the proposed image recommendation algorithm based on target alternating attention and user affiliation network (TAUA). Section 4 presents the experiment datasets, comparison approach, and evaluation indexes and analyzes the experimental results. Section 5 summarizes the main research contents and introduces future work, especially to improve the interpretability and diversity of image recommendations to carry out more in-depth research.

2. Related Work

This paper focuses on image feature extraction and user data processing and utilization. Currently, the main strategies to be solved include multi-dimensional information fusion, such as using deep learning for feature extraction, capturing important information using attention, and fusing multi-domain data by knowledge transfer. The following summarizes the research status and development trends of the three aspects related to our study.

2.1. Image Feature Data Extraction

Multi-attribute image classification is multi-task learning (MTL). The simplest method is to split it into simple independent single tasks for learning and then merge the results [11]. This means using several different convolutional neural networks to learn information, such as category, color, and style, from item images. However, for item images, if multiple attributes contained in an image are decomposed into independent issues, the correlation between multiple attributes will be ignored. For example, low-level features such as the edges of images are shared among multiple attributes.

Many scholars have studied multi-attribute image classification. Liu et al. divided human body images into several regions and then extracted the color histogram features of these regions to classify them [12]. Bossard et al. integrated the low-level features of

histograms of oriented gradient (HOG), speeded up robust features (SURFs), local binary patterns (LBPs), etc., and adopted Random Forest to achieve the classification of clothing multi-attribute images [13]. Li et al. constructed a deep learning framework—which recognizes multiple attributes jointly (DeepMAR)—model for pedestrian multi-attribute image classification [14]. Ak et al. used a method based on the combination of unsupervised segmentation and convolutional neural networks for multi-attribute e-commerce image retrieval [15].

The development of neural network image feature extraction technology can well solve the inconvenience caused by manual extraction [16]. Most of the current feature extraction methods are based on convolutional neural networks, which are widely used in the field of computer vision [17,18]. Their excellent characteristics, such as local connection, weight sharing, pooling operation, and multi-layer structure, have attracted the attention of many researchers. They automatically extract features from data through multi-layer nonlinear transformation, having strong expression ability and learning ability.

Abdulnabi et al. proposed a joint multi-task learning algorithm for better prediction of attributes in images using a deep convolutional neural network (CNN) [19]. Binary semantic attributes were learned through the multi-task CNN model. Krizhevsky et al. proposed the AlexNet deep convolutional neural network and applied the convolutional neural network to the automatic extraction of image features, forming abstract high-level features by combining low-level single features of the image and then using classifiers such as Softmax to classify the extracted features [20]. McAuley et al. used convolutional neural networks to learn semantic information about clothing and jewelry images and then recommended clothing matching to provide users with matching suggestions [21]. Some studies extracted item image features by using a convolutional neural network in combination with recommendation models, alleviating data sparsity and cold start problems of recommender systems to a certain extent [22,23]. Compared with the above method, which only considers the image features obtained by convolutional neural networks, Yu et al. also utilized the aesthetic features in images to improve the recommendation effect [24]. Geng et al. proposed a convolutional neural network model to learn the unified representation of users and images in social content networks so that the similarity between users and images can be measured and then recommend images [25].

The features extracted using traditional content-based recommendation algorithms include low-level features, such as color, textures, and high-level features, such as pyramid histogram of oriented gradients (PHOGs) [26]. Wang et al. introduced basic image low-level feature representation technology for color, texture, and shape features, and they proposed that the image characteristics with a single feature may lead to unsatisfactory image retrieval ability [27]. Some studies suggested that extracting image style features can better describe image semantic information. Li et al. believed that style features could better represent the overall image features and proposed a global style and local matching contrast learning network (GLCNet) for remote sensing image (RSI) semantic segmentation [28]. Gatys et al. found that the representation of content and style in convolutional neural networks is separable, and they proposed that style representation is a multi-scale representation including multi-layer neural networks [29].

Cao et al. designed a multi-task learning method based on an improved convolutional neural network [30]. The network used two channels to train the network simultaneously, and each channel was responsible for the learning tasks of different attributes in the image. The two learning tasks helped each other learn by sharing underlying parameters, which improved the convergence speed of the network and the generalization ability of the model.

Existing research has made a lot of contributions in terms of the high complexity of image feature extraction and the need for domain experts and manual intervention. However, the traditional convolutional neural networks only train and classify the dimensions of category and color of item images, ignoring important information, such as fabric, style, and brand of items. They cannot meet the actual needs of image feature extraction on e-commerce platforms. Designing a neural network that can extract multi-dimensional

features of images and introducing style features is conducive to more fully extracting image information for recommendation tasks, and mining user preferences with style features to make recommendations from more angles.

2.2. Attention Captures Pixel Contextual Information

Attention is the main means to address the problem of information overload, which is widely used in various tasks. Squeeze-and-Excitation networks enhance representation by modeling channel relationships in the attention [31]. Chen et al. used several attention masks to fuse feature maps or predictions of different branches [32]. Vaswani et al. applied the self-attention model to machine translation [33].

The attention model was first applied in the image recommendations to generate dense pixel-level contextual information; a point-wise spatial attention network (PSANet) realized long-distance contextual aggregation in scene analysis by learning point position-sensitive contextual dependency and two-way information propagation mode [34]. Non-local network utilizes self-attention to directly capture the long-distance dependence by calculating the interaction between any two locations, thus obtaining the complete image contextual information [33,35].

The complete image dependency provides useful contextual information to enhance image feature representation. Deeplabv2 proposes that atrous spatial pyramid pooling (ASPP) uses filters to detect the incoming convolutional feature layers at multiple sampling rates and effective field of view, to capture objects and image context at multiple scales [7]. Densely connected atrous spatial pyramid pooling (DenseASPP) densely connects a set of atrous convolutional layers to generate multi-scale features that cover larger sizes and densely cover that scale range [36]. Pyramid scene parsing network (PSPNet) uses a pyramid pool to mine global contextual information based on contextual aggregation in different regions [37]. Recently, Zhao et al. proposed a point-based spatial attention network, which aggregates remote contextual information by connecting all other locations with an adaptive predictive attention map [34]. Conditional random field (CRF) and Markov random field (MRF) are also used to capture the long-distance dependence of semantic segmentation [7,38]. Object context for semantic segmentation (OCNet) and dual attention network (DANet) use non-local blocks to obtain contextual information [35,39,40]. Huang et al. proposed the CCNet model to collect the contextual information of all pixels on the cross path and capture the complete image correlation through further loop operation [10]. However, this method is not comprehensive enough to extract and may ignore key information, resulting in low accuracy in extraction and high computational complexity.

Existing studies have made a lot of contributions in capturing image contextual information by using attention and can aggregate remote contextual information. However, for each pixel point, capturing pixel contextual information is not comprehensive enough, which may ignore key information, resulting in low accuracy in extraction and insufficient integrity and reliability in the semantic feature description. If attention can be proposed to collect the complete pixel contextual information more accurately and comprehensively, it can not only enrich the item information but can also enhance the semantic feature description of the image and mine the users' interest more comprehensively.

2.3. Cross-Domain User Social Relationships

Cross-domain recommender systems (CDRSs) can assist recommendations in the target domain based on the knowledge learned from the source domain [41]. The existing cross-domain data fusion, respectively, fuses the users' characteristics in different storage sources, and user data from multiple platforms are further fused through the corresponding relationship of users to obtain the cross-domain data of users from multiple platforms. The existing cross-domain data mining methods use cross-domain user identification, cross-domain recommendations, and other methods to extract the value of cross-domain data [42,43].

The cross-domain recommendation is proposed to solve the problem of data sparsity, and the rich training data in the source domain are used to improve the recommendation accuracy of the sparse domain. To improve the quality of recommendations, most of the existing cross-domain recommendation methods extract the domain-shared features among multiple related domains or integrate transfer learning with recommendation models. Some methods are based on matrix factorization (MF) and its variants, while others utilize clustering algorithms [44–47].

Ahmed et al. proposed a trust perception cross-domain deep neural matrix factorization (TCrossDNMF) model, which can address the user cold start problem in the cross-domain scenario of user overlap in the e-commerce system [48]. Yu et al. aligned the potential factors between the two domains based on pattern matching and transferred the user preferences of the auxiliary domain to update the original user potential vector in the target domain [49]. Ouyang et al. constructed a multi-graph according to user behavior in different fields and proposed a multi-graph neural network to learn cross-domain app embedding [50]. Liu et al. designed a new framework, a deep adversarial and attention network (DAAN), which considers both domain-shared knowledge and domain-specific knowledge across domains [51].

Users' decisions are affected by personal preferences and social relationships at the same time, so some studies are based on the homogeneity hypothesis; that is, users connection to each other in social networks tend to have similar preferences [52].

Feng et al. integrated user social networks into a random walk model with restart and conducted various tests on the internal correlation among group members to better describe group preferences and improved the performance of the group recommender system [53]. Li et al. proposed a social network recommendation method combining social tags and trust relationships [54]. Based on probability matrix factorization, major information related to social trust relationships, item tag information, and user rating matrix was collected, and all data resources from different dimensions were connected through shared user potential space (or item potential space). Yuan et al. proposed a unified framework to appropriately incorporate the influence of social relationships into the recommendations through the guidance of friends (friends who have a strong influence on users) and susceptibility (willingness to be affected) mining [55].

Traditional research on recommendations based on social relationships mostly focuses on analyzing the multifaceted impact of friends on user decisions and enriching user individual preferences. However, the structure of the social network has not changed. Although user purchases for others are taken into account, the recommendation algorithm has not been effectively applied. All the items of user historical behaviors are regarded as their preferences. As a result, when building the recommendation model, if there are items that do not belong to the users themselves, there may be a certain impact on the calculation of user preferences, reducing the accuracy of the recommendations. If users' historical behaviors can be more fully utilized, a user preference framework that is more in line with the real application of the recommendation system can be designed to more accurately simulate the real scene, test users' interests, balance the diversity of recommendations while improving the accuracy of the recommendation system, and help users avoid the information cocoon effect.

To sum up, although the development of image recommender systems is relatively mature, there are still some problems in the current research. Therefore, this paper combines neural network, attention, and user data to propose TAUA.

3. Methodology

The overall framework of TAUA is shown in Figure 1, which is divided into three steps: (1) build a MCNN to extract multi-dimensional features of category, color, and style of images; (2) construct the target alternating attention, aggregate the pixel contextual information of surrounding pixel points, and enhance the item features; (3) extract user

preferences, build affiliated users and user affiliation network, extract cross-fusion data between users and items, and recommend for users by cross-domains.

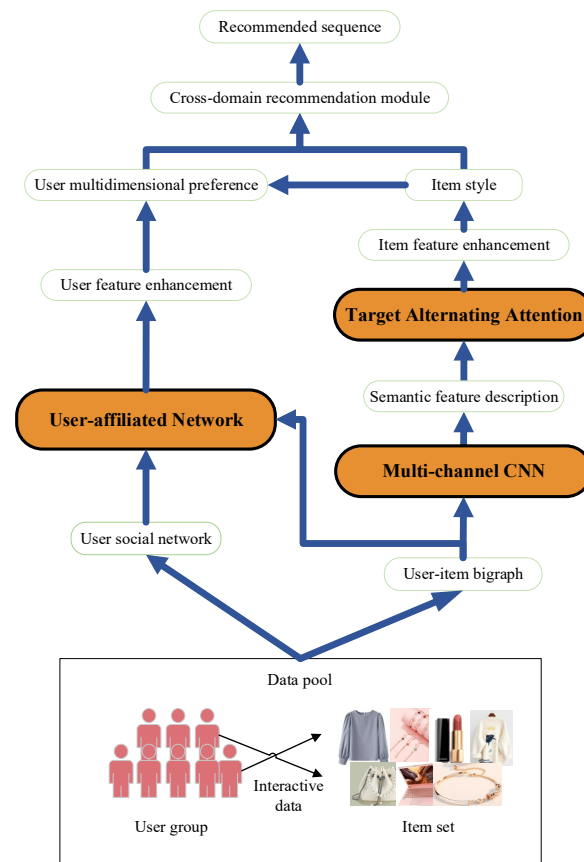


Figure 1. The overall framework of TAUU.

Firstly, the algorithm uses a distributed web crawler to obtain all kinds of item images and user data from the item pages of the e-commerce websites and store them in the database as the source data of the data pool. Then, the images in the data pool are preprocessed and input into the MCNN to extract the multi-dimensional features of the images. When the user inputs an item image, the MCNN method is also used to extract the multi-dimensional features of the target image to obtain the semantic feature description and image feature map of the image. The style dimension of the product image is used as the input of the cross-domain recommendations. At the same time, target alternating attention is utilized, image feature map is input into the module, and the pixel contextual information is aggregated by image correlation to enhance the item information.

The user social records and historical purchase preferences are fused to obtain the users' cross-platform data. The long-term and short-term preferences of the user are extracted, and the fusion preferences of the user are obtained by combining the long-term and short-term preferences. The affiliated users are mined from user historical preference items, and the user affiliation network is constructed based on the user social network. When the user interacts with an item, the algorithm automatically determines which affiliated user the item conforms to and adds it to the corresponding part of the user affiliation network to obtain the user multi-dimensional preferences. User preferences and item features are taken as the basis for cross-domain recommendations, and the implicit cross-fusion data between users and items are found through common dimension characteristics in the user affiliation network and image features. Based on the style dimension, this paper migrates the style features to more domains.

A flow chart of the image recommendation algorithm based on target alternating attention and user affiliation network is shown in Figure 2.

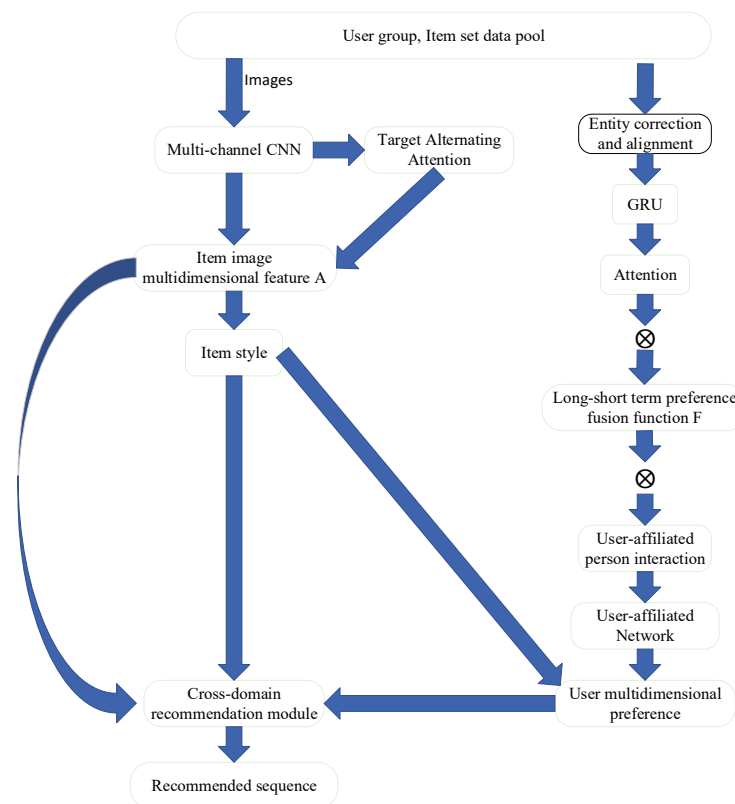


Figure 2. Flow chart of TAUA.





3.1. Multi-Channel Convolutional Neural Network Architecture

Given the problem that traditional convolutional neural networks cannot simultaneously extract features from multiple dimensions contained in e-commerce images, the network designed in this paper has multiple channels, each of which corresponds to the learning task of different dimensions of improving the feature extraction ability of multi-dimensional images.

Each user has their color or style preferences when shopping (extracted by attention in the user preference extraction section below), so users not only pay attention to the categories of items but also pay attention to the color and style of items and other information. Multi-attribute image means that an e-commerce image contains multiple dimensions. For example, a skirt can be described as a black preppy skirt, a dress can be described as a white French dress, a T-shirt can be described as a white leisure T-shirt, and a handbag can be described as a white ins handbag. The example are shown in Table 1. Each dimension describes e-commerce images from different angles and levels [19].

For the three dimensions of e-commerce image category, color, and style, the extracted features contribute to the classification of these three dimensions. Some prior distributions or model parameters can be shared in the learning tasks of these three dimensions, and these prior distributions or model parameters can be transferred during the training process. Fine-tuning is a common method in migration learning. It refers to using existing parameters to initialize new network parameters and transfer some tasks in the pre-training model to other tasks. In this way, the network can start learning from a good initial point, which can greatly save time when training new tasks.

Table 1. Dimension description of item image.

Item Image	Description
	black preppy skirt
	white French dress
	white leisure T-shirt
	white ins handbag

The MCNN is shown in Figure 3. The former part of the network, like the traditional network, has four convolutional layers, and each convolutional layer is connected with a pooling layer. The three dimensions share parameters in the first four convolution layers. From the fourth pooling layer, the network is divided into three channels. Each channel consists of two convolutional layers, one pooled layer, three fully connected layers, and a final Softmax classifier. The first channel is trained and classified according to the dimension of category, the second channel is trained and classified according to color, and the third channel is trained and classified according to style. Because the maximum pooling can learn the texture structure of the image, all network pooling modes are maximum pooling. In the latter part of the network, the network parameters of the three channels are the same, but the first channel outputs 12-dimensional vectors at the last fully connected layer, corresponding to 12 categories of the kind of dimension; the second channel outputs 3-D vectors, corresponding to the RGB value of the color dimension; the third channel outputs 12-dimensional vectors, corresponding to 12 categories of the style dimension. These three vectors will be input into three Softmax classifiers for classification, and the higher the output value, the greater the probability of belonging to the corresponding category.

When an image is input, three learning tasks can be carried out simultaneously; that is, the categories of kind, color, and style of the image can be predicted simultaneously through three classifiers.

This paper focuses on clothing, digital, and other items on JD, Taobao, Amazon, and other platforms. One-hot coding is carried out for the major categories.

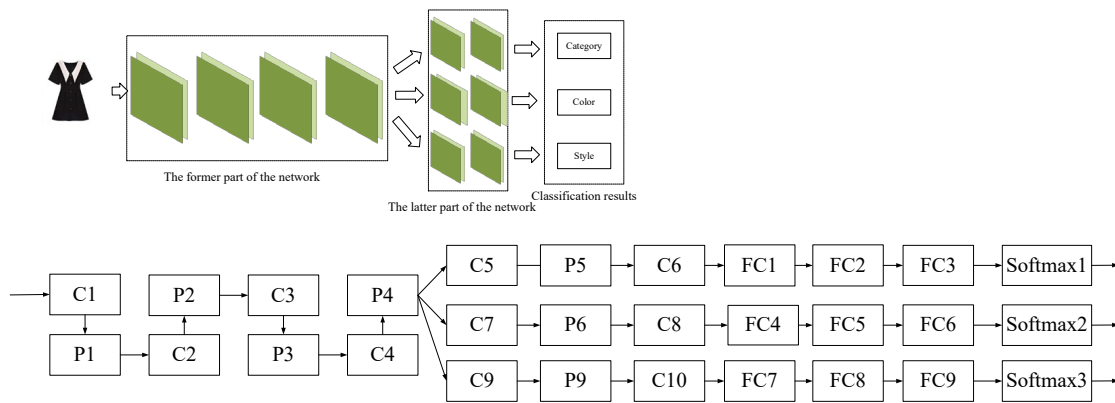


Figure 3. The architecture of the MCNN.

The training set image represented as $(w_1, x_1, y_1, z_1), (w_2, x_2, y_2, z_2), (w_3, x_3, y_3, z_3), \dots, (w_n, x_n, y_n, z_n)$, including $w_i \in R^{200 \times 200}, x_i \in R^{12}, y_i \in R^3, z_i \in R^{12}$. w_i is a third-order tensor with a size of 200×200 for each image. x_i is a 12-dimensional binary vector, and each dimension corresponds to 12 categories of images, the 12 categories are listed in Table 2, including women’s wear, ornaments, etc. Among them, others represents the item category that does not belong to the above 11 categories; y_i uses RGB color mode; z_i is a 12-dimensional binary vector, and each dimension corresponds to 12 styles of images, the 12 styles are listed in Table 3, including simple, retro, etc. Among them, others represents the item styles that does not belong to the above 11 styles. The expression of x_i and z_i using one-hot encoding is shown in Tables 2 and 3.

Table 2. Category one-hot encoding form.

Category	x_i	Category	x_i
women’s wear	(1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)	ornament	(0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
makeup	(0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0)	bag	(0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)
men’s wear	(0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0)	sport	(0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0)
appliance	(0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0)	furniture	(0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0)
mother-child	(0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0)	shoes	(0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0)
digital	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0)	others	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1)

Table 3. Style one-hot encoding representation.

Style	z_i	Style	z_i
simple	(1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)	retro	(0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
ins	(0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0)	preppy	(0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)
Korean	(0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0)	niche	(0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0)
French	(0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0)	luxurious	(0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0)
nordic	(0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0)	leisure	(0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0)
Japanese	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0)	others	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1)

x_i, y_i and z_i respectively, represent the category, color, and style of the image, such as when $x_m = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$, $y_m = (0, 0, 0)$ and $z_m = (0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0)$, the image w_m is a black Korean women’s image.

All images in the graphic library are classified according to category, color, and style using an MCNN network, and corresponding images in the database are marked. When a user uploads an interested item image, the MCNN determines the features of category, color, and style of the image $A = (x_i, y_i, z_i)$.

3.2. Target Alternating Attention

To aggregate more complete image information, this paper proposes target alternating attention (TAA), which aggregates the pixel contextual information of central pixel points in each surrounding layer according to the target to enhance item features, obtain important information more efficiently, and enhance the integrity and reliability of semantic feature description.

Target alternating attention collects pixel contextual information in a hierarchical direction to enhance the pixel-level representation ability. A frame diagram of target alternating attention is shown in Figure 4. Each pixel in the image is traversed from top to bottom and from left to right. For each pixel, the pixel contextual information of the center pixel around the pixel is mined from inside to outside according to the target, until the outermost pixel is traversed, to obtain more accurate recommendation results.

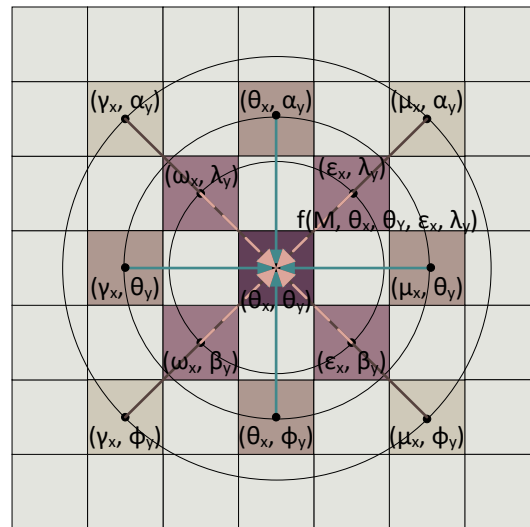


Figure 4. Structure of target alternating attention.

The input image is processed by the MCNN to generate the feature map X with space size $H \times W$. Given X , we first apply the convolutional layer to obtain the reduced-dimension feature map $H(0)$, and then send the feature map $H(0)$ into the target alternating attention. After the first loop, a new feature map $H(1)$ is generated, which aggregates the information of the four pixels of upper left, upper right, lower left, and lower right. To obtain richer and denser pixel contextual information, we send the feature map $H(1)$ again into the target alternating attention and output feature map $H(2)$. This process is repeated R times until the boundary pixel point position of the image, so that each position in the feature map $H(R)$ actually collects information from all pixels.

Target alternating attention collects information about four pixels per loop. Given the local feature map $H(0) \in \mathbb{R}^{C \times W \times H}$, this module firstly applies two convolutional layers with 1×1 filter to $H(0)$ to generate two feature maps, Q and K , where $\{Q, K\} \in \mathbb{R}^{C' \times W \times H}$. C' is the number of channels, less than C for dimension reduction.

After obtaining the feature maps Q and K , the attention map $M \in \mathbb{R}^{2 \times (H+W-4) \times W \times H}$ is further generated through Affinity. At each position u of the spatial dimension of feature map Q , the vector $Q_u \in \mathbb{R}^{C'}$ is obtained. At the same time, the set $\Omega_u \in \mathbb{R}^{2 \times (H+W-4) \times C'}$ can be obtained by extracting four feature vectors from K : upper left, upper right, lower left, lower right, or up, down, left, and right on each ring from position u outward. $\Omega_{i,u} \in \mathbb{R}^{C'}$ is the i th element of Ω_u . The Affinity operation is defined as follows:

$$d_{i,u} = \sqrt{(Q_u - \Omega_{i,u})(Q_u - \Omega_{i,u})^T} \tag{1}$$

where the $d_{i,u} \in D$ is the feature correlation between Q_u and $\Omega_{i,u}$, $i = [1, \dots, |\Omega_u|]$, $D \in \mathbb{R}^{2 \times (H+W-4) \times W \times H}$. Then, we apply the softmax layer on channel dimension D to calculate the attention map M .

Another convolutional layer with a 1×1 filter is applied to $H(0)$ to generate $V \in \mathbb{R}^{C \times W \times H}$ for feature adaptation. At each position u of the spatial dimension in the feature map V , we obtain the vector $V_u \in \mathbb{R}^C$ and a set of $\Phi_u \in \mathbb{R}^{2 \times (H+W-4) \times C}$. The set Φ_u is a set of four feature vectors: upper left, upper right, lower left, lower right, or up, down, left, and right, on each outward loop of V . The pixel contextual information is collected by the Aggregation operation:

$$H_u(j) = \begin{cases} \sum_{i \in |\Phi_u|} M_{i,u} \Phi_{i,u} + H(j-1), j \geq 1 \\ H(0), j = 0 \end{cases} \quad (2)$$

where $H_u(j)$ is the output feature mapping $H(j) \in \mathbb{R}^{C \times W \times H}$ at the j th layer circulation position u . $M_{i,u}$ is the scalar value in M at channel i and position u . Adding pixel contextual information to local feature H can enhance the representation of local feature and pixel.

Target alternating attention can be expanded into an R loop. In the first loop, the target alternating attention takes the feature map $H(0)$ extracted from the MCNN network as input and outputs the feature map $H(1)$, where $H(0)$ and $H(1)$ have the same shape. In the j th loop, the attention takes the feature map $H(j-1)$ as input and outputs the feature map $H(j)$. As shown in Figure 4, the target alternating is equipped with R loops, which can obtain complete image pixel contextual information from all pixels to generate new feature maps with dense and rich pixel contextual information.

$M(j)$ is represented as the attention map of the j th layer loop. For any pixel in the image (x, y) , from the position x', y' to the weight of the pixel $M_{j,x,y}$ mapping function is defined as $M_{j,x,y} = f(M, x, y, x', y')$, the loop j for the feature map $H(j)$ in any position u , propagation path of pixel contextual information in spatial dimension:

$$H_u(j) \leftarrow \left(\prod_{i=1}^j M_{j,x,y} \right) \cdot H_\theta \quad (3)$$

where \leftarrow is added to operation; $\theta \in \Phi_u$ is the pixel point for which information is to be collected for each loop of the target.

In e-commerce images, different colors often have specific styles, so after the pixel contextual information of the image is aggregated through the alternating attention, the image style is associated with color and category channels to a certain extent, which is used to enhance the extraction and representation of image style features.

3.3. User Affiliation Network

To make full use of the implicit social relationships between users, the user affiliation network is defined to realize more dimensional and efficient links between users with the help of more media. Through the construction of a cross-domain user affiliation network, the algorithm can automatically identify which affiliated users the items belong to, distinguish them from the real person, enhance the user information, and realize more accurate and comprehensive recommendations. For users, recommendations from multi-angle and multi-aspect not only improve the diversity of e-commerce recommendations but also help users break the information cocoon effect. For business, the accuracy and diversity of the recommender system play a crucial role in the e-commerce behavior of users, which determines whether the item can be found by users.

3.3.1. User Preference Combination Framework

Many user interests may change over time and may be triggered by specific contexts or time requirements. The long-term preference sequence of users is very rich and relatively stable, reflecting the overall trend of user interests, but there is a lot of redundant

information. The short-term preference sequence of users can more accurately reflect the changes in user interests in a short period, which plays a major role in the prediction, but it can also be easily affected by a single item. To make the recommender system accurately predict user interests, this paper combines long-term and short-term preferences, which can not only accurately grasp the overall trend of user preferences but also effectively reflect the evolution of user demands.

For the item set I interacting with users, GRUs are used to extract user short-term preferences, and different weights are assigned to user long-term preferences and immediate interests for item attributes through attention. Weighted calculation is used to model the final results of short-term interests and long-term preferences at the same time to obtain the vector representation of user preferences. When a user purchases an item, the system adds corresponding descriptions of category, color, and style to the user preference record according to the text description of the item in the item library and updates the user preference in time.

Among all the historical behaviors of users, only a part of historical behaviors can effectively affect the predictive attributes of the current recommended items, and each historical behavior has different impacts and contributions to user preferences. Therefore, attention is adopted to extract the importance of the user historical behaviors to the current recommended, namely the difference in weight.

For example, during festivals, people will buy some items in line with the festive atmosphere, but these items are rarely purchased at ordinary times, which may be different from the category, color, and style of the items purchased by users. The attributes of these items will have a negative impact on the accuracy of predicting user preferences. Therefore, when using attention to extract user long-term preferences, in the non-festive period, less attention is paid to items with festive significance; that is, less weight is given. During festive times, increased attention and greater weight are given to items that fit with the times. Similarly, since many products are only used during certain seasons, the attention assigns different weights depending on the season and item information.

t represents festive and season attributes, different users have different purchase preferences in different festival seasons, and each user is personalized to represent whether there is a special purchase preference at a certain time t , such as birthday, etc. Then, the probability p of user u buying a certain item i is a function of time t :

$$p_u = f(i, S, t) \quad (4)$$

where S is the quantity of item i purchased by all users at time t , and the larger b is, the more likely users are to buy similar items again at time t .

Taking the user short-term preference h_t extracted in chronological order as the input of Attention, the influence weight α of each item on the current recommendations in the historical behavior is a function of time t :

$$\alpha = \text{sigmoid}(f(p)) \quad (5)$$

The user long-term preference $L = \{l_1, l_2, \dots, l_m\}$ is a function of the user historical behavior of items and short-term preferences:

$$l = f(\alpha, h) \quad (6)$$

The user long-term preference $L = \{l_1, l_2, \dots, l_m\}$ and user short-term preference h are combined in a non-linear manner to obtain user long-term and short-term preference fusion F :

$$F = \sigma(f(h, l)) \quad (7)$$

where σ is a non-linear function.

Since the user long-term preference can reflect the user preferred style, the user preferred style can be found in the extraction of the user long-term preference, and the

style of users can be compared with the style of the item image in the recommendations. The cross-fusion data of user items can be extracted in the following for cross-domain recommendations.

3.3.2. User Affiliation Network

The schematic diagram of the affiliation network is shown in Figure 5. For each user, their affiliated users include both indirect and virtual users. In Figure 5, the pink figure means the real user, and the gray figure is the affiliated user. If users buy an item that does not fit their age profile, a virtual person is built for them. The preference matrix of affiliated users is constructed for each real person node. When the user purchases items, the algorithm automatically identifies which node belongs to according to the item information. When making recommendations, in addition to recommending the user’s own favorite items, users can also be recommended items in line with the preferences of other affiliated users. For example, if a middle-aged woman buys a children’s dress, a daughter’s affiliated user will be added to her affiliation network. The preference information of the affiliated user will be recorded according to the color and style of the dress, and recommend children’s items with the same or similar style according to the purchased dress to the user.

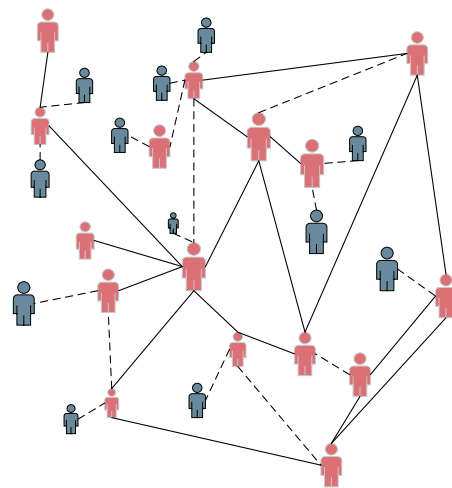


Figure 5. User affiliation network.

Definition 1. *Virtual users.* If a real user purchases an item that does not belong to him or her, a corresponding affiliated user is constructed, such as father, mother, child, friend, etc. sub represents the affiliated user node; then, the affiliated user node sub is:

$$sub = ext\{I | I \in (i_b - i_u)\} \tag{8}$$

where ext is the extraction operation, i_b is all the items purchased by user u , i_u is the items purchased by user u that conform to user identity information, and I is the set of items selected by the same affiliated user.

Definition 2. *Indirect user.* For user u_a , if there is another user u_b with the same or similar preferences as u_a in the category, color, and style of items, u_b is the indirect user of u_a , represented by ind ; then, the indirect user node ind of user u_a is:

$$ind = \{u_b | A_{u_b} \sim A_{u_a}\} \tag{9}$$

where \sim indicates similar preferences.

The preference relationship between the indirect and virtual user relative to the user itself is the affiliated relationships. The affiliated relationships of indirect users refer to the idea of collaborative filtering. If users have similar preferences and they have no

social relationship, they are indirect users of each other. The affiliation of virtual users is information about what the user has purchased for them. When making recommendations, the virtual user considers not only the user themselves but also the affiliated relationships of the user as a part of the user data, and makes recommendations separately from the user preference.

Definition 3. *Increasing preference.* The increasing preference pays attention to the changes in the interests of affiliated users, recommends items they continue to like in their historical preferences, and conducts interest testing with a small number of novel items that have never appeared, which is the key to interact and influence with the surrounding users.

Whether between users or between users and affiliated users, purchase behavior is crucial to the generation and change in the relationship chain between users. the relationship chain means that real users have the same preference, real users buy items for affiliated users, and the identity of affiliated users is consistent with the algorithm judgment.

Through the items purchased by users for different affiliated users, the preference information of affiliated users is obtained and, based on this, the user affiliation network is constructed to enhance the user preferences.

Definition 4. *Affiliated user preference attributes.* In the set of items purchased by the user for others, judge which affiliated users belong to according to the item features and add to the corresponding attribute; then, the preference attribute of affiliated user $w = \{w_1, \dots, w_n\}$ is:

$$w = \{A(I) | I \in (i_b - i_u)\} \quad (10)$$

In the user affiliation network, the interaction between users and the purchase of items for affiliated users is an explicit relationship, while the relationships between users with the same preference and affiliated users with the same identity are implicit.

Definition 5. *User affiliation network.* The user affiliation network, such as self, parents, friends, and children, is constructed according to the age and purchase information of real users. Each affiliated user is a node of the network, where R stands for affiliation network and pri stands for real user node. Then, user affiliation network R is:

$$R[i][j] = \begin{cases} 1, \{i \in j | i \in sub, j \in pri\} \\ 2, \{F_i = F_j | i, j \in pri\} \\ 0, rest \end{cases} \quad (11)$$

where 1 represents a solid line in the figure, which means the node itself has a social relationship; 2 represents a dotted line in the figure, which means the user's indirect or affiliated user; 0 means there is no edge in the figure.

The preferences of affiliated users will also affect the purchase demand of users. The user knowledge is enhanced according to the preferences of affiliated users to predict the preferences of other users with the same affiliated users and recommend the items purchased by the user to other users while recommending the items matching the identity or complementing the affiliated users to the user.

Based on the user and item dimension characteristics, the category, color, and style characteristics extracted above are predicted. In the affiliation network, the affiliated user is considered to be an independent entity in the recommendations, and user preferences are mined from the dimension perspective to make user preferences clear, which is conducive to more accurate discovery of user interests.

Definition 6. *User linkage relationship.* a is the affiliated user of A , b is the affiliated user of B , and a, b belong to the affiliated user of the same identity. Then, when A buys item v for a , the algorithm will automatically recommend the v to B and recommend the matching item for A . For example, if A is a 30-year-old female, a and b are children, and A buys ice skates for a , sports clothes are recommended for A and ice skates for B .

$$Rec(A) = \psi(F_A, w_a, v) \tag{12}$$

$$Rec(B) = \Phi(v) \tag{13}$$

where ψ and Φ are non-linear relations.

3.3.3. User Item Joint Recommendations

The phenomenon of information cocoon is quite common in recommender systems. The main reason is that individuals pursue personalized subjective needs, and the development of algorithm recommendation technology makes it more obvious.

User IDs and item IDs are encoded as one-hot, and the attribute data of the item are encoded as multi-hot, meaning that an item may correspond to multiple dimensions.

Definition 7. *User item joint recommendations.* Mine the similar attribute between user item target/source and recommend items with similar item style attributes for users with the same or similar attributes, as shown in Figure 6. If the target user is a 30-year-old female, and the historical preference is a simple bag, and there is another 30-year-old female in the user database who likes simple dress, simple dress is recommended for the target user.

$$Rec = F(u_s)_{species} \left[F(u_s)_{style} \cap F(u_t)_{style} \right] - i_s \tag{14}$$

where u_s is user source, $F(u_s)_{species}$ is the category preference of user source, $F(u_s)_{style}$ is the style preference of user source, u_t is user target, $F(u_t)_{style}$ is the style preference of user target, and i_s is item source.

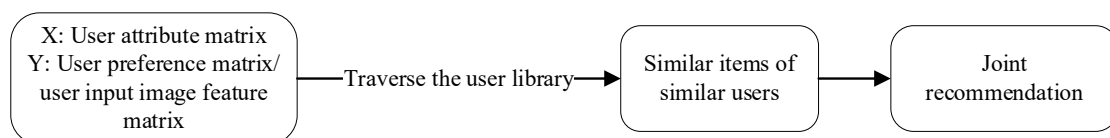


Figure 6. User item joint feature recommendations.

Jaccard distance is used to calculate the similarity between user target and user source, item target and item source, and Top-n recommendation is generated according to the similarity.

$$similarity(u_t, u_s) = \frac{|u_t \cap u_s|}{|u_t \cup u_s|} \tag{15}$$

$$similarity(i_t, i_s) = \frac{|i_t \cap i_s|}{|i_t \cup i_s|} \tag{16}$$

4. Experiment

In this section, we conduct a series of experiments to evaluate recommendation performance and demonstrate that our TAUA algorithm has superior performance compared to other good-performing methods. We conducted experiments on the Amazon and Flickr datasets and compared the performance of other models.

4.1. Datasets

This paper used Amazon Product Data (<http://snap.stanford.edu/data/amazon/productGraph/>, accessed on 26 March 2023) as the experimental dataset [56]. Amazon

Product Data is a public dataset. It recorded 82.83 million reviews and ratings of 9.35 million items from 20.98 million Amazon platform users from May 1996 to July 2014. The dataset includes reviews, item metadata, and link dataset files. Reviews include rating, text, and helpfulness vote, and metadata include description, category information, price, brand, and image features. This paper only used a subset dataset of item metadata and user interaction record to conduct experiments, including user ID, item ID, item image, etc. We also used the OpenImages5 dataset (<https://storage.googleapis.com/openimages/web/index.html>, accessed on 26 March 2023), which contains 15,440,132 objects each annotated with tags, spanning more than 600 categories [57]. However, since this dataset does not include user historical behavior information required in the recommendation task, this article used the officially published API to obtain user information from Flickr6 (<https://www.flickr.com/services/api>, accessed on 26 March 2023) [58].

4.2. Experimental Scheme

The experiments in this study were conducted on a desktop computer with 8 GB of RAM, an Intel Core i5-8250 CPU, and an Nvidia RTX 2070 GPU. The operating system was Windows 11 64-bit, and the software was Python 3.9 and MySQL 8.0. The models were implemented based on Tensorflow2.6.0, and Adam optimizer was used for training.

TAUA involves cross-domain recommendations; we referred to some cross-domain recommendation dataset settings based on previous work [59]. Specifically, for the Amazon dataset, since no images of items were given in the dataset and the label description cannot meet the needs of our experiments, we used crawler technology to obtain the original image of the commodity and the labels of category, color, and style to supplement the data information in the dataset. In addition, user's comment time in the dataset was approximately regarded as the interaction time between the user and item, which was regarded as the basis for holiday and season attributes.

For the Flickr dataset, we crawled the interactive feedback between the user and the item. Since the labels in the OpenImages5 dataset did not contain descriptions of image color and style, we used the MCNN model pre-trained on the Amazon dataset to extract and supplement the color and style labels of the image. Since users usually did not give their own personal information in e-commerce recommendations, we used the user affiliation network to find the user identity of the main user and infer the users' gender, age, and other information. The dataset contained many kinds of items, so we made recommendations in the feature dimension through the image features and labels extracted by the algorithm.

The datasets were further processed to create datasets suitable for our experiment. Users with less than 5 interactions and items with less than 10 interactions were removed from this article to ensure the validity of each user and item. In order to evaluate the performance of the proposed method, we designed different test schemes based on Table 4. First, we set different user sizes in the experiment. Sub-datasets with user sizes of 1000 were taken and represented as Amazon and Flickr. The data were randomly divided into the training set (80%) and the test set (20%), run five times, and the average value of the five experimental results was taken as the final evaluation value.

Table 4. Description of the processed datasets.

Datasets	User	Item	Interactions	Object Label
Amazon	1000	13,748	109,856	500
Flickr	1000	9492	86,147	550

MCNN vs. Dual CNN: In order to test the excellent classification ability of MCNN in extracting image features with the addition of style channels, we compared MCNN with dual-channel convolutional neural networks. Dual channel convolutional neural network extracted the category and color features of images, and MCNN introduced style features.

First, data preprocessing was conducted, and images were reshaped to 200×200 as the input of MCNN. The activation function adopted by each layer was Relu function, the learning rate was set to 0.00001, and the network was trained for a maximum of 100 iteration rounds. At the same time, in order to prevent overfitting and improve the generalization ability of the model, we used the dropout layer to normalize the data, and the ratio was set to 0.2. The network parameters of MCNN are shown in Tables 5 and 6.

Table 5. The former part of the MCNN parameters.

Network Layer	Kernel Size	Kernel Number	Output Dimension
Convolution layer 1	5×5	32	$200 \times 200 \times 32$
Pooling layer 1	2×2	-	$100 \times 100 \times 32$
Convolution layer 2	5×5	64	$100 \times 100 \times 64$
Pooling layer 2	2×2	-	$50 \times 50 \times 64$
Convolution layer 3	5×5	128	$50 \times 50 \times 128$
Pooling layer 3	2×2	-	$25 \times 25 \times 128$
Convolution layer 4	5×5	256	$25 \times 25 \times 256$
Pooling layer 4	2×2	-	$12 \times 12 \times 256$

Table 6. The latter part of the MCNN parameters.

Network Layer	Kernel Size	Kernel Number	Output Dimension
convolution layer 5	5×5	128	$12 \times 12 \times 128$
pooling layer 5	2×2	-	$6 \times 6 \times 128$
convolution layer 6	5×5	256	$6 \times 6 \times 256$
Fc1	-	-	2048
Fc2	-	-	512
Fc3	-	-	12
convolution layer 7	5×5	128	$12 \times 12 \times 128$
pooling layer 7	2×2	-	$6 \times 6 \times 128$
convolution layer 8	5×5	256	$6 \times 6 \times 256$
Fc4	-	-	2048
Fc5	-	-	3
Fc6	-	-	$256 \times 256 \times 256$
convolution layer 9	5×5	128	$12 \times 12 \times 128$
pooling layer 9	2×2	-	$6 \times 6 \times 128$
convolution layer 10	5×5	256	$6 \times 6 \times 256$
Fc7	-	-	2048
Fc8	-	-	512
Fc9	-	-	12

TAA vs. CCNet: To evaluate the accuracy and advance of target alternating attention, we compared the ability of the TAA and CCNet to extract pixel contextual information. CCNet collected contextual information by means of the cross, and TAA used target to obtain pixel contextual information. Hollow convolution was used in the convolution layer, and the step size of the output was set to 8. We used the ploy decay strategy, in which the initial learning rate was $lr \times \left(1 - \frac{iter}{max_iter}\right)^{power}$, where power = 0.9, momentum = 0.9, and weight attenuation is 0.00001.

UAN vs. long-term and short-term preferences: In order to measure the effectiveness of the user affiliation network, we conducted a comparative experiment between UAN and the users' long-term and short-term preference model. On the basis of extracting users' long-term and short-term preferences, UAN structured a user affiliation network and increased the preferences of the affiliated user. The time of user comments was used as a basis for distinguishing between short long-term and short-term preferences and time t . In the experiment, the number of GRU units was set to 32, the regularization coefficient was set to 1×10^{-6} , the vector representation dimension of users and items was set to 32, and the batch_size was set to 32.

Finally, in order to evaluate the performance difference between the TAUA and other algorithms, we compared our method with the following baselines, including BPR, VBPR, DVBPR, ACF, VPOI, LASSO, and VSM [58,60–65]:

1. BPR is a classic top-N recommendation method, which makes ranking optimization for each user preference. It is based on user image interaction, regardless of visual characteristics. This approach is considered the strongest baseline in the field of recommendations;
2. VBPR is an extension of BPR; it uses pre-trained CNN to extract visual features of item images and further combines image features with potential features for recommendations. In this approach, each image is treated as a single object;
3. DVBPR is an end-to-end training image feature and recommendation model integrating CNN and MF models;
4. ACF models project-level and component-level attention with two attention networks, and it simultaneously uses images and image regions to achieve a better recommender system. In this method, each image is divided into many regions of the same size;
5. VPOI uses visual content to enhance POI recommendations, extracts features from images through CNN, and uses them to guide the learning process of potential features between users and POI;
6. LASSO builds personalized models for users based on their favorite images;
7. VSM has fine-grained semantic features of images, and each image contains many semantic objects to better model image representation and user preferences.

Table 7 shows a comparison between the TAUA algorithm and the baselines.

Table 7. Comparison between TAUA with baseline methods. \checkmark means that the algorithm considers the factor, \times means the opposite.

Algorithms	Visual Feature	Style	Context	Affiliated Relation
BPR	\times	\times	\times	\times
VBPR	\checkmark	\times	\times	\times
DVBPR	\checkmark	\checkmark	\times	\times
ACF	\checkmark	\times	\times	\times
VPOI	\checkmark	\times	\times	\times
LASSO	\times	\times	\times	\times
VSM	\checkmark	\times	\times	\times
TAUA	\checkmark	\checkmark	\checkmark	\checkmark

The Comparison between TAUA with baseline methods is shown in Table 7. For all baselines, the results based on the recommended performance maintain the optimal setting of the hyper-parameters. Specifically, for VBPR, $\gamma_\theta = 10$, γ_E is always set to 0, and the dimension of visual feature is set to 4096, which is consistent with the original paper. The optimal learning rate of BPR and VBPR is 0.01. For ACF, the image is divided into 7×7 regions, as suggested in [63]. For VPOI, set $\alpha = 0.001$, $K = 10$, $\gamma_1 = \gamma_2 = 1$, $r = 5$. VGG16 is used for pre-training on ImageNet to initialize the weight to 3. For VSM, the learning rate is set to 0.1, and the regularization and minimum batch are set to 0.001 and 128, respectively. The potential vector dimension is 128. The optimum parameters proposed in the original literature are used in the experiment.

4.3. Performance Comparison

Three evaluation indexes of recommendation algorithms are used in the experiment, including F, NDCG, and AUC [66–68]. The three indexes were calculated for top@5 and top@10 respectively. Precision measures the accuracy of predicting positive sample results, and Recall measures the recall rate of results. The formula is as follows:

$$P = \frac{TP}{TP + FP} \tag{17}$$

$$R = \frac{TP}{TP + FN} \quad (18)$$

where TP represents items that the users like among the items predicted to be of interest to the users. FP represents the items predicted to be of interest to the users that the users do not like. FN represents the items that the users like among the items that are not predicted to be of interest to the users. However, Precision and Recall are often contradictory. Therefore, F is used to comprehensively consider Precision and Recall, so that both of them can reach the highest at the same time to achieve a balance:

$$F = \frac{2 * P * R}{P + R} \quad (19)$$

Normalized Discounted Cumulative Gain (NDCG) is an evaluation index that takes into account the return list to evaluate the accuracy of the list. The value ranges [0, 1]. The larger the value, the better the recommendation effect. AUC measures the probability that a model will predict a positive sample as a positive example rather than a negative sample as a positive example. The formula is as follows:

$$NDCG = \frac{DCG}{IDCG} \quad (20)$$

$$AUC = \frac{\sum_{ins_i \in positive} rank_{ins_i} - M * (M + 1) / 2}{M * N} \quad (21)$$

where DCG refers to Discounted Cumulative Gain, $IDCG$ refers to optimal DCG , and $NDCG$ is used to evaluate the accuracy of ranking. $rank_{ins_i}$ represents the sequence number of samples with the i th smallest probability score; M and N represent the number of positive and negative samples, respectively. AUC is used to evaluate the ranking quality of sample prediction.

4.4. Result

The performance comparison of all methods on Amazon Product Data and Flickr datasets is shown in Table 8. Compared with other performances, the performance of the TAUA algorithm is not obvious when the number of recommendations is small, but it can effectively recommend more items.

Table 8. Algorithm performance comparison.

Datasets	Models	F@5	F@10	NDCG@5	NDCG@10	AUC
Amazon	BPR	0.0241	0.0347	0.0393	0.0468	0.5310
	VBPR	0.0438	0.0512	0.0508	0.0541	0.7054
	DVBPR	0.0394	0.0355	0.0459	0.0517	0.6134
	ACF	0.0453	0.0547	0.0472	0.0559	0.7258
	VPOI	0.0447	0.0538	0.0525	0.0582	0.7328
	LASSO	0.0218	0.0323	0.0377	0.0445	0.5218
	VSM	0.035	0.0563	0.0479	0.0578	0.717
	TAUA	0.0449	0.0569	0.0497	0.0587	0.7413
	BPR	0.0177	0.0305	0.0242	0.0329	0.5543
	VBPR	0.0239	0.0439	0.0282	0.0317	0.6140
Flickr	DVBPR	0.0196	0.0328	0.0254	0.0336	0.5642
	ACF	0.0297	0.0418	0.0333	0.0410	0.6468
	VPOI	0.0248	0.0440	0.0371	0.0429	0.6673
	LASSO	0.0169	0.0296	0.0235	0.0314	0.5235
	VSM	0.0262	0.0441	0.0356	0.0417	0.6560
	TAUA	0.0268	0.0442	0.0324	0.0431	0.6706

Figures 7 and 8 show the NDCG of the proposed TAUA algorithm compared to other algorithms. Compared with VPOI and VSM, the TAUA algorithm also takes into account the pixel contextual information on the image and the users' affiliated user preference. The results show that the performance of TAUA is obviously better than other algorithms, which is more in line with the requirements of image recommendations.

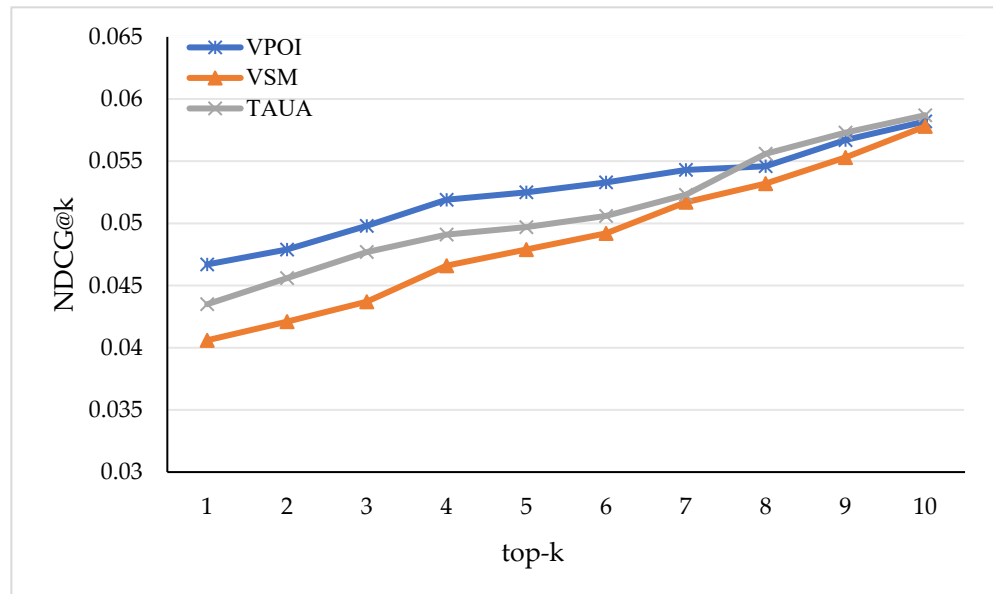


Figure 7. Top-k image recommendations NDCG@k performance comparison on Amazon datasets.

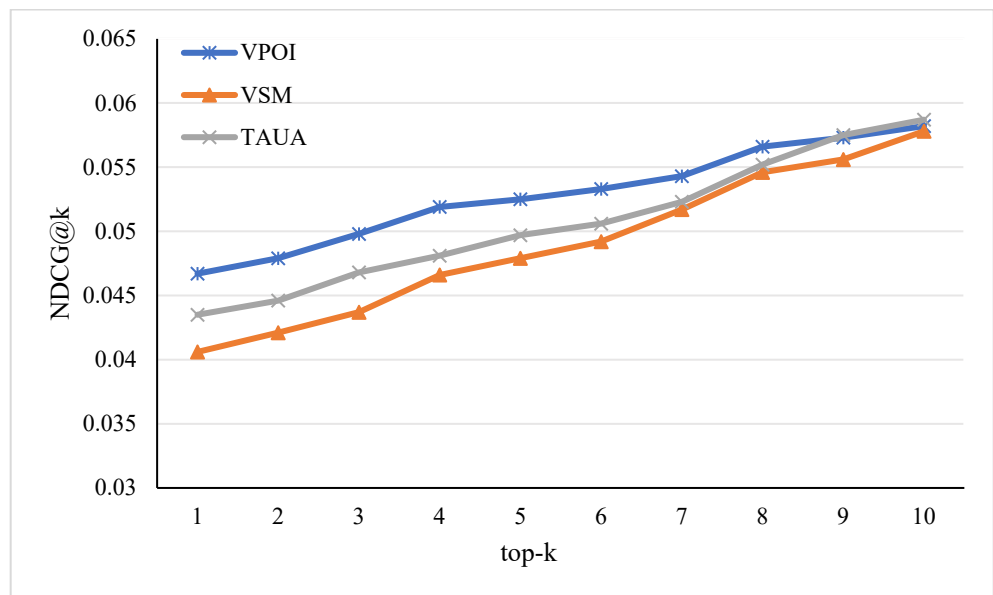


Figure 8. Top-k image recommendations NDCG@k performance comparison on Flickr datasets.

In order to prove the effectiveness of each module in the proposed TAUA algorithm, we carried out a comparative experiment. Specifically, MCNN is compared with dual-channel convolutional neural networks to compare their ability to classify image features. Compared with the two-channel convolutional neural network, the MCNN adds the style channel to extract the multi-dimensional features of the image. The experimental results are shown in Figure 9. The classification results of MCNN under different learning rates are shown in Table 9.

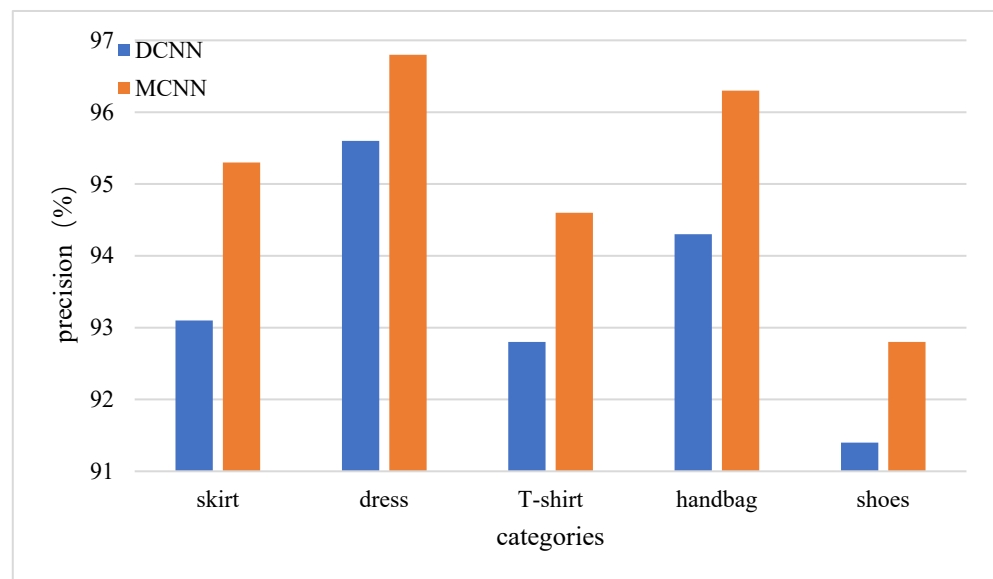


Figure 9. Comparison of classification accuracy between MCNN and dual-channel convolutional neural network.

Table 9. Performance of MCNN classification results.

Attribute	Learning Rate	Precision (%)
category	0.001	92.3
color	0.001	95.9
style	0.001	90.4
category	0.0001	94.6
color	0.0001	96.3
style	0.0001	92.7
category	0.00001	95.3
color	0.00001	96.8
style	0.00001	91.9

In order to prove the accuracy of TAA in collecting pixel contextual information, we conducted a comparison experiment between TAA and CCNet. Table 10 proves that TAA is superior to CCNet in aggregating image pixel contextual information.

Table 10. Comparison between TAA and CCNet aggregated pixel contextual information performance.

Datasets	Method	mIOU (%)
Amazon	CCNet	81.3%
Amazon	TAA	82.1%
Flickr	CCNet	80.2%
Flickr	TAA	80.8%

Based on the extraction of users’ long-term and short-term preferences, the affiliation network adds affiliated users, which not only considers indirect users but also sets up virtual characters. The image recommendation algorithm based on UAN is compared with the image recommendation algorithm based on users’ long-term and short-term preferences. Figure 10 shows that UAN is effective in processing users’ preferences, and it makes more effective use of users’ historical behaviors to improve the diversity of recommendations.

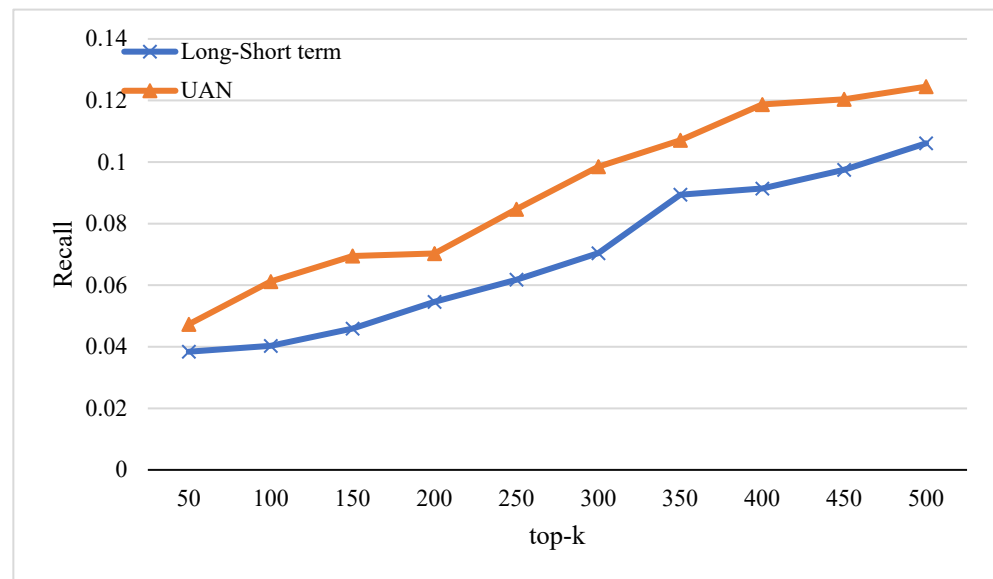


Figure 10. Comparison of top-k recommendation recall rate between image recommendation algorithm based on UAN and image recommendation algorithm based on users' long-term and short-term preferences.

4.5. Ablation Experiment

To further analyze the roles played by MCNN, user affiliation network (UAN), and target alternating attention TAA in TAUA algorithm, TAUA ablation experiments were conducted: (1) image features were extracted by traditional convolutional neural network instead of MCNN and applied to the algorithm in this study; (2) we made use of traditional user long-term and short-term preferences to extract the framework to obtain user interests for recommendations; (3) pixel contextual information is not collected through TAA. The experimental results are shown in Tables 11 and 12. Based on this, the following conclusions can be shown:

Table 11. Performance ablation of TAUA in Amazon dataset.

Model	F@10	NDCG@10	AUC
TAUA	0.0569	0.0587	0.7413
CNN + UAN + TAA	0.0532	0.0563	0.7258
MCNN + UAN	0.0557	0.0581	0.7306
MCNN + TAA	0.0548	0.0579	0.7295

Table 12. Performance ablation of TAUA in Flickr dataset.

Model	F@10	NDCG@10	AUC
TAUA	0.0442	0.0431	0.6706
CNN + UAN + TAA	0.0428	0.0417	0.6493
MCNN + UAN	0.0436	0.0429	0.6614
MCNN + TAA	0.0433	0.0426	0.6537

4.6. Analysis and Discussion

In this paper, we propose a TAUA model to provide accurate and diverse recommendation lists for e-commerce image recommendations. As can be seen from the results in Table 8, compared with the baseline model, the TAUA algorithm in this paper generally achieved the best performance in all evaluation indexes. Specifically, in the Amazon Product Data dataset, the F@10, NDCG@10, and AUC of our algorithm are 4.02%, 5.00%, and 2.143% higher than ACF, respectively, and 5.76%, 0.86%, and 1.16% higher than that of the VPOI.

On the Flickr dataset, our algorithm outperforms the ACF model by 5.74%, 7.565.12%, and 3.68% in F@10, NDCG@10, and AUC, respectively, and outperforms the VPOI model by 0.45%, 0.47%, and 0.49%. The TAUA algorithm is superior to VBPR in the recommendation performance, which shows that extracting deeper-level image information in the image recommendation task can better represent the image. The specific discussion and analysis are as follows:

1. When $k = 5$, the proposed TAUA algorithm is slightly lower than the existing algorithm in terms of F and NDCG. As the number of recommendations increases, the performance of the TAUA algorithm is better than that of other algorithms. This is mainly due to the accuracy and diversity of our proposed algorithm. Since the accuracy and diversity are contradictory, MCNN and TAA reduce the recall to some extent, while UAN reduces the accuracy. Therefore, when the number of recommendations is small, there will be a large difference between accuracy and diversity. When the number of recommendations increases, accuracy and recall of the algorithm gradually tend to balance, both reach the highest simultaneously, and the algorithm performance reaches the optimum.
2. With an increase in k , the overall performance of the TAUA algorithm is better than VBPR, indicating that our algorithm is effective for image recommendations. The TAUA algorithm has made contributions in extracting image features and pixel context, obtaining more accurate image semantic descriptions, mining item features in a deeper and more comprehensive way and predicting user preferences.
3. The performance of the TAUA algorithm is better than other algorithms on the whole, which shows that the algorithm we proposed is effective. From the perspective of item and user, this paper extracts item features and user interests. Mainly, item features extracted through MCNN and TAA are used as the basis for judging user preferences, which reduces the deviation in understanding item and user preferences and ensures the consistency of data content utilization.

Compared with BPR, LASSO, and VSM algorithms, which only consider the users' preference for an item, TAUA takes the style of the items as one important users' style preference, introduces the affiliation network, divides the users' historical behavior into the preference of different virtual users, and mines the users' implicit preferences. VBPR, DVBPR, and VPOI algorithms extract image features through CNN for recommendations, but the traditional CNN only focuses on the single dimension of the image; the TAUA algorithm extracts the category, color, and style features of the image through multiple channels and extracts features from more dimensions for recommendations. DVBPR and ACF divide the image into many regions and only pay attention to the local information of the image. The algorithm we proposed captures the pixel contextual information on the whole image by using the target alternating attention, which pays attention to both the local information and the global information in the image.

We compared the influences of different learning rates on the accuracy of image feature extraction on MCNN. With the increase in learning rate, the classification accuracy of MCNN is improved. When the learning rate is 0.00001, the performance of MCNN is optimal. Under the three learning rates, the classification accuracy of color dimension is higher, while the classification accuracy of style dimension is lower. This is because the features of style dimension are relatively more complex and require more iterations for training. Comparing the classification ability of MCNN and DCNN to extract image features, MCNN has higher accuracy for the same kind of items, indicating that increasing the dimension of style can improve the performance of image classification. In the same dataset, TAA has higher mIOU than CCNet, which proves that TAA has higher detection accuracy and can collect pixel contextual information of images more comprehensively. The UAN is introduced to make use of the historical behavior information of users from the perspective of user-affiliated relationships, and UAN has better performance in the top-k recommendation Recall. Moreover, with an increase in the number of recommended items, the Recall increases, indicating that the UAN plays a certain role in improving the diversity

of recommendations. This is because UAN predicts user needs from both indirect users and virtual users.

We also conducted ablation experiments on the TAUA algorithm. Based on Tables 9 and 10, the following conclusions can be obtained:

1. The MCNN is replaced by the traditional convolutional neural network, and the AUC performance of the algorithm decreases on both datasets. Therefore, it can be concluded that the MCNN has great potential in extracting image features.
2. Instead of collecting pixel contextual information through TAA, only MCNN and UAN are used for recommendations, and the performance of the TAUA algorithm is slightly reduced. Therefore, TAA has a certain effect in collecting pixel contextual information.
3. UAN of TAUA is deleted, and we directly used the traditional user preference extraction framework to add it to the model. It can be seen that the performance of the two datasets decreased compared with the TAUA algorithm, which means that the proposed UAN is effective.

5. Conclusions and Future Work

By inputting various attributes and features of users and items, the recommendation system outputs a list of recommendations ranked according to user preferences. User information, item information, and contextual information are the main data sources in recommender systems. However, for image recommendations, image feature extraction and user preference mining are very important, which affect the performance of the recommender system. Although some methods for image feature extraction have achieved certain effects, they still fail to fully excavate deep information, and their effects need to be further improved. The development of deep learning and neural networks brings opportunities to recommender systems, but it requires large-scale data and the calculation cost is relatively high, which limits its application in the recommendation domain to a certain extent.

In this study, we apply knowledge transfer and attention to recommender systems. New attention is designed based on the intrinsic relationships between image pixels. Based on this, an image recommendation algorithm based on target alternating attention is proposed for better recommend e-commerce users from the perspective of images. Specifically, a user item joint recommendation method is designed to realize cross-domain recommendations by mining hidden cross-fusion preferences between users and items. Experimental results on the Amazon and Flickr datasets show that the proposed method is effective in image processing and recommendations and successfully improves the ability of image recommendations in the e-commerce field.

TAUA, as a deep-learning-based recommendation method, provides positive guidance in the field of image recommendations and can be applied to more fields. However, the multi-dimensional image feature extraction designed in this paper only trains and classifies the categories, colors, and styles of images, which still falls short of the actual needs of e-commerce platforms. In future work, in addition to style features, more useful user features and items can be explored. The number of MCNN channels can be expanded. Target alternating attention collects more accurate pixel contextual information, but the computational complexity still needs to be reduced. In the future, we will try to design a more concise attention mechanism to further reduce the complexity of image recommendation algorithms.

Author Contributions: Conceptualization, S.W. and D.Q.; methodology, S.W., S.Y. and Y.L.; software, D.Q.; validation, J.D. and C.W.; formal analysis, Y.L.; investigation, S.W.; data curation, S.Y.; writing—original draft preparation, S.W. and S.Y.; writing—review and editing, J.D. and C.W.; supervision, C.W.; project administration, C.W.; funding acquisition, C.W. All authors have read and agreed to the published version of the manuscript.

Funding: The authors wish to acknowledge the support of the National Natural Science Foundation of China (No. 61902016), the Education and Research Project of Beijing University of Civil Engineering and Architecture (No. Y2009), the Postgraduate Education and Teaching Quality Improvement Project of Beijing University of Civil Engineering and Architecture, China (No. J2023002, J2022005), and BUCEA Post Graduate Innovation Project (No. PG2023086, PG2022080, PG2022082).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article. We conducted experiments on two benchmark datasets: Amazon Product Data (link: <http://snap.stanford.edu/data/amazon/productGraph/>, accessed on 26 March 2023), and OpenImages5 (link: <https://storage.googleapis.com/openimages/web/index.html>, accessed on 26 March 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Warren, J. *Big Data: Principles and Best Practices of Scalable Realtime Data Systems*; Manning: Shelter Island, NY, USA, 2015.
2. Fisher, M.; Smiley, A.H.; Grillo, T.L. Information without knowledge: The effects of Internet search on learning. *Memory* **2022**, *30*, 375–387. [[CrossRef](#)] [[PubMed](#)]
3. Ricci, F.; Rokach, L.; Shapira, B. Recommender systems: Introduction and challenges. In *Recommender Systems Handbook*; Ricci, F., Rokach, L., Shapira, B., Eds.; Springer: Boston, MA, USA, 2015; pp. 1–34.
4. Bollen, D.; Knijnenburg, B.P.; Willemsen, M.C.; Graus, M. Understanding choice overload in recommender systems. In Proceedings of the Fourth ACM Conference on Recommender Systems, Barcelona, Spain, 26–30 September 2010; pp. 63–70.
5. Ji, L. How to crack the information cocoon room under the background of intelligent media. *Int. J. Soc. Sci. Educ. Res.* **2020**, *3*, 169–173.
6. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
7. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [[CrossRef](#)] [[PubMed](#)]
8. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
9. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7151–7160.
10. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 603–612.
11. Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv* **2017**, arXiv:1706.05098.
12. Liu, S.; Song, Z.; Wang, M.; Xu, C.; Lu, H.; Yan, S. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In Proceedings of the 20th ACM International Conference on Multimedia, Nara, Japan, 29 October–2 November 2012; pp. 1335–1336.
13. Bossard, L.; Dantone, M.; Leistner, C.; Wengert, C.; Quack, T.; Van Gool, L. Apparel classification with style. In Proceedings of the Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Republic of Korea, 5–9 November 2012; Revised Selected Papers, Part IV 11. pp. 321–335.
14. Li, D.; Chen, X.; Huang, K. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In Proceedings of the 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, 26–29 November 2017; pp. 111–115.
15. Ak, K.E.; Lim, J.H.; Tham, J.Y.; Kassim, A.A. Efficient multi-attribute similarity learning towards attribute-based fashion search. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1671–1679.
16. Aich, S.; Yamazaki, M.; Taniguchi, Y.; Stavness, I. Multi-scale weight sharing network for image recognition. *Pattern Recognit. Lett.* **2020**, *131*, 348–354. [[CrossRef](#)]
17. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
18. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [[CrossRef](#)]

19. Abdulnabi, A.H.; Wang, G.; Lu, J.; Jia, K. Multi-task CNN model for attribute prediction. *IEEE Trans. Multimed.* **2015**, *17*, 1949–1959. [[CrossRef](#)]
20. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [[CrossRef](#)]
21. McAuley, J.; Targett, C.; Shi, Q.; Van Den Hengel, A. Image-based recommendations on styles and substitutes. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, 9–13 August 2015; pp. 43–52.
22. Li, Z.; Chen, H.; Ni, Z.; Deng, X.; Liu, B.; Liu, W. ARPCNN: Auxiliary Review-Based Personalized Attentional CNN for Trustworthy Recommendation. *IEEE Trans. Ind. Inform.* **2022**, *19*, 1018–1029. [[CrossRef](#)]
23. Xiong, M.T.; Feng, Y.; Wu, T.; Shang, J.X.; Qiang, B.H.; Wang, Y.N. TDCTFIC: A novel recommendation framework fusing temporal dynamics, CNN-based text features and item correlation. *IEICE Trans. Inf. Syst.* **2019**, *102*, 1517–1525. [[CrossRef](#)]
24. Yu, W.; Zhang, H.; He, X.; Chen, X.; Xiong, L.; Qin, Z. Aesthetic-based clothing recommendation. In Proceedings of the 2018 World Wide Web Conference, Lyon, France, 23–27 April 2018; pp. 649–658.
25. Geng, X.; Zhang, H.; Bian, J.; Chua, T.-S. Learning image and user features for recommendation in social networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 4274–4282.
26. Sha, D.; Wang, D.; Zhou, X.; Feng, S.; Zhang, Y.; Yu, G. An approach for clothing recommendation based on multiple image attributes. In Proceedings of the Web-Age Information Management: 17th International Conference, WAIM 2016, Nanchang, China, 3–5 June 2016; Part I 17. pp. 272–285.
27. Wang, S.; Han, K.; Jin, J. Review of image low-level feature extraction methods for content-based image retrieval. *Sens. Rev.* **2019**, *39*, 783–809. [[CrossRef](#)]
28. Li, H.; Li, Y.; Zhang, G.; Liu, R.; Huang, H.; Zhu, Q.; Tao, C. Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [[CrossRef](#)]
29. Gatys, L.A.; Ecker, A.S.; Bethge, M. A neural algorithm of artistic style. *arXiv* **2015**, arXiv:1508.06576. [[CrossRef](#)]
30. Cao, Z.; Shaomin, M.; Yongyu, X.; Dong, M. Image retrieval method based on CNN and dimension reduction. In Proceedings of the 2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), Jinan, China, 14–17 December 2018; pp. 441–445.
31. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [[CrossRef](#)]
32. Chen, L.-C.; Yang, Y.; Wang, J.; Xu, W.; Yuille, A.L. Attention to scale: Scale-aware semantic image segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 3640–3649.
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In *Advances In Neural Information Processing Systems 30, Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017*; Curran Associates, Inc.: New York, NY, USA, 2018.
34. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Loy, C.C.; Lin, D.; Jia, J. Psanet: Point-wise spatial attention network for scene parsing. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 267–283.
35. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
36. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3684–3692.
37. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–25 July 2017; pp. 2881–2890.
38. Liu, Z.; Li, X.; Ping, L.; Chen, C.L.; Tang, X. Semantic Image Segmentation via Deep Parsing Network. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015.
39. Yuan, Y.; Huang, L.; Guo, J.; Zhang, C.; Chen, X.; Wang, J. Ocnnet: Object context network for scene parsing. *arXiv* **2018**, arXiv:1809.00916.
40. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154.
41. Gahier, A.K.; Gujral, S.K. Cross Domain Recommendation Systems using Deep Learning: A Systematic Literature Review. In Proceedings of the International Conference on Innovative Computing & Communication (ICICC), Delhi, India, 20–21 February 2021.
42. Shu, K.; Wang, S.; Tang, J.; Zafarani, R.; Liu, H. User identity linkage across online social networks: A review. *Acm Sigkdd Explor. Newsl.* **2017**, *18*, 5–17. [[CrossRef](#)]
43. Khan, M.M.; Ibrahim, R.; Ghani, I. Cross domain recommender systems: A systematic literature review. *ACM Comput. Surv. (CSUR)* **2017**, *50*, 1–34. [[CrossRef](#)]
44. Singh, A.P.; Gordon, G.J. Relational learning via collective matrix factorization. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV, USA, 24–27 August 2008; pp. 650–658.
45. Pan, W.; Xiang, E.; Liu, N.; Yang, Q. Transfer learning in collaborative filtering for sparsity reduction. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 22 February–1 March 2022; pp. 230–235.
46. Li, B.; Yang, Q.; Xue, X. Can movies and books collaborate? cross-domain collaborative filtering for sparsity reduction. In Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence, Pasadena, CA, USA, 11–17 July 2009.

47. Li, B.; Yang, Q.; Xue, X. Transfer learning for collaborative filtering via a rating-matrix generative model. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; pp. 617–624.
48. Ahmed, A.; Saleem, K.; Khalid, O.; Rashid, U. On deep neural network for trust aware cross domain recommendations in E-commerce. *Expert Syst. Appl.* **2021**, *174*, 114757. [CrossRef]
49. Yu, X.; Hu, Q.; Li, H.; Du, J.; Gao, J.; Sun, L. Cross-domain recommendation based on latent factor alignment. *Neural Comput. Appl.* **2022**, *34*, 3421–3432. [CrossRef]
50. Ouyang, Y.; Guo, B.; Tang, X.; He, X.; Xiong, J.; Yu, Z. Mobile app cross-domain recommendation with multi-graph neural network. *ACM Trans. Knowl. Discov. Data (TKDD)* **2021**, *15*, 1–21. [CrossRef]
51. Liu, H.; Guo, L.; Li, P.; Zhao, P.; Wu, X. Collaborative filtering with a deep adversarial and attention network for cross-domain recommendation. *Inf. Sci.* **2021**, *565*, 370–389. [CrossRef]
52. McPherson, M.; Smith-Lovin, L.; Cook, J.M. Birds of a Feather: Homophily in Social Networks. *Annu. Rev. Sociol.* **2001**, *27*, 415–444. [CrossRef]
53. Feng, S.; Zhang, H.; Cao, J.; Yao, Y. Merging user social network into the random walk model for better group recommendation. *Appl. Intell.* **2018**, *49*, 2046–2058. [CrossRef]
54. Li, H.; Zhang, S.; Hu, Y.; Shi, J.; Zhong, Z.M. Research of social recommendation based on social tag and trust relation. *Clust. Comput.* **2017**, *21*, 933–943. [CrossRef]
55. Yuan, T.; Cheng, J.; Zhang, X.; Liu, Q.; Lu, H. How friends affect user behaviors? An exploration of social relation analysis for recommendation. *Knowl.-Based Syst.* **2015**, *88*, 70–84. [CrossRef]
56. Ni, J.; Li, J.; McAuley, J. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 188–197.
57. Krasin, I.; Duerig, T.; Alldrin, N.; Ferrari, V.; Abu-El-Haija, S.; Kuznetsova, A.; Rom, H.; Uijlings, J.; Popov, S.; Veit, A. Openimages: A Public Dataset for Large-Scale Multi-Label and Multi-Class Image Classification. Available online: <https://github.com/openimages> (accessed on 26 March 2023).
58. Guo, G.B.; Meng, Y.; Zhang, Y.F.; Han, C.Y.; Li, Y.J. Visual Semantic Image Recommendation. *IEEE Access* **2019**, *7*, 33424–33433. [CrossRef]
59. Zhu, Y.; Ge, K.; Zhuang, F.; Xie, R.; Xi, D.; Zhang, X.; Lin, L.; He, Q. Transfer-meta framework for cross-domain recommendation to cold-start users. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event Canada, 11–15 July 2021; pp. 1813–1817.
60. Rendle, S.; Freudenthaler, C.; Gantner, Z.; Schmidt-Thieme, L. BPR: Bayesian personalized ranking from implicit feedback. *arXiv* **2012**, arXiv:1205.2618.
61. He, R.; McAuley, J. VBPR: Visual bayesian personalized ranking from implicit feedback. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 22 February–1 March 2022.
62. Kang, W.-C.; Fang, C.; Wang, Z.; McAuley, J. Visually-aware fashion recommendation and design with generative image models. In Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM), New Orleans, LA, USA, 18–21 November 2017; pp. 207–216.
63. Chen, J.; Zhang, H.; He, X.; Nie, L.; Liu, W.; Chua, T.-S. Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Tokyo, Japan, 7–11 August 2017; pp. 335–344.
64. Wang, S.; Wang, Y.; Tang, J.; Shu, K.; Ranganath, S.; Liu, H. What your images reveal: Exploiting visual contents for point-of-interest recommendation. In Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3–7 April 2017; pp. 391–400.
65. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288. [CrossRef]
66. Sundheim, B.M. Overview of the fourth message understanding evaluation and conference. In Proceedings of the 4th Conference on Message Understanding, McLean, VA, USA, 16–18 June 1992.
67. Järvelin, K.; Kekäläinen, J. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst. (TOIS)* **2002**, *20*, 422–446. [CrossRef]
68. Lobo, J.M.; Jiménez-Valverde, A.; Real, R. AUC: A misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.* **2008**, *17*, 145–151. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.