Expert Systems With Applications 213 (2023) 119167

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Fashion intelligence system: An outfit interpretation utilizing images and rich abstract tags

Ryotaro Shimizu [a,b,*], Yuki Saito [b], Megumi Matsutani [b], Masayuki Goto [c]

[a] Graduate School of Creative Science and Engineering, Waseda University, 3-4-1 Ohkubo, Shinjuku-ku, Tokyo 169-8555, Japan
[b] ZOZO Research, Kioi Tower 22F, 1-3 Kioicho, Chiyoda-ku, Tokyo 102-0094, Japan
[c] School of Creative Science and Engineering, Waseda University, 3-4-1 Ohkubo, Shinjuku-ku, Tokyo 169-8555, Japan

ARTICLE INFO

Keywords:
Fashion intelligence
Fashion interpretation
Visual-semantic embedding
Outfit image
Image retrieval
Attribute activation map

ABSTRACT

In recent years, it has become common for consumers to familiarize themselves with the latest fashion trends through the internet and engage in their own fashion-inspired shopping activities. Therefore, making fashion-inspired shopping and browsing activities (internet surfing in the fashion domain) comfortable is essential because it leads to interactions in the fashion industry. However, fashion is a fuzzy and complex domain that contains many abstract elements, and this ambiguity and complexity can hinder users' deep interest in the fashion industry. Therefore, we define a novel technology and domain called "fashion intelligence" and propose a system based on a visual-semantic embedding method for automatically learning and interpreting fashion and obtaining answers to users' questions. Our proposed method can embed the abundant abstract tag information in the same projective space as outfit images. Mapping of images and tags in a projective space helps search for outfit images using fashion-specific abstract words. In addition, visually estimating the degree of relevance between images and tags helps interpret abstract words. As a result, this research helps decrease fashion-specific ambiguity and complexity and supports the marketing activities and fashion choices of both experts and non-experts.

1. Introduction

In recent years, it has become common for consumers to be acquainted with the fashion outfits of others through social network services (SNSs) and E-commerce (EC) sites and engage in their own fashion-item-purchasing activities. In addition to multi-topic-type SNSs and EC sites, such as Twitter (Twiter, Inc., 2022), Facebook (Meta Platforms, Inc., 2022a), Instagram (Meta Platforms, Inc., 2022b), and Amazon (Amazon.com, Inc., 2022), fashion-specific services, such as WEAR (ZOZO, Inc., 2022a), Snap Fashion (Snap Vision Ltd., 2022), and ZOZOTOWN (ZOZO, Inc., 2022b), are also used extensively. Users' searches for full-body outfit images on these services' applications (apps) or EC sites are crucial activities for refining their fashion-based shopping experiences, with respect to fashion and making purchasing decisions. Therefore, making fashion-based browsing (internet surfing in the fashion domain) comfortable is important because it develops relations in the fashion industry.

However, fashion is a fuzzy and complex domain that contains many abstract elements, and it can be challenging for consumers to understand and interpret fashion, especially for non-experts, because abstract expressions such as "casual", "formal", "street", and "cute" are normally used when explaining fashion. For example, questions such as "what factor makes this outfit casual?", "how casual is this outfit?", and "what kind of outfit would it be if this outfit was made a little more formal?" are difficult to answer, especially for non-experts (and not easy for experts either). This difficulty in interpreting fashion is often encountered on the internet because users cannot rely on human experts. Thus, the ambiguity and complexity of online shopping can make it difficult for many users to engage with new lines of outfits and hinder their deep interest in the fashion industry. Therefore, it is expected that the automatic discovery of answers to such questions will greatly contribute to the fashion industry by making users' online full-body-outfit-image-retrieval and fashion-item-purchasing more comfortable. Simultaneously, it is expected to broaden the perception of users and help in interpreting fashion outfits and encouraging their interests, not always for business purposes. In other words, there is a need for a "fashion intelligence system" that helps to broaden the perceptions of and interests in fashion, rather than a "business intelligence system" that aims to plan a marketing strategy (see Fig. 1).

* Corresponding author at: Graduate School of Creative Science and Engineering, Waseda University, 3-4-1 Ohkubo, Shinjuku-ku, Tokyo 169-8555, Japan.
E-mail addresses: shi3mizu8-r@fuji.waseda.jp (R. Shimizu), yuki.saito@zozo.com (Y. Saito), megumi.matsutani@zozo.com (M. Matsutani), masagoto@waseda.jp (M. Goto).

https://doi.org/10.1016/j.eswa.2022.119167
Received 23 May 2022; Received in revised form 22 August 2022; Accepted 27 October 2022
Available online 7 November 2022

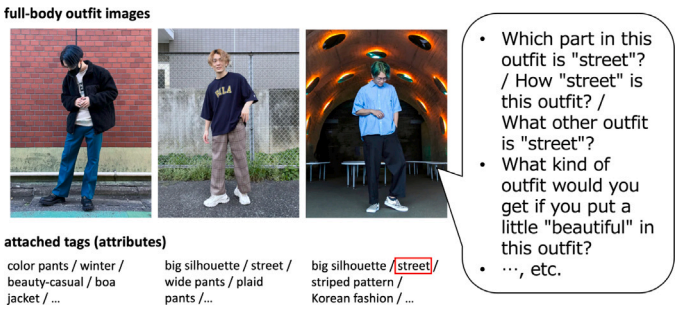

Fig. 1. Image of fashion abstract problem.

Therefore, this study proposes a fashion intelligence system for online outfit coordination based on a visual-semantic embedding method for automatically learning and interpreting fashion and obtaining answers to users' questions. Our proposed method can embed the abundant abstract tag information in the same projective space as the outfit images. Calculating similarities between images and tags in a mapped space helps search outfit images using fashion-specific abstract words. Furthermore, visually estimating the degree of relevance between images and tags makes it possible to interpret abstract words. In this study, we focus on full-body outfit images, which include an extensive range of backgrounds even after detecting a person in a rectangular form, due to the shape of the object (person). Therefore, we introduce foreground-centered learning and background regularization terms that utilize grid weight maps obtained by semantic segmentation. Moreover, because the target datasets contain abstract expressions, such as "casual", "formal", and "cute", which are one of the major reasons why the fashion domain is complex and difficult to interpret, we include sensitive negative sampling based on a latent class model in the proposed system (see Fig. 2).

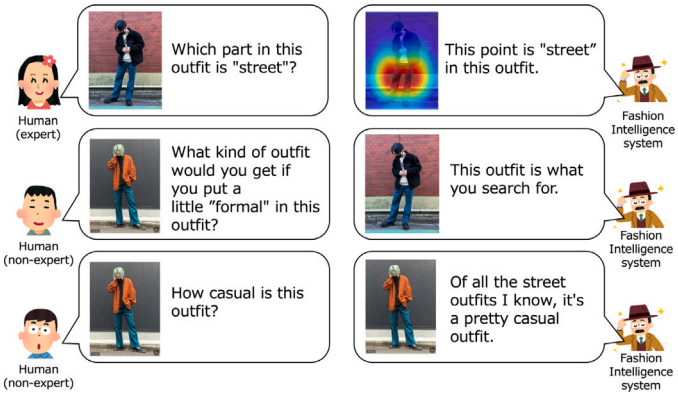

Fig. 2. Image of fashion intelligence system.

Furthermore, results from multifaceted evaluation experiments and analyses using accumulated data from real-world fashion services suggest that the proposed system has valuable applications in the industry. Thus, this research helps decrease fashion-specific ambiguities and complexities (including the dependence on personal judgments) and supports users (experts and non-experts alike) in fashion-related marketing activities and choices.

The main contributions of this work are summarized as follows. (1) We define a novel technology and research area called "fashion intelligence". (2) We propose a fashion intelligence system (based on a visual-semantic embedding method) enabling the embedding of full-body outfit images and rich fashion-specific abstract tags in the same projective space. (3) The proposed system is applied to multiple datasets accumulated from actual fashion-related services, and its efficacy is verified by the results of multifaceted evaluation experiments and analyses. (4) On the basis of the multifaceted experimental results and analyses, we introduce the application of the proposed system to fashion marketing strategy planning.

2. Related work

2.1. Application of artificial intelligence in the fashion industry

Numerous studies have proposed various artificial intelligence (AI) methods for application in the fashion and apparel industries (Giri et al., 2019). For example, several methods have been proposed to improve the efficiency of the production process and supply chain and to increase sales (Riahi et al., 2021), especially in fabric selection and evaluation (Arora & Majumdar, 2022; Kim et al., 2022), AI utilization in the manufacturing process and distribution (Chen et al., 2014; Lee et al., 2016; Lu et al., 2010), and recommendation systems on online shopping sites (Guan et al., 2016). In addition, explainable recommendation and image retrieval studies are also aimed at improving sales. The common point between these existing studies and technologies is that they are related to technologies that are utilized in the flow from manufacturing to sales (supply chain). In other words, they are intended to support business decision making, improve business efficiency, and replace the role of experts. These studies are included in the framework of business intelligence (Liang & Liu, 2018) in a wide sense, and do not go beyond that concept.

By contrast, we propose "fashion intelligence", which is defined as intelligence about fashion. More specifically, it involves diverse knowledge obtained by analyzing information for various fashion-related decisions and a mechanism to obtain such knowledge. Whereas conventional studies have not gone beyond the business intelligence framework, our definition of fashion intelligence envisions the generation and discovery of new knowledge by targeting fashion, which is evaluated and imaged differently depending on people's preferences, values, and cultural backgrounds. In other words, it involves the question of "how to handle ambiguity in human evaluations and judgments", and the handling of this ambiguity is also our primary focus. In this area, we define a "fashion intelligence system" as a mechanism that facilitates the discovery of new knowledge and the creation of new values about fashion by automatically interpreting fashion and collaborating with people (users) through a series of dialogues.

Thus, it is clear that the concepts underlying the fashion intelligence system proposed in this study differ fundamentally from those associated with conventional business intelligence systems, such as those used by companies to make business decisions. Fashion intelligence systems are primarily used by ordinary consumers and experts who support consumers (although they can, of course, be used meaningfully for corporate activities) and provide functions that support the discovery of knowledge and values about fashion. In other words, a fashion intelligence system is not an AI that is introduced from a perspective that is easily evaluated objectively, such as profit or efficiency, or an AI that replaces the role of experts. Instead, it aims to strongly support end users who wish to revel in the fashion world by acquiring knowledge and making new discoveries.

2.2. User support with explainable recommendations in the fashion domain

The concept of explainable recommendations has been proposed and is currently actively being studied as a way to support online user decision-making (Zhang & Chen, 2020). In this research field, beyond the recommendation system's decision to "recommend this item to this user", the reasons "why this item is recommended to this user" are also given by the system. These studies are based on the argument that by showing the reason for recommendations, the opacity in online purchasing with recommender systems is eliminated, and user satisfaction is improved (Chen et al., 2019; Zhang & Chen, 2020; Zhang et al., 2018). Various innovative technologies have been proposed specifically for the fashion field; for example, graph neural network-based methods

have been proposed that output the reasons for a recommendation and its strength based on large amounts of peripheral information (Shimizu et al., 2022; Zhan et al., 2022). Moreover, convolutional neural network (CNN)-based methods exist that indicate where a user is likely to be interested in the image, using item image information and textual information (Chen et al., 2019; Hou et al., 2019; Lin et al., 2020).

In particular, many CNN-based explainable recommendation approaches predict the areas on item images in which users will be interested by the similarity between the embedded representation of the image, the user's embedded representation, and the attention score (Chen et al., 2019; Hou et al., 2019). In other words, decision-making support text such as "this user will prefer the vicinity of the v-neck of this item, so let us recommend this item" can be obtained. However, users often find it difficult intuitively to understand the text "this point on this item is good" while browsing items with such additional explanations. For example, it is rare that a user develops a preference for shirt A by a process where the user first wonders "What do I prefer on this?", and then the area on the item that the user will prefer is highlighted by a recommender system or an expert. In most cases, when a user comes to prefer shirt A, the reason for their preference has already been decided. Therefore, although it can be extremely useful information for product designers and developers, it is doubtful that a user will be persuaded owing to the presented reasons (areas on each item image) (see Fig. 3).

Fig. 3. Image of explainable recommendation (He & McAuley, 2016; Hou et al., 2019).

In contrast, the decision support that users (especially non-experts) truly desire is to interpret abstract and difficult aspects of the fashion domain, such as "where is the casual (formal) point on this fashion item (outfit)?" Especially online, users, experts, and non-experts must interpret fashion on their own, which can be based on their own experiences. Accordingly, they can decide what they need to wear when they go out on the day, what is fashionable, and what is unfashionable. While if the answers are easily obtained, it will increase users' understanding of fashion, motivate them for the clothes, and increase their purchasing desire, these are quite difficult questions for non-experts to answer (and not easy for experts, either). Even then, this support is not the target problem in explainable recommendation technology.

In addition, the structures of many models have become complicated in realizing this complicated multimodal learning with item image information, user information, and, in some cases, textual information within one model. However, it is hard to operate these models constantly, from a practical viewpoint of learning complexity and computational cost. For robust, practical support applications, a simple and easy-to-understand model structure is preferred as far as possible.

2.3. User support with fashion image retrieval

Fashion image retrieval is one of the most active research areas in the field of image processing. Several studies (Chopra et al., 2019; Hussain et al., 2021; Park et al., 2019) provide complete content-based search techniques, searching from query images for other similar images. Moreover, a study has performed cross-domain fashion image retrieval, e.g., search a professional's photo (used in an online store) from the user's photo of the item (Gajic & Baldrich, 2018; Gao et al., 2020; Ibrahimi et al., 2019). Furthermore, a study was conducted on image retrieval techniques to search the daily (realway) clothes similar to the query image of a person on the runway (Vittayakorn et al., 2015). Other studies exist that learn words (attributes) as side-information and utilize them for improving image retrieval accuracy (Corbière et al., 2017). Dong et al. (2021) proposed an image retrieval method that learns attributes (e.g., "lapel design", "neckline design", "collar design") as side-information and focuses on these specific areas. Furthermore, research on techniques for searching for individual items that match fashion items from the query images of the individual items has been actively conducted (Han et al., 2020; Saito et al., 2020).

Moreover, there are studies on techniques for searching images by manipulating images and attributes (Ak et al., 2021; Hou et al., 2021; Shin et al., 2019). These studies have resulted in, for example, if the word "short-sleeve" is added to an image of a long-sleeved blue shirt, an image of a short-sleeved blue shirt will be obtained as a search result. This is a useful technique for improving online image retrieval efficiency and contributes to improving user satisfaction.

Fig. 4. Depiction of the image retrieval system by calculating the similarities between images and words (Ak et al., 2021, 2018; Guo et al., 2019).

However, the focus of most of the existing studies is clothes retrieval, not outfit (full-body) retrieval. Full-body images, including single images with multiple products, have an extensive range and complex background, and targeting them is a challenging task (Zhao et al., 2019). In previous studies, the words that were added or removed from images were specific (non-abstract) words (such as "short-sleeve" and "black"). This contributes to improving the efficiency of the search function, but not to support users' understanding and interpretation of fashion. For instance, in Fig. 4, if the word "short-sleeve" is added to a long-sleeved blue shirt, it is easy, even for non-expert users, to answer "short-sleeved blue shirt". Conversely, because of its difficulty, what users (especially non-experts) truly want to be supported on is to obtain the answers to abstract and somewhat complex questions, such as "what kind of fashion outfit would you recommend if I want to make this fashion outfit a little more casual (formal)?" Obtaining answers to these questions is very helpful in deciding what clothes users will wear on the day and what items they will purchase. In Banerjee et al. (2022), item search is realized by considering only eight types of selectable abstract category information, such as "party", "outdoor", and "summer". However, this previous study only dealt with a limited number (eight) of information categories. In this study, we focus on a system that includes a search function for full-body outfit images by utilizing more than 1000 types of tags, and a visual interpretation function via an attribute activation map (AAM). As we mentioned, it is extremely difficult for non-experts to play this role on their own (and it is not easy for experts either).

2.4. Fashion concept discovery

Visual-semantic embeddings (VSE) have been widely researched for image-caption and image-description retrieval (Faghri et al., 2018; Wu et al., 2019), visual question-answering (Ren et al., 2015b), hashing task (Frome et al., 2013), person re-identification (Gao et al., 2021) and image descriptions generation (Karpathy & Fei-Fei, 2015). In the fashion domain, VSE has been used for text-based and individual clothes image retrieval (Chen & Bazzani, 2020; Tautkute et al., 2019), and learning outfit compatibility (individual outfit item matching) (Wang et al., 2021). In this study, we focus on attribute-based full-body outfit image retrieval.

VSE, included in Han et al. (2017) as a discovery concept, is a method of embedding in the same projective space a fashion item image and a specific word contained in the item description. As a result, VSE in Han et al. (2017) makes it possible to search for fashion images by calculating similarities between individual item images and simple words, and to find specific points on the target item image that are highly related to the word (thereby creating an AAM). These multiple functions constitute the advantage of VSE. The main task of Han et al. (2017) is the automatic discovery of concepts such as "size", "color", and "shoulder shape". For example, words such as maxi and mini belong to the concept of "size", and words such as "one-shoulder" and "no-sleeves" refer to "shoulder shape". By utilizing the embedded representations of images and words acquired in this concept discovery process, it is possible to calculate similarities between images and words and discover parts of images with high relevance to related words. Compared to other attribute-based individual clothes image retrieval methods, the strength of VSE in Han et al. (2017) is that it not only executes image-retrieval tasks but also creates an AAM due to the simplicity of its model structure.

This fashion concept discovery process consists of three processes: "visual-semantic embedding training", "spatially-aware concept discovery", and "concept subspace learning", with image and word embedding being mainly performed in the "visual-semantic embedding training" step. The strength of this research includes not only the results of searching and extraction of attention parts but also the relatively simple structure of the VSE model. Specifically, only (1) the embedded representation of each word, (2) the transform matrix for converting the output obtained from a CNN into the embedded representation of the image, and (3) the parameters included in the CNN itself are targeted for update. The simplicity of the model contributes to the ease of implementation and management, the ease of updating parameters, and the realization of short computational times. In other words, it leads to the realization of model application to actual business.

However, the main task of this previous research was to discover concepts automatically, and the target image data was an image of an individual fashion item (similar to other studies on explainable recommendations and fashion image retrieval). Moreover, because the target image data are related to products for sale, it is usually a beautiful image taken by an expert photographer with a simple background. Besides, due to the shape of the item, most rectangles, after detecting the area where the fashion item appears, contain almost no background. Furthermore, the target word data is simple (as in other studies on explainable recommendation and fashion image retrieval) and incorporates concrete words, such as "maxi", "mini", "one-shoulder", and "no-sleeves". As a result, the image retrieval function can realize simple search tasks that even non-experts can easily answer, as in other studies. Therefore, the task of extracting attention parts on images that are highly related to words is simple enough that even non-experts can easily answer, such as "no-sleeves in this item is here" (see Fig. 5).

Fig. 5. Contributions of the previous visual-semantic embedding method in fashion concept discovery (Han et al., 2017; Liu et al., 2016b).

In contrast, the target data in this study are images that include full-body outfits. Since most general users post those images on SNS, there can be a variety of backgrounds. In addition, since full-body outfits are targeted, the rectangle after object detection always includes an extensive range and variety of backgrounds, depending on the pose and shape of the person. However, unlike in previous studies, tag data added for non-commercial purposes to convey images of outfits are targeted. Therefore, many abstract expressions are also included, in addition to specific tags (see Fig. 6).

VSE model	image	attribute information	sample
in concept discovery	- for each item - for sells / professional post - with simple and small range of background	- words (simple in the item description) - easy-to-understand expressions given by professionals for commercial purposes	 multicolor / ¾-sleeve / square neck / dress
in this study	- full-body outfit - not for sells / mostly non-professional post - with complex and wide range of background	- tags (various variety attached to images) - including an abstract expression given by a general user for non-commercial purposes	 big silhouette / street / striped pattern / over-size / wide pants / Korean fashion / 2021 / shirt style / Chongjin gorge / spring-style

Fig. 6. Comparison with previous visual-semantic embedding in fashion concept discovery (Han et al., 2017; Liu et al., 2016b).

In fashion concept discovery, individual models (processes) were learned for each category of fashion items, such as "dress", "top", and "pants". In contrast, in our proposal, a set (full-body outfit) including all of these items is collectively learned with one model. This is greatly beneficial because the number of models to be managed is reduced in the actual application of the models to businesses. However, considerable ingenuity is required to learn all the information contained in full-body outfit images and abstract tags all at once and efficiently.

2.5. Motivation for this study

In summary, the motivations for this study, based on the review of related works, are given as follows:

- To accurately quantify full-body outfit images and related abundant tags, including abstract tags, in one comprehensive model.
- To achieve image retrieval that is difficult for users (both experts and non-experts) by calculating full-body outfit images and tags similarities from fashion-specific abstract expressions.
- To clarify the parts in the image that are highly related to each tag (thereby creating an AAM) and make fashion-specific abstract tags visually interpretable.

Based on the above motivations, this study proposes a visual-semantic embedding model that enables the interpretation of fashion attribute information (including abstract aspects) and decreases dependencies on human experiences or knowledge in the fashion field. We present an application of the model that broadens users' perceptions and supports their online purchasing and browsing activities. We propose a novel fashion intelligence system that makes it possible to perform tasks that are difficult for users (both experts and non-experts), which were not targeted in conventional studies on explainable recommendations and fashion image retrieval.

3. Proposed system

3.1. Problem definition

In this study, we propose a system that supports the visual interpretation of fashion information by embedding full-body outfit images and tag information in the same projective space to quantify the information. By completing this research, users (especially non-experts) will be able to understand fashion on their own, which will lead to increased motivation for fashion and purchasing desire. In other words, it is expected to improve the usability of all SNS and EC sites handling fashion items, including fashion-related posts. Furthermore,

it is expected to broaden the understanding and perception of users (including experts), and help in planning marketing strategy.

The image data targeted in this study are full-body fashion pictures of single subjects (persons) taken in front of a wide variety of backgrounds. Even if the smallest rectangle reflecting only one person is extracted from these images, a wide and complex background is included in the rectangle due to the problem of the shape of the person. In particular, areas such as the left and right areas of the neck, between the legs, and the left and right areas of the legs are all part of the background. In addition, background patterns vary broadly (especially in SNS data), and many cases exist where images other than the subject are not one-tone backgrounds. Therefore, if the entire contents of a rectangle are learned as they are, the background becomes noise, and sensitive learning cannot be realized.

Multiple tags are attached to each image as attribute information from post users. This tag information includes not only specific and simple tags (e.g., "denim", "skirt", and "t-shirt") but also abstract tags (e.g., "spring-style", "formal", "casual", and "office-casual"). Regarding their characteristics, specific tags, once attached, are always correct regardless of the sensibility of post users. Conversely, the characteristics of abstract tags, whether attached or not, are uncertain and depend on the sensibility of post users. For example, in the sensibility of post user A, if image A is completely "casual", it can be correct that the tag "casual" is attached. In contrast, if post user B feels that image A is only partially casual, the "casual" tag may not be attached by this user. For post user C, if the expression "holiday-style" seems more appropriate than "casual", "holiday-style" would be attached rather than "casual". In this way, a target image includes not only specific tags but also abstract tags. The abstract expressions are one of the major reasons why users (especially non-experts) find the fashion domain difficult.

3.2. Visual-semantic embedding

In the proposed system, it is possible to embed specific and abstract tags in the same space as full-body outfit images. To handle a complex and extensive background included in a rectangle and abundant tag information, including abstract expressions, the proposed system introduces some inventiveness to enable sensitive learning. The entire model structure is shown in Fig. 7. In the following sections, a systematic description of the proposed model is provided.

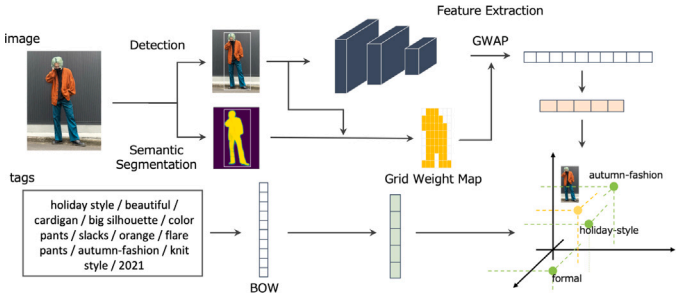

Fig. 7. Structure of a prototype of our visual-semantic embedding model proposal.

3.2.1. Person detection

Using a trained object detection model, the smallest rectangle reflecting only one person is extracted from an entire posted image related to a full-body outfit image. Single Shot MultiBox Detector (SSD) (Liu et al., 2016a) and Faster-RCNN (Ren et al., 2015a) are generally used for person detection (we chose SSD, but this is not the main point of this research). By utilizing these models, the smallest rectangle, where only one person is reflected, and the background scraped off as much as possible, is obtained.

3.2.2. Grid weight map based on semantic segmentation

There are several studies on person foreground segmentation (Liang et al., 2022; Liu et al., 2021). From the perspective of ease of handling, in this study, semantic segmentation is performed by using a trained Fully Convolutional Network (FCN) (Long et al., 2015). During semantic segmentation, we determine the probability of whether or not a cell corresponds to a "person". This output is obtained for each pixel. Based on the value calculated for each pixel, an average value is determined for each grid, where grid refers to an area obtained when a target image is divided vertically into I and horizontally into J. If the average value of the probabilities calculated for each grid is equal to or greater than a threshold α, the weight corresponding to the (i, j)th grid is set to 1.0, and the others are set to 0.0. A grid weight map corresponding to the foreground is obtained by normalizing the weight obtained through this procedure. Moreover, a grid weight map corresponding to the background is obtained in the same step with the 1.0 and 0.0 settings swapped.

3.2.3. Loss function

By transferring the trained CNN and re-training with a target dataset, an extractor for acquiring image features is obtained. Several types of CNN models can be used as extractors; however, the complexity of the dataset and the allowable computational cost are the determining factors. In this research, the GoogleNet Inception V3 model (Szegedy et al., 2016) is used as an extractor on the basis of the original study (Han et al., 2017).

Global weighted average pooling (GWAP) (Qiu, 2018) is performed using the image features extracted from the CNN and the grid weight map. This implies replacing the global average pooling (GAP) layer remaining after the final convolutional layer of a general CNN with the GWAP layer. To extract high-quality image features, the idea of delicately handling the features for each grid is effective (Xiang et al., 2022). This GWAP operation significantly reduces the problem of capturing noise caused by the GAP layer, which is ignored by numerous related methods (Han et al., 2017; Park et al., 2019). By assuming K to be the dimension of the embedded space, the embedded representation of image $\mathbf{x}_\mathrm{f} \in \mathbb{R}^K$ in the foreground can be calculated as follows:

$$\mathbf{x}_\mathrm{f} = \mathbf{W}_\mathrm{I}\mathbf{f}_\mathrm{f}, \tag{1}$$

$$\mathbf{f}_\mathrm{f} = \sum_{i,j} g_{(i,j)}\mathbf{q}_{(i,j)}, \tag{2}$$

where $\mathbf{W}_\mathrm{I} \in \mathbb{R}^{K\times D}$ is the transform matrix for converting the image feature vector extracted from the CNN to image embedded representation (D is the number of dimensions of the final convolutional layer of the CNN), $\mathbf{f}_\mathrm{f} \in \mathbb{R}^D$ is a weighted image feature vector for the foreground acquired via the grid weight map obtained from the GWAP layer, $g_{(i,j)}$ is a grid weight for the (i, j)th grid corresponding to the foreground, and $\mathbf{q}_{(i,j)} \in \mathbb{R}^D$ is the foreground image feature vector for the (i, j)th grid obtained from the final convolution layer of the CNN. By contrast, we calculate the embedded representation of the image in the background $\mathbf{x}_\mathrm{b} \in \mathbb{R}^K$ as follows:

$$\mathbf{x}_\mathrm{b} = \mathbf{W}_\mathrm{I}\mathbf{f}_\mathrm{b}, \tag{3}$$

$$\mathbf{f}_\mathrm{b} = \sum_{i,j} g'_{(i,j)}\mathbf{q}_{(i,j)}, \tag{4}$$

where $\mathbf{f}_\mathrm{b} \in \mathbb{R}^D$ is a weighted image feature vector for the background acquired via the grid weight map obtained from the GWAP layer and $g'_{(i,j)}$ is a grid weight for the (i, j)th grid corresponding to the background.

In addition, tags attached to datasets can vary in frequency. For example, seasonal tags such as "spring-style" are frequently added to the entire set of images posted in spring. By contrast, tags such as "beret" are attached only to the images of people wearing them; therefore, they are rarely attached to a whole dataset. Typically, infrequent tags are likely to be important factors that characterize an image (i.e., differentiate from other images). To use this property in this study,

we obtained the embedded representation of the tags $\mathbf{v} \in \mathbb{R}^K$ attached to a target image using Eq. (5) by weighting the tags so that the less frequent tags have a greater effect.

$$\mathbf{v} = \sum_t w_t \mathbf{a}_t, \tag{5}$$

$$\mathbf{a}_t = \mathbf{W}_\mathrm{T} \mathbf{e}_t, \tag{6}$$

$$w_t = \frac{1/\log_e(N_t+1)}{\sum_t 1/\log_e(N_t+1)}, \tag{7}$$

where $\mathbf{a}_t \in \mathbb{R}^K$ is an embedded representation for the tth single tag among the tags attached to the target image, $\mathbf{W}_\mathrm{T} \in \mathbb{R}^{K \times C}$ is a transform matrix for converting tag feature vectors with tag embedded representations (C is the number of tags in the entire dataset), $\mathbf{e}_t \in \mathbb{R}^C$ is an one-hot vector for the tth single tag, and N_t indicates the total attachment frequency of the tth attached tag to the target image in the entire batch dataset.

Using these features, we realize an operation that focuses only on the foreground (i.e., foreground-centered learning) and is robust to the noise of the wide and complex background (background regularization). As a result, the loss function is defined as follows:

$$\begin{aligned}\mathcal{L}(\Theta) = & \sum \max\left(0, m - s(\mathbf{x}_\mathrm{f}^+, \mathbf{v}^+) + s(\mathbf{x}_\mathrm{f}^+, \mathbf{v}^-)\right) \\ & + \sum \max\left(0, m - s(\mathbf{v}^+, \mathbf{x}_\mathrm{f}^+) + s(\mathbf{v}^-, \mathbf{x}_\mathrm{f}^+)\right) \\ & + \beta \sum \max\left(0, m + s(\mathbf{x}_\mathrm{b}^+, \mathbf{v}^+)\right), \end{aligned} \tag{8}$$

where $\Theta = \{\mathbf{W}_\mathrm{I}, \mathbf{W}_\mathrm{T}, \mathbf{V}\}$ is a set of parameters to be optimized, $\mathbf{V}$ is a parameter set contained in CNN, $s(\mathbf{x}, \mathbf{y})$ indicates the cosine similarity between vectors $\mathbf{x}$ and $\mathbf{y}$, and $\beta > 0$ is a positive hyperparameter to adjust the importance of the background regularization term. Furthermore, the superscript sign + of A^+ indicates that A is a variable related to the positive sample, and $-$of A^- indicates that A is a variable related to the negative sample.

By updating each parameter to optimize the loss function in Eq. (8), we obtain a transform matrix for mapping the embedded representation corresponding to each tag and the image features obtained from the CNN in the same space as the tag.

3.2.4. Negative sampling based on the latent class model

Among the optimization steps, for each set of an image and tags (positive samples), a set of an image and tags (negative) is sampled (negative sampling). Here, as in many other studies, if sampling is performed with equal probability (i.e., completely random) from all the sets of images and tags, it becomes difficult to realize efficient and sensitive optimization. Devising negative sampling to fit the target problem contributes to the realization of efficient and effective learning (Daghaghi et al., 2021; Moon et al., 2019). For example, many studies in areas such as natural language processing (Chen et al., 2018; Stergiou et al., 2017) and graph convolution network (Yang et al., 2020; Zhang et al., 2019) have devised negative sampling mainly with the aim to reduce the computation amount. In this study, we additionally improve negative sampling for learning abstract tags, in terms of effectiveness, efficiency, and sensitivity.

In this study, the first problem is that, as shown on the left in Fig. 8, candidates including duplicated tags with positive tags and candidates not including them are treated equally. For instance, if the tag set containing "autumn-fashion", "street-style", and "casual" is a positive sample, then the other set, set 1 = ("winter-fashion", "formal", "black-tone", and "suits") and set 2 = ("casual", "spring-fashion", "holiday", and "jeans"), is a candidate for negative sampling. In this case, set 1 should be sampled because there is a duplication of tags between a positive sample and set 2. Therefore, in the proposed system, the tag sets that include a duplicated tag with the positive sample tags are excluded from being candidates for negative sampling.s

If, as shown on the right in Fig. 8, the set containing "autumn-fashion", "street-style", and "casual" is a positive sample, and set 1 = ("winter-fashion", "formal", "black-tone", and "suits"), and set 2 =

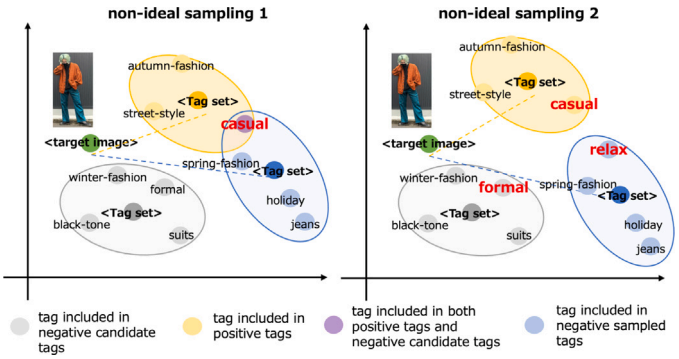

Fig. 8. Image of a negative sampling.

("relax", "spring-fashion", "holiday", and "jeans") are the candidates for negative sampling; here, both sets have no duplication of tags with the positive tag set. However, since set 2 contains similar tags (e.g., "casual" and "relax"), set 1 should be sampled compared to set 2. This problem often happens in the situation that target tags data include abstract tags (in this case, the post user in the set 2 chose the tag "casual" instead of "relax"). Therefore, in the proposed model, data related to tags are applied to a latent class model (Blei et al., 2003; Hofmann, 1999), which is a powerful method for extracting necessary information while compressing complex data (Goto, 2019; Kim et al., 2020), and the topic distribution related to the tags is acquired for each candidate. Moreover, negative sampling is performed based on the distance of distributions (Kullback–Leibler divergence) between the positive sample and all negative sample candidates.

By the two proposed methods, the following sample probability p_i is given to the ith sampling candidate. Eq. (9) is a definition formula of probability $p_i^{(1)}$ that excludes tag sets duplicate with a positive sample, and Eq. (10) is a definition formula of probability $p_i^{(2)}$ that considers the distribution of tags similarity between a positive sample and each candidate for negative sampling.

$$p_i^{(1)} = \begin{cases} 0.0 & (\text{if } \mathbf{e}^{+\mathsf{T}} \mathbf{e}_i > 0), \\ 1.0 & (\text{if } \mathbf{e}^{+\mathsf{T}} \mathbf{e}_i = 0), \end{cases} \tag{9}$$

$$p_i^{(2)} = \begin{cases} 0.0 & (\text{if } \mathbf{e}^{+\mathsf{T}} \mathbf{e}_i > 0), \\ \mathrm{D}_\mathrm{KL}\left(q_\mathrm{lda}(Z|T^+), q_\mathrm{lda}(Z|T_i)\right) & (\text{if } \mathbf{e}^{+\mathsf{T}} \mathbf{e}_i = 0), \end{cases} \tag{10}$$

where $\mathbf{e}^+ \in \mathbb{R}^C$ is a bag-of-words representation of the positive tag set, $\mathbf{e}_i \in \mathbb{R}^C$ is a bag-of-words representation for the tag set of the ith candidate, $\mathrm{D}_\mathrm{KL}(A, B)$ expresses Kullback–Leibler divergence between A and B, $q_\mathrm{lda}(Z|T^+)$ is a topic distribution for the positive sample, and $q_\mathrm{lda}(Z|T_i)$ is a topic distribution for the ith candidate obtained from Latent Dirichlet Allocation (LDA) based on the attached tags T_i. Furthermore, the probability p_i is normalized and used for sampling.

As a result, for the negative tag set, the tag set that does not have duplicate tags with the positive tags and has a distant meaning from the attached tags is sampled. In this way, the tag information is effectively used to realize more efficient and sensitive optimization.

3.3. Creating an attribute activation map

The degree of relevance between the arbitrary single tag embedded representation $\mathbf{a} \in \mathbb{R}^K$ and the arbitrary image embedded representation $\mathbf{x} \in \mathbb{R}^K$ is expressed as follows:

$$\begin{aligned} s(\mathbf{a}, \mathbf{x}) &= \sum_k a_k x_k \\ &= \sum_k a_k \sum_d W_{I_{kd}} f_{f_d} \\ &= \sum_k a_k \sum_d W_{I_{kd}} \sum_{i,j} g_{(i,j)} q_{(i,j)_d} \end{aligned}$$

$$= \sum_{i,j} g_{(i,j)} \sum_{k} a_k \sum_{d} W_{I_{kd}} q_{(i,j)_d}, \tag{11}$$

where a_k and x_k are the values in the kth dimension of tag embedded representation $\mathbf{a}$ and the image embedded representation $\mathbf{x}$, respectively. In addition, $W_{I_{kd}}$, f_{f_d} and $q_{(i,j)_d}$ are the values in the dth dimension of the kth vector $\mathbf{W}_{\mathrm{I}_k} \in \mathbb{R}^D$ in the transform matrix $\mathbf{W}_{\mathrm{I}}$ for the image, the output of the GWAP layer $\mathbf{f}_{\mathrm{f}} \in \mathbb{R}^D$ for the foreground, and the (i,j)th grid obtained from the final convolution layer of the CNN $\mathbf{q}_{(i,j)} \in \mathbb{R}^D$, respectively.

Based on Eq. (11), the degree of relevance between the tag and the (i,j)th grid on the image is expressed as follows:

$$M(i,j) = \sum_{k} a_k \sum_{d} W_{I_{kd}} q_{(i,j)_d}, \tag{12}$$

and a set of $M(i,j)$ for all grids is called attribute activation map (AAM) between the image and the tag.

By using AAM, it is possible to visually understand where the target tag is related to the target image. For example, if a user wants to know "what is the casual point on a full-body outfit image A?", it can be visually grasped by measuring the degree of relevance between image A and the tag "casual".

3.4. Image retrieval

Image retrieval is realized by adding (positive) and subtracting (negative) a tag to the query image, and is given by the following Eq. (13).

$$\mathbf{x_o} = \underset{\mathbf{x}}{\operatorname{argmax}}\ s\left(\sum_{i,j} \delta_{q_{(i,j)}} \left(\mathbf{W}_{\mathrm{I}}\mathbf{q}_{\mathrm{q}_{(i,j)}} + \mathbf{a}_{\mathrm{p}} - \mathbf{a}_{\mathrm{n}}\right), \mathbf{x}\right), \tag{13}$$

$$\delta_{q_{(i,j)}} = \begin{cases} 1.0 & (\text{if } M_n^q(i,j) g_{(i,j)} > 0), \\ 0.0 & (\text{if } M_n^q(i,j) g_{(i,j)} \le 0), \end{cases} \tag{14}$$

where $\mathbf{x_o} \in \mathbb{R}^K$ is the image embedded representation of the search result, $\delta_{q_{(i,j)}}$ is a binary output indicator of whether the (i,j)th grid in the query image is the computed operation target, $\mathbf{q}_{\mathrm{q}_{(i,j)}} \in \mathbb{R}^D$ is the output for the (i,j)th grid of the query image obtained from the final convolution layer of the CNN, $\mathbf{a}_{\mathrm{p}} \in \mathbb{R}^K$ is the embedded representation of the positive tag, $\mathbf{a}_{\mathrm{n}} \in \mathbb{R}^K$ is the embedded representation of the negative tag, $M_n^q(i,j)$ represents the degree of relevance between the (i,j)th grid on query image and the negative tag, and $g_{q_{(i,j)}}$ is the weight of the (i,j)th grid in the query image.

By this calculation, a negative tag is subtracted only in an area highly related to the specific grid on the query image, and a positive tag is added to that area instead of a negative tag. The image of this operation is shown in Fig. 9.

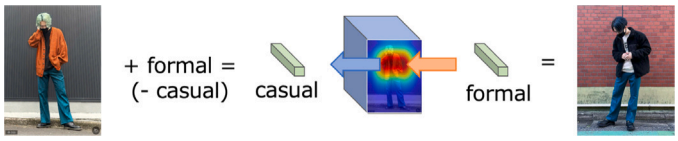

Fig. 9. Image of calculation between tag and image for retrieval.

By this operation focusing on the individual grid, it is possible to acquire an image in which changes have been made only to the parts that should be changed on the image. This makes it possible to perform calculations that do not make excessive changes to the entire image in the full-body outfit image of a person coordinating multiple clothes.

4. Experimental evaluation

To confirm its effectiveness, the proposed system was applied to actual posted full-body outfit image data and the tags information attached to each image accumulated in WEAR (ZOZO, Inc., 2022a), a fashion-specific SNS. An evaluation experiment was conducted to confirm the effectiveness of the proposed system in terms of (1) loss transition, (2) similarity between the embedded representation of the image and the attached tag, (3) relevance score between foreground/background and each tag (AAM), and (4) evaluation questionnaire for image retrieval results for both experts and non-experts.

4.1. Experimental settings

Table 1 summarizes the two types of datasets used in the experiment.

Table 1
Summary of dataset features.

	dataset-1	dataset-2
Number of snaps	18,050	15,740
Number of unique tag	1,753	1,104
Gender of the subject	women	women
Background	Contains considerable noise	Contains some noise
Situation assumption	SNS	SNS/EC-site
Sample		

With dataset-1, we verified whether learning could proceed robustly for data whose background is extremely complicated (and extensive range). With dataset-2, we verify whether the proposed system is effective even for data with a relatively simple background (but extensive range). Dataset-1 is an image dataset posted by general users, and dataset-2 is an image dataset posted by professional users. In both cases, only one person is shown in an image, and the extensive background is included.

The embedded representation dimension included in the VSE model is set to 64, the learning rate is 0.001, the number of epochs is 50, the batch size is 32, threshold α for the grid weight map is 0.1, the parameters for regularization terms β is 0.1, and margin m is set to 0.2. Furthermore, VGG16 (Liu & Deng, 2015) (a CNN with a relatively large number of parameters) and GoogleNet Inception V3 (a CNN with a relatively small number of parameters) pre-trained on ImageNet (Deng et al., 2009) are respectively used as extractor for assessing the impact of CNN. As preprocessing, SSD (extractor: MobileNet V2 (Sandler et al., 2018)) trained on Open Images (Kuznetsova et al., 2020) is used for object detection, and FCN (extractor: ResNet) trained on COCO (Lin et al., 2014) is used for semantic segmentation. The number of LDA topics used for negative sampling was set to 8 for dataset-1 and 7 for dataset-2 by prior experiments.

4.2. Loss transition

By checking the loss transition for each epoch, we confirm the effectiveness of (1) foreground-centered learning by the grid weight map and (2) negative sampling according to the duplicated relation and the prior distribution that reflects the co-occurrence of the attached tags. These two methods are commonly proposed for efficient and sensitive learning. First, to confirm the individual improvement effects in detail, a comparison of the previous method (Han et al., 2017) and proposed methods that changed only the grid weight map (foreground-centered learning) is shown in Figs. 10–13. Here, GAP represents the previous method, and GWAP(threshold) represents the completely proposed method. Furthermore, as a comparison method, GWAP(original) that excludes the threshold part of the grid weight map and uses each grid's probability of the foreground softly, and GWAP(center) that learns only the 3/4 range of the center all images are applied.

From Figs. 10 and 11, for datasets with complex and wide range backgrounds (dataset-1), learning hardly progresses with the previous method (VGG16), whereas with the proposed method and the comparison methods (including foreground-centered learning), it can be seen that learning progresses relatively smoothly for each epoch. Moreover,

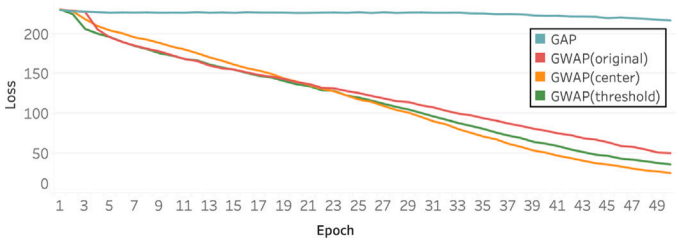

Fig. 10. Transition of loss for each epoch (about grid weight map/dataset-1, VGG16).

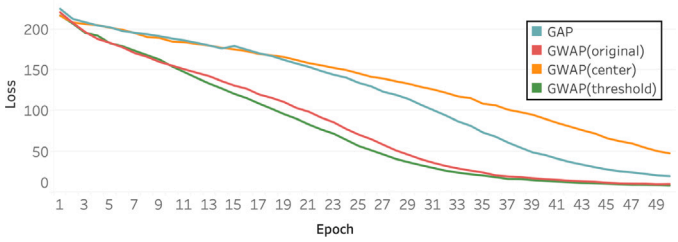

Fig. 11. Transition of loss for each epoch (about grid weight map/dataset-1, Inception V3).

in the case of Inception V3, it can also be seen that learning progresses more efficiently in the proposed method and the GWAP(original) (including foreground-centered learning and given the position of a foreground accurately) than in the previous method. From these results, it can be suggested that the proposal is a robust method for background noise by foreground-centered learning with a grid weight map. In contrast, it is considered that the reason why GWAP(center) did not proceed smoothly in dataset-1 is that the images in dataset-1 include various human poses, and in many images, the foreground is not always in the 3/4 range of the center.

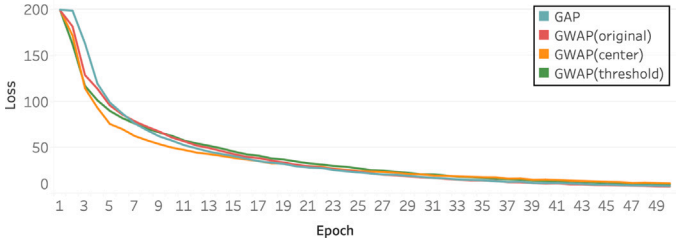

Fig. 12. Transition of loss for each epoch (about grid weight map/dataset-2, VGG16).

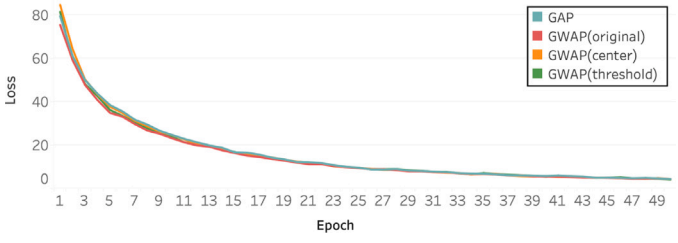

Fig. 13. Transition of loss for each epoch (about grid weight map/dataset-2, Inception V3).

From Figs. 12 and 13, we could not confirm the remarkable improvement effect of the loss transition by the grid weight map for the dataset with a relatively simple background. Therefore, in the next section, we will confirm the validity of the proposal from another angle by confirming the acquired embedded representation in more detail with respect to dataset-2. All the following experiments will also show the experimental results of dataset-2 and Inception V3, which are relatively proper to learn, and the effectiveness could not be clearly confirmed only by the loss transition.

Next, to verify whether the proposed negative sampling is effective, the result of the method in which only the negative sampling part is changed (no grid weight map) from the previous method is shown in Fig. 14. Here, random represents a method of completely randomly selecting negative sampling (previous method), p1 represents a method for performing negative sampling using Eq. (9), and p2 represents a method using Eq. (10).

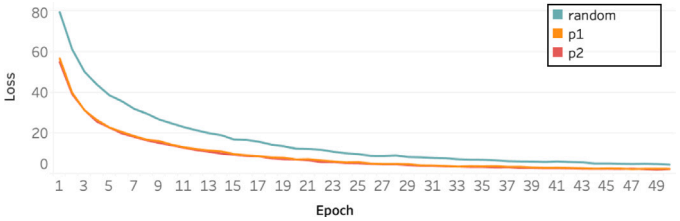

Fig. 14. Transition of loss for each epoch (about negative sampling/dataset-2, Inception V3).

From Fig. 14, it was clearly confirmed that efficient learning was realized by the changes made further to the negative sampling. This result suggests that the proposal can select appropriate negative sampling compared to the previous method. It can be seen that effective and sensitive learning is indispensable for the embedding of the full-body outfit image and tags, including abstract tags targeted in this study; thus, the rules for excluding duplicate tags and the topic distribution of attached tags by LDA contribute.

4.3. Detailed verification of the effectiveness of the grid weight map

From Figs. 12 and 13, for a dataset with a relatively simple background, the effect of foreground-centered learning by the grid weight map could not be confirmed clearly only by the viewpoint for the transition of loss. Therefore, here, we confirm the validity of the proposal by 2 types of detailed analyses of the obtained embedded representation from another angle.

4.3.1. Similarity between images and attached tags

The effectiveness of the proposed method (the effect of foreground-centered learning) using the similarity of embedded representation for the image and the attached tags obtained from the proposed method is verified. If it is accurately embedded in the space of the mapped destination, the image would be embedded around the attached tag. Based on this assumption, in this experiment, the images attached to each tag are defined as the ground truth, and the negative images are obtained randomly from other images. The number of negative images is 10 times the number of ground truth. Under this circumstance, top-K $= 5, 10, 15$ images that are particularly similar to the embedded representation of each tag are acquired and evaluated how many ground truths are included in them. The evaluation indicators are Precision and NDCG. The experiment is repeated 30 times, and the average of the results is shown in Table 2. Moreover, the t-test was performed between the results of the GAP and other methods, and the significance level was set to 5%.

Table 2
Summary of evaluation values for top-K images selected using similarity for each tag.

	Precision			NDCG		
top-K	5	10	15	5	10	15
GAP	0.581	0.571	0.555	0.532	0.551	0.549
GWAP (original)	0.602**	0.607**	0.603**	0.554**	0.581**	0.587**
GWAP (center)	**0.700****	**0.663****	**0.653****	**0.637****	**0.645****	**0.646****
GWAP (threshold)	0.659**	0.631**	0.621**	0.604**	0.614**	0.615**

From Table 2, it can be seen that GWAP exceeds GAP on all indicators, and learning by focusing on the foreground is an effective method for accurately embedding the foreground and tags in the same space.

In particular, GWAP (center) has the highest accuracy. In dataset-2, a foreground is included in the central 3/4 area of most images (the background is not included at all). It is considered that this result is caused by the fact that learning proceeds without problems even if the arm or legs and the edge of the clothes are completely ignored. This result further suggests that accurate learning can be achieved by completely ignoring the background rather than considering the edges of the clothes instead of including some background. However, since it is not desirable to ignore the area at the edge of the clothes for fashion interpretation, GWAP(threshold) is considered to be the most effective method, and GWAP(threshold) is used comprehensively in the subsequent experiments.

4.3.2. Attribute activation map

Based on the results so far, the effectiveness of foreground-centered learning using the grid weight map was observed. In this section, the effectiveness of background regularization, which is the third main change proposed in this study, is verified. Unlike the previous two (foreground-centered learning/negative sampling), background regularization is a method proposed to reduce the relevance of the background and the tag. By decreasing the degree of relevance between the background and the tag, the probability that the background will be colored in the colored AAM will decrease, and improved user satisfaction will be expected.

Specifically, in this experiment, the degree of relevance between all the images contained in the data and all the individual tags attached to them is calculated. Then, the degrees of relevance for each of the foreground and the background area are observed individually. Since the background of dataset-2 is relatively simple, there must always be no relevance between the tag and the background. In other words, if the embedded representations of the image and the tag are learned properly, the embedded representations that the background and the tag have a low relevance and the foreground and tag have a high relevance should be acquired. First, Table 3 shows the result of the average of the relevance between each grid of foreground/background and the attached tags. Here, GAP represents a method that does not use the grid weight map at all, GWAP represents a method in which only foreground-centered learning is performed, and GWAP+reg represents a method in which background regularization is combined with foreground-centered learning. Moreover, the t-test was performed for the results of foreground/background, respectively, between the results of GAP and other methods, and the significance level was set to 5%.

Table 3
Overall average of the relationship between tags and foreground, background.

	Foreground	Background
GAP	0.060	−0.127
GWAP(threshold)	0.063**	−0.131**
GWAP(threshold)+reg	**0.089****	**−0.186****

From Table 3, unless background regularization is added, it can be seen that even if learning focusing on the foreground is performed, the degree of relevance will be as high as that of the previous method. In contrast, with background regularization, a higher relevance is observed between the grids of the foreground and the attached tags, and a lower relevance is observed between the grids of the background and the attached tags. This result suggests that the contribution of the regularization of the background with the grid weight map makes it possible to give priority to the foreground over the background and embed the image and the attached tag closer to each other.

Fig. 15 (Fig. 16) shows the result of calculating the degree of relevance between the tag and each grid included in the foreground (background) and comparing GWAP+reg and GWAP. Each cell represents a grid, and the numbers in it represent the ratio that the grid is the foreground (background). It is orange if the GWAP+reg shows excellent results with the 5% level of significance, it is blue if the GWAP shows excellent results with 5% significance, and it is white if a difference is not statistically significant at the 5% level. In the foreground, each grid should have high relevance to the attached tags. On the contrary, in the background, each grid should have low relevance to the attached tags.

0.33%	12.14%	63.33%	93.68%	92.46%	61.72%	15.70%	0.83%
3.99%	43.68%	82.31%	97.16%	97.40%	89.82%	72.82%	32.00%
38.34%	73.91%	92.39%	99.38%	99.42%	96.36%	88.31%	66.07%
34.52%	76.40%	95.10%	99.55%	99.61%	99.55%	99.61%	54.93%
18.56%	76.78%	97.11%	99.48%	99.68%	98.31%	85.33%	40.35%
16.15%	65.16%	93.92%	98.42%	99.11%	94.60%	69.91%	31.96%
10.31%	52.37%	87.51%	96.58%	97.54%	91.48%	58.65%	22.69%
21.22%	49.69%	76.59%	89.17%	93.79%	76.10%	41.14%	20.24%

Fig. 15. T-test results with and without background regularization for individual grid average of the relevance between tags and foreground.

99.67%	87.86%	36.67%	6.32%	7.54%	38.28%	84.30%	99.17%
96.01%	56.32%	17.69%	2.84%	2.60%	10.18%	27.18%	68.00%
61.66%	26.09%	7.61%	0.62%	0.58%	3.64%	11.69%	33.93%
65.48%	23.60%	4.90%	0.45%	0.39%	0.45%	0.39%	45.07%
81.44%	23.22%	2.89%	0.52%	0.32%	1.69%	14.67%	59.65%
83.85%	34.84%	6.08%	1.58%	0.89%	5.40%	30.09%	68.04%
89.69%	47.63%	12.49%	3.42%	2.46%	8.52%	41.35%	77.31%
78.78%	50.31%	23.41%	10.83%	6.21%	23.90%	58.86%	79.76%

Fig. 16. T-test results with and without background regularization for individual grid average of the relevance between tags and background.

From Figs. 15 and 16, on most grids of both the foreground and background, better results can be obtained with background regularization. There were some grids that led to worse results; however, there were no grids that led to worse results for both the foreground and background simultaneously throughout all grids. Therefore, as a result of the background regularization, it can be seen that the embedded representations for tags are learned so that they are far from the background, and as a result, the representations can be learned to be closer to the foreground. From these results, it was clearly confirmed that foreground-centered learning and background regularization using the grid weight map contributed to learning that emphasized the foreground and did not emphasize the background.

Considering that AAM is displayed to the user in the application to the real-world service, the colored background always causes user distrust. In GWAP+reg, the relevance between the background grids and the tag became low. This result suggests that the proposed method, in which the background is less likely to be colored, is more effective than the method without background regularization when considering the application of real-world services from the viewpoint of AAM.

Furthermore, since the result of semantic segmentation is obtained in the proposed system in preprocessing, it is possible to set the degree of relevance to the background area to 0 before the results of AAM show to the users. Moreover, in our experiments, by giving the highest priority to reducing the loss as much as possible, the weight parameter β related to the background regularization term was set to 0.1. By increasing this parameter, it seems to be possible that the range in which the proposed system is superior can be expanded.

4.4. Image retrieval evaluation

Using the proposed system (including foreground-centered learning, background regularization, and negative sampling with Eq. (10)), some tags are added (or subtracted) to evaluation images, and it is verified by a user questionnaire. For evaluation purposes, 21 randomly sampled full-body outfit images were used for image retrieval, and a questionnaire was conducted. One abstract tag attached to the target image was subtracted, and another abstract tag was added instead. This tag was chosen arbitrarily. This questionnaire confirms whether the proposed system can accurately perform image retrieval operations (via image and tag calculation) on images containing a wide range of backgrounds and tags, including abstract tags, compared to the comparison method (Han et al., 2017). A sample of the questionnaire is shown in Fig. 17.

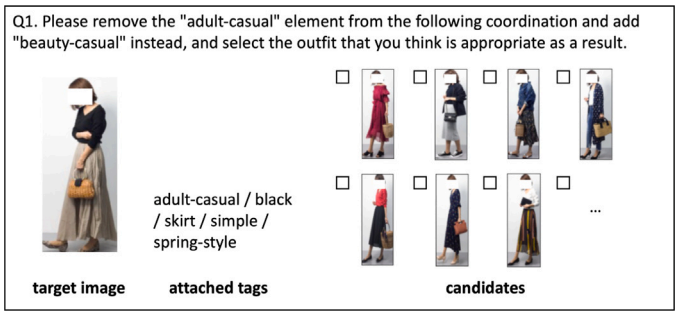

Fig. 17. A sample evaluation questionnaire.

More specifically, in each question, the top five images of the search results were selected from the comparison method and the proposed system. From a total of 10 images, the images evaluated by the subjects were correct as the search results were selected (with duplication allowed). For all questions, we calculated the average of the number of selected images and compared the results. In other words, 5.0 was found to be the best score. Experts have been defined for four types, as shown in Table 4. However, of the four types, subjects who have multiple applicable elements are included in multiple groups. Under these conditions, the summarized result of a survey for a total of 58 subjects is shown in Table 4.

Table 4
Result of questionnaire on outfit image retrieval.

	Number	Comparison	Proposal
Experts: subjects have an experience working in fashion shops or companies	10	0.919	**1.880**
Experts: subjects have a confident in fashion and fashion knowledge	13	0.915	**1.695**
Experts: subjects have an experience posting fashion items on SNS or blogs	17	1.103	**1.885**
Experts: subjects sometimes called fashionable from around them	24	1.009	**1.781**
Non-experts	21	1.199	**1.927**

From Table 4, it can be seen that the proposed system achieves more appropriate image retrieval results regardless of the group. Compared to other groups, subjects belonging to a group that is confident in fashion tend not to judge many images as correct answers; however, the proposed system still showed better results than the comparison method.

Furthermore, some examples of image retrieval results are shown in Fig. 18.

First, for clarity, the above two examples of using the image retrieval function with specific tags are shown. For example, if the "khaki" tag was removed from a query image (wearing jeans and a khaki shirt) that had "khaki" and "casual" tags attached, and a "yellow" tag was added instead, it would acquire the tag of images wearing jeans and a yellow shirt, while this full-body outfit image would remain in the casual category. In contrast, if the "white" tag was removed from a query image (wearing a white shirt and black trousers) with a "white" and "formal" tag, and a "yellow" tag was added; instead, it would acquire the tag of images of wearing black trousers and a yellow shirt, while this full-body outfit image would remain in the formal atmosphere. These results suggest that the operations using specific tags can be reasonably performed to some extent. Apart from the above two examples, some examples of the results of image retrieval using abstract tags (which is one of the main tasks of this research) are shown. For example, if the "casual" tag was removed from the query image on the left and the "formal" tag was added instead, the images including black trousers and a green shirt were acquired, and these full-body outfit images remained a formal atmosphere. Furthermore, if the "adult-casual" was added, the green shirt was changed to a long-length green item, and the full-body outfit images that looked more mature overall were acquired. From these results in Fig. 18, it can be seen that the proposed system obtains appropriate results for image retrieval.

By utilizing such an image retrieval function, even non-expert users can interpret fashion-specific abstract tags. It is expected that this interpretability will increase users' interest in and motivation for fashion and lead to purchasing behavior. Furthermore, it is also expected that experts will make new discoveries, and their perceptions will be broadened.

5. Additional analysis

Since the proposed model can acquire embedded representations for each image and tag in the same projective space, it is possible to perform multifaceted analyses other than analysis related to image retrieval. Moreover, analyzing these results is also useful for the interpretation/understanding of fashion and planning marketing strategies. This section analyzes in more detail the embedded representations obtained by applying the proposed system to actual data. Moreover, we clarify the usefulness of applying the proposed model to real-world services.

5.1. Ranking-based image retrieval using the relevance score

In addition to the abovementioned image retrieval function using images and words, it is also possible to search for images with a particularly high (low) relevance to the tag. An example is shown in Fig. 19.

For example, outfits with a higher relevance score with the "yellow" tag tend to have a higher proportion of the yellow part in the entire image. On the contrary, if the score is low, outfits including only a few parts of yellow appear. In the previous system, it was only possible to show all images with the tag "yellow" in a batch; however, by using this function, it is possible to search for clothes based on the user's purpose, such as "I want to find a yellow atmosphere as a whole", or "I want to incorporate yellow only for one point". Images can also be searched by the degree of relevance to abstract expressions, such as "casual", "office-casual", "beauty-casual", and "adult-casual", and scenes, such as "wedding-party". For example, denim tends to be included especially for casual outfits, and light-colored long coats and long skirts tend to be included for less casual outfits. Outfits that are highly relevant to the wedding also tend to include one-tone dresses that are not too bright, and outfits with low relevance tend to include patterned dresses and dresses that are too bright in color. By utilizing this result, the user can find out the outfits that should be worn especially for the wedding-party and can know the outfits that are more suitable for the office.

Fig. 18. An example of image retrieval.

5.2. Visual interpretation of abstract tags by AAM

By using AAM, it is possible to grasp which region on each image is relevant to a tag. By utilizing this function, it is possible to support the visual interpretation of the meaning of abstract fashion-specific tags. In this section, some tags, including abstract tags, are visually interpreted by utilizing AAM. An example of the result is shown in Fig. 20.

As shown in Fig. 20, it is possible to visually interpret which area on the image the tag attached to each full-body outfit image has a high degree of relevance. For example, regarding the AAM between the specific tag "yellow" and each image, the importance of actual yellow regions on the image becomes high; thus, appropriate results can be obtained. In addition, for "office-casual" clothing, shoes seem to be important for all images. Moreover, it can be seen that the rightmost image is particularly "office-casual" for the full-body.

In this way, it has become possible to visually interpret abstract fashion terms that are difficult for users to interpret and their relevance to the full-body outfit image. This function is expected to improve the understandability of coordinated images and interest in fashion and encourage users to study the fashion, purchase clothes, etc., in real-world services.

5.3. Checking the average of AAM

In the previous section, AAM was calculated for individual image and individual tag combinations. By calculating AAM for all images

Fig. 19. An example of ranking-based image retrieval using relevance score.

Fig. 20. An example of visual interpretation of tags using AAM.

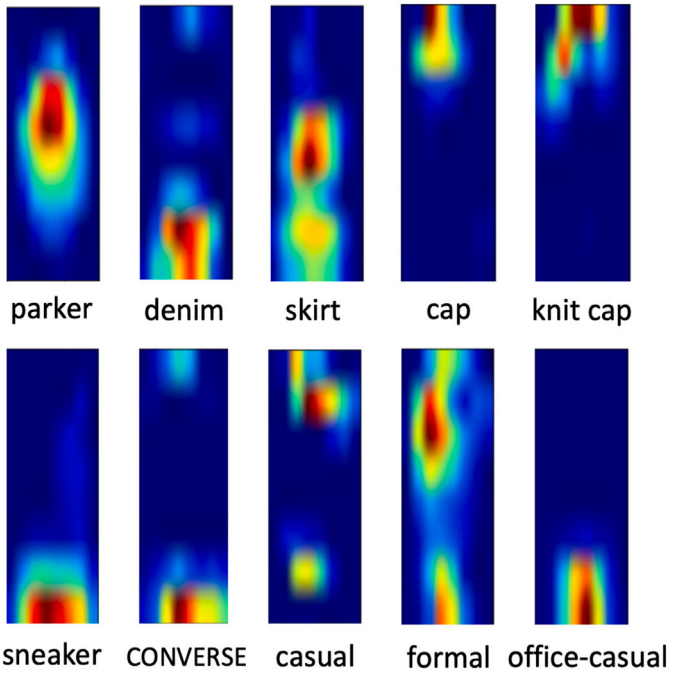

Fig. 21. An example of average AAM for each tag.

with individual tags and averaging them, it is possible to understand in which area on the image the tag tends to relate. Fig. 21 shows some examples of average of AAM.

From Fig. 21, it can be seen that specific tags such as sneakers, denim, and dresses tend to be more relevant in appropriate areas. From this result, it is suggested that, at least for specific tags, the embedded representations can be properly obtained while considering the relevance of the image features. On the contrary, it is difficult to judge from this figure whether abstract tags have been learned accurately or not. However, for example, it can be seen that the feet are important for "office-casual". Based on this result, when we confirmed the images that the tag "office-casual" is attached, there were many images wearing high-heel. In this way, it is possible to use the average of AAM to grasp the average of the features in all the images with the tag all at once and then perform an interpretation such as checking individual images. This allows users to know in advance that they should pay attention to feet when they have a question, "What is office casual?"

Moreover, for expressions such as "casual" where the points differ greatly depending on the combination of multiple clothes, it is doubtful what the average of AAM can be used for. This result may not be useful in marketing, but it may be one of the testaments of successful learning.

6. Discussion

6.1. Considerations based on experimental results and additional analyses

Considering the above experimental results, we can provide the following summary.

- The proposed system can properly learn even for complicated datasets (the full-body outfit images, a wide range and noisy backgrounds, and various tags, including abstract tags) compared to the conventional method.
- Each major improvement point (foreground-centered learning, background regularization, and negative sampling) contributes to appropriate learning.
- The proposed system was able to realize a more accurate image search by observing the image retrieval results using the calculation operation of obtained embedded representations.

Furthermore, in the additional analysis for utilizing the obtained results, we clarified that the following analysis can be achieved by applying the proposed system to real-world service data.

- By visualization of the degree of relevance between each image (area) and each tag, it is possible to interpret which area corresponds to the tag, even for abstract tags.
- By visualizing the average value of the degree of relevance for each tag and all the images that the tag attached, the proposed system can embed tags and images in the same space appropriately to some extent.

These analyses confirmed that the proposed system was able to perform the learning of fashion-specific knowledge contained in complex datasets. Furthermore, the possibility to utilize the obtained results in various ways in real-world services by analyzing the results from various angles was suggested. These multi-faceted features are expected to make it possible to support users online. Therefore, it can be said that the purpose of this research has been achieved.

6.2. Comparison with conventional studies

As mentioned in the related research section, the reason for recommending "the user seems to prefer this area in this item" can be obtained by the explainable recommendation methods using image information. However, it is unclear whether such information is what the users really desire. By contrast, the system proposed in this study makes it possible for users to obtain the information they truly desire to obtain by themselves because of fashion-specific abstract expressions. Specifically, it is possible to obtain information such as "the casual point of this outfit is this area" or "if you make this outfit a little more casual, the answer is this outfit". This study makes it possible to truly help users, help them understand fashion online, motivate them to buy, and help them make convincing purchases.

In addition, the main task of the conventional image retrieval task was to search for similar items from the single clothes images dataset or to search for similar outfits from the outfit images dataset. These are easy tasks even for non-experts to understand. Moreover, there were also tasks such as Fig. 4 that search for the single item image by adding an image and specific words (such as "short-sleeves" and "black"); however, this is also an easy task even for a non-expert. In contrast, in this study, full-body outfit image retrieval by adding a tag containing abstract expressions peculiar to the fashion domain is realized. It also provides other functions such as a visually understanding support function for fashion-specific expressions using AAM. Therefore, compared to the conventional studies in the fashion clothes retrieval field, our study is useful because it makes it possible to truly help the users with multiple functions according to the point you really desire.

Based on these facts, this study proposes a novel task of "automatically interpreting the fashion domain", which is completely different from the conventional explainable recommendation and image retrieval task. This point is one of the major contributions of this research, and in this sense, this study is valuable.

6.3. Model structure considerations

The major feature of the proposed model structure is that person area detection and semantic segmentation need to be performed as preprocessing and are used as inputs. As for the person area detection, it is generally applied as a preprocessing in many other studies. In contrast, for foreground and background classification, besides semantic segmentation, a method of learning semantic segmentation task with the same model or a method of detecting regions for each clothes item part in advance and learning them individually can be considered. Alternatively, a method of introducing a self-attention mechanism to distinguish the foreground and the background in the model can be considered. However, with a view to practical application, this study focuses on the simplicity of the model. Furthermore, in the target data, the foreground is a person, and if the grid in which the person is reflected can be fortunately detected, the foreground/background separately learning becomes possible. Since it is easy to use a public pre-trained model for semantic segmentation recently, it is possible to eliminate the trouble of preparing a separate dataset for semantic segmentation task. Furthermore, if accurate foreground and background information are obtained in advance, it is better to be able to focus on the main task of "fashion interpretation" rather than learning multiple tasks at the same time. The accuracy of the preprocessing was confirmed only visually this time. However, unlike the image where a person is small and in the corner, only one person is shown in the center, so the semantic segmentation in pre-progressing is not a difficult task. Therefore, in this study, it can be suggested that the accuracy of this task will be not a problem.

Commonly, the exclusion of background during learning is one of the major issues in image processing. Foreground-centered learning using the grid weight map and background regularization proposed in this study can be applied to all models using CNN (including the GAP layer). Therefore, it can be considered that the proposal of this versatile method is one of the major contributions of this research.

Furthermore, the advantage of being able to utilize the grid weight map obtained separately is not only the possibility to focus on the original purpose to interpret fashion, but also the ability that the degree of relevance between tag and background in the AAM can be replaced to zero even if it becomes high. When the function for presenting AAM to the user is installed in the real-world online service, if the importance of the background area becomes high, it leads to distrust by the user. In other words, the ability to use a grid weight map for AAM is an extremely significant advantage in practical applications.

6.4. Loss function and regularization term considerations

In the experiment conducted in this study, the hyperparameters α and β were set to 0.1 and 0.1. The transition of the loss term and the regularization term at that time is as shown in Fig. 22.

From Fig. 22, it can be seen that the regularization term had almost no adverse effect on the whole loss function because both of them are steadily decreasing in this condition and setting. If α is set too large, the area where people are shown will also be treated as the background. Therefore, the threshold should be set by comparing the obtained grid weight map with the corresponding image. Moreover, if β is set too large, the subtask that keeps the background and the

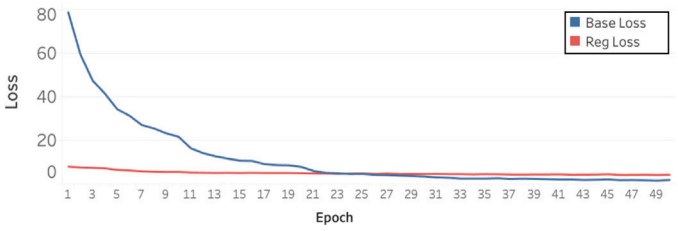

Fig. 22. Variation of loss terms and regularization items with number of epochs.

tag away will interfere with the main task that brings the foreground and the tag closer Fig. 23. Since there is a risk that the entire loss function is adversely affected by the background regularization term, it is necessary to decide while checking the transition of the entire loss and the background regularization term.

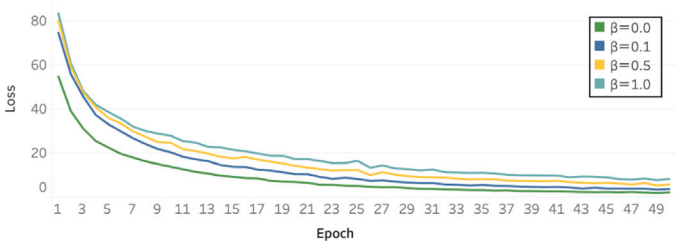

Fig. 23. Variation of loss terms with number of epochs and hyperparameter β.

6.5. Image retrieval improvement considerations

To realize more sensitive fashion interpretation learning, it is conceivable to embed images for each clothing part (collar, sleeves, torso, legs, etc.) individually, or to perform search calculations for each part individually. However, learning each item individually and clarifying the relationship between each item and each tag is not a simple task, and the model will definitely become complicated. Specifically, finally, there are problems such as how to combine the features of each part. In contrast, the multifaceted evaluation experiments and analyses in this study suggest that the proposed system may be useful to some extent, so it can be considered that this is sufficient as a starting point in this field "fashion intelligence". After this study, when improving the accuracy, it is necessary to consider the improvement while recognizing that the improvement of the sensitive learning is a trade-off with the complexity of the model.

To combine self-attention mechanism for embedding attributes (Shimizu et al., 2022; Zhang et al., 2021) and hashing (tag complementation) algorithm (Jin et al., 2020; Liu et al., 2018; Yan et al., 2019; Zhang & Peng, 2019) simultaneously as learning a model for visual-semantic embedding can also be considered. However, since each factor is a research area that has been deeply tackled as an independent field, comprehensive learning of these technologies with one model leads to the complexity of the model. For example, for hashing, it is necessary to properly learn visual-semantic embedding while ensuring the accuracy of tag completion. In brief, it is necessary to consider the trade-off between accuracy and model complexity.

6.6. Considerations regarding actual service application

Regarding the image retrieval experiments on dataset-2 and interpretation with AAM, we obtained quite convincing results. However, in another experiment (not published in this paper) on dataset-1, it was difficult to obtain an intuitive image retrieval result and AAM. In other words, even if the learning progresses well against ground truth tags, the excessive variety of attached tag noise, background, and pose makes learning extremely difficult. Accurate interpretation for datasets that contain such large amounts of noise is reserved for future research. In contrast, it is now clear that accurate interpretations and useful results can be obtained by properly selecting a dataset in advance (as in dataset-2). That means it is possible to support users of EC sites (dealing with images with little noise) with the proposed model. Moreover, it can also be used on SNS by carefully selecting the data (as in dataset-2). Therefore, it can be suggested that the proposed system can be fully used in real-world applications.

7. Conclusions and future work

In this study, we defined a novel technology and research area called "fashion intelligence" and proposed a fashion intelligence system that enables the interpretation of abstract fashion attribute information. The proposed system is based on a visual-semantic embedding method, including foreground-centered learning, background regularization, and negative sampling according to a duplicated relation and prior distribution that reflects the co-occurrence of attached tags. The proposed system utilizes methods such as image retrieval by calculation of rich tags (including abstract tags) and full-body outfit image similarities and visual interpretation support by AAM. The effectiveness of the proposed system was demonstrated by applying it to real-world service datasets and conducting evaluation experiments. Multiple analyses of experimental results indicate a method (with the proposed system) that helps decrease fashion-specific ambiguity and complexity (including the dependence on personal judgments) and supports users in understanding and interpreting fashion items. Furthermore, real-world system application, supporting fashion studies, and online user purchasing activities is expected.

A possible future research direction will include the building of models that learn discrete fashion parts and building a multimodal model that simultaneously performs tag complementation. Furthermore, our future work also incorporates proposals for new learning and evaluation methods for quantifying the abstract aspects of the fashion domain, which will contribute to the development of new "fashion intelligence" systems utilizing machine learning.

CRediT authorship contribution statement

Ryotaro Shimizu: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Project administration. **Yuki Saito:** Conceptualization, Funding acquisition. **Megumi Matsutani:** Writing – review & editing, Funding acquisition. **Masayuki Goto:** Conceptualization, Writing – review & editing, Funding acquisition, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgment

This work was supported by JSPS, Japan KAKENHI Grant Number 21H04600.

References

Ak, K. E., Lim, J. H., Sun, Y., Tham, J. Y., & Kassim, A. A. (2021). FashionSearchNet-v2: Learning attribute representations with localization for image retrieval with attribute manipulation. CoRR arXiv:2111.14145, https://arxiv.org/abs/2111.14145.

Ak, K. E., Lim, J. H., Tham, J. Y., & Kassim, A. A. (2018). Efficient multi-attribute similarity learning towards attribute-based fashion search. In *Proceedings of the 2018 IEEE winter conference on applications of computer vision (WACV)* (pp. 1671–1679). http://dx.doi.org/10.1109/ICCV48922.2021.01193.

Amazon. com, Inc. (2022). Amazon. Retrieved from https://www.amazon.co.jp/. Accessed May 20, 2022.

Arora, S., & Majumdar, A. (2022). Machine learning and soft computing applications in textile and clothing supply chain: Bibliometric and network analyses to delineate future research agenda. *Expert Systems with Applications*, *200*, Article 117000. http://dx.doi.org/10.1016/j.eswa.2022.117000.

Banerjee, D., Dhakad, L., Maheshwari, H., Chelliah, M., Ganguly, N., & Bhattacharya, A. (2022). Recommendation of compatible outfits conditioned on style. In *Advances in information retrieval* (pp. 35–50). Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-99736-6_3.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. http://dx.doi.org/10.5555/944919.944937.

Chen, Y., & Bazzani, L. (2020). Learning joint visual semantic matching embeddings for language-guided retrieval. In *Proceedings of the computer vision – ECCV 2020* (pp. 136–152). http://dx.doi.org/10.1007/978-3-030-58542-6_9.

Chen, X., Chen, H., Xu, H., Zhang, Y., Cao, Y., Qin, Z., & Zha, H. (2019). Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval* (pp. 765–774). http://dx.doi.org/10.1145/3331184.3331254.

Chen, M.-K., Wang, Y.-H., & Hung, T.-Y. (2014). Establishing an order allocation decision support system via learning curve model for apparel logistics. *Journal of Industrial and Production Engineering*, *31*(5), 274–285. http://dx.doi.org/10.1080/21681015.2014.951406.

Chen, L., Yuan, F., Jose, J. M., & Zhang, W. (2018). Improving negative sampling for word representation using self-embedded features. In *Proceedings of the eleventh ACM international conference on web search and data mining* (pp. 99–107). http://dx.doi.org/10.1145/3159652.3159695.

Chopra, A., Sinha, A., Gupta, H., Sarkar, M., Ayush, K., & Krishnamurthy, B. (2019). Powering robust fashion retrieval with information rich feature embeddings. In *Proceedings of the 2019 IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 326–334). http://dx.doi.org/10.1109/CVPRW.2019.00045.

Corbière, C., Ben-Younes, H., Ramé, A., & Ollion, C. (2017). Leveraging weakly annotated data for fashion image retrieval and label prediction. In *Proceedings of the 2017 IEEE international conference on computer vision workshops* (pp. 2268–2274). http://dx.doi.org/10.1109/ICCVW.2017.266.

Daghaghi, S., Medini, T., Meisburger, N., Chen, B., Zhao, M., & Shrivastava, A. (2021). A tale of two efficient and informative negative sampling distributions. *Vol. 139*, In *Proceedings of the 38th international conference on machine learning* (pp. 2319–2329). https://proceedings.mlr.press/v139/daghaghi21a.html.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: a large-scale hierarchical image database. In *Proceedings of the 2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). https://ieeexplore.ieee.org/document/5206848.

Dong, J., Ma, Z., Mao, X., Yang, X., He, Y., Hong, R., & Ji, S. (2021). Fine-grained fashion similarity prediction by attribute-specific embedding learning. *IEEE Transactions on Image Processing*, *30*, 8410–8425. http://dx.doi.org/10.1109/TIP.2021.3115658.

Faghri, F., Fleet, D. J., Kiros, J. R., & Fidler, S. (2018). Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the british machine vision conference 2018* (pp. 1–13). http://bmvc2018.org/contents/papers/0344.pdf.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., & Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. *Vol. 2*, In *Proceedings of the 26th international conference on neural information processing systems* (pp. 2121–2129). https://dl.acm.org/doi/10.5555/2999792.2999849.

Gajic, B., & Baldrich, R. (2018). Cross-domain fashion image retrieval. In *Proceedings of the 2018 IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 1950–1952). http://dx.doi.org/10.1109/CVPRW.2018.00243.

Gao, Y., Kuang, Z., Li, G., Luo, P., Chen, Y., Lin, L., & Zhang, W. (2020). Fashion retrieval via graph reasoning networks on a similarity pyramid. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–15. http://dx.doi.org/10.1109/TPAMI.2020.3025062.

Gao, Z., Wei, H., Guan, W., Nie, W., Liu, M., & Wang, M. (2021). Multigranular visual-semantic embedding for cloth-changing person re-identification. http://dx.doi.org/10.48550/arXiv.2108.04527, CoRR arXiv:2108.04527.

Giri, C., Jain, S., Zeng, X., & Bruniaux, P. (2019). A detailed review of artificial intelligence applied in the fashion and apparel industry. *IEEE Access*, *7*, 95376–95396. http://dx.doi.org/10.1109/ACCESS.2019.2928979.

Goto, M. (2019). Latent class models on business analytics. In *Proceedings of the 2019 IEEE international conference on big data, cloud computing, data science & engineering* (pp. 142–147). http://dx.doi.org/10.1109/BCD.2019.8885147.

Guan, C., Qin, S., Ling, W., & Ding, G. (2016). Apparel recommendation system evolution: an empirical review. *International Journal of Clothing Science and Technology*, *28*(6), 854–879. http://dx.doi.org/10.1108/IJCST-09-2015-0100.

Guo, S., Huang, W., Zhang, X., Srikhanta, P., Cui, Y., Li, Y., Adam, H., Scott, M. R., & Belongie, S. (2019). The imaterialist fashion attribute dataset. In *Proceedings of the 2019 IEEE/CVF international conference on computer vision workshop (ICCVW)* (pp. 3113–3116). http://dx.doi.org/10.1109/ICCVW.2019.00377.

Han, X., Song, X., Yao, Y., Xu, X.-S., & Nie, L. (2020). Neural compatibility modeling with probabilistic knowledge distillation. *IEEE Transactions on Image Processing*, *29*, 871–882. http://dx.doi.org/10.1109/TIP.2019.2936742.

Han, X., Wu, Z., Huang, P. X., Zhang, X., Zhu, M., Li, Y., Zhao, Y., & Davis, L. S. (2017). Automatic spatially-aware fashion concept discovery. In *Proceedings of the 2017 IEEE international conference on computer vision* (pp. 1472–1480). http://dx.doi.org/10.1109/ICCV.2017.163.

He, R., & McAuley, J. (2016). Ups and downs: modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th international conference on world wide web* (pp. 507–517). http://dx.doi.org/10.1145/2872427.2883037.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 289–296). http://dx.doi.org/10.1145/312624.312649.

Hou, Y., Vig, E., Donoser, M., & Bazzani, L. (2021). Learning attribute-driven disentangled representations for interactive fashion retrieval. In *Proceedings of the 2021 IEEE/CVF international conference on computer vision* (pp. 12127–12137). http://dx.doi.org/10.1109/ICCV48922.2021.01193.

Hou, M., Wu, L., Chen, E., Li, Z., Zheng, V. W., & Liu, Q. (2019). Explainable fashion recommendation: A Semantic Attribute Region guided approach. In *Proceedings of the 28th international joint conference on artificial intelligence* (pp. 4681–4688). http://dx.doi.org/10.5555/3367471.3367694.

Hussain, S., Zia, M. A., & Arshad, W. (2021). Additive deep feature optimization for semantic image retrieval. *Expert Systems with Applications*, *170*, Article 114545. http://dx.doi.org/10.1016/j.eswa.2020.114545.

Ibrahimi, S., van Noord, N., Geradts, Z., & Worring, M. (2019). Deep metric learning for cross-domain fashion instance retrieval. In *Proceedings of the 2019 IEEE/CVF international conference on computer vision workshop* (pp. 3165–3168). http://dx.doi.org/10.1109/ICCVW.2019.00390.

Jin, S., Zhou, S., Liu, Y., Chen, C., Sun, X., Yao, H., & Hua, X.-S. (2020). SSAH: Semi-supervised adversarial deep hashing with self-paced hard sample generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(07), 11157–11164. http://dx.doi.org/10.1609/aaai.v34i07.6773.

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the 2015 IEEE conference on computer vision and pattern recognition* (pp. 3128–3137). http://dx.doi.org/10.1109/CVPR.2015.7298932.

Kim, H., Jung, W.-K., Park, Y.-C., Lee, J.-W., & Ahn, S.-H. (2022). Broken stitch detection method for sewing operation using CNN feature map and image-processing techniques. *Expert Systems with Applications*, *188*, Article 116014. http://dx.doi.org/10.1016/j.eswa.2021.116014.

Kim, S., Park, H., & Lee, J. (2020). Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis. *Expert Systems with Applications*, *152*, Article 113401. http://dx.doi.org/10.1016/j.eswa.2020.113401.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., & Ferrari, V. (2020). The open images dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, *128*, 1956–1981. http://dx.doi.org/10.1007/s11263-020-01316-z.

Lee, C., Choy, K., Ho, G., & Lam, C. (2016). A slippery genetic algorithm-based process mining system for achieving better quality assurance in the garment industry. *Expert Systems with Applications*, *46*, 236–248. http://dx.doi.org/10.1016/j.eswa.2015.10.035.

Liang, Z., Guo, K., Li, X., Jin, X., & Shen, J. (2022). Person foreground segmentation by learning multi-domain networks. *IEEE Transactions on Image Processing*, *31*, 585–597. http://dx.doi.org/10.1109/TIP.2021.3097169.

Liang, T.-P., & Liu, Y.-H. (2018). Research landscape of business intelligence and big data analytics: A bibliometrics study. *Expert Systems with Applications*, *111*, 2–10. http://dx.doi.org/10.1016/j.eswa.2018.05.018.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: common objects in context. In *Proceedings of the computer vision – ECCV 2014* (pp. 740–755). https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48.

Lin, Y., Ren, P., Chen, Z., Ren, Z., Ma, J., & de Rijke, M. (2020). Explainable outfit recommendation with joint outfit matching and comment generation. *IEEE Transactions on Knowledge and Data Engineering*, *32*(8), 1502–1516. http://dx.doi.org/10.1109/TKDE.2019.2906190.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. E., Fu, C., & Berg, A. C. (2016). SSD: Single shot MultiBox detector. *Vol. 9905*, In *Proceedings of the computer vision – ECCV 2016* (pp. 21–37). Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-46448-0_2.

Liu, S., & Deng, W. (2015). Very deep convolutional neural network based image classification using small training sample size. In *Proceedings of the 2015 3rd IAPR asian conference on pattern recognition* (pp. 730–734). http://dx.doi.org/10.1109/ACPR.2015.7486599.

Liu, L., Du, X., Zhu, L., Shen, F., & Huang, Z. (2018). Learning discrete hashing towards efficient fashion recommendation. *Data Science and Engineering*, *3*, 307–322. http://dx.doi.org/10.1007/s41019-018-0079-z.

Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016). Deepfashion: powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1096–1104). http://dx.doi.org/10.1109/CVPR.2016.124.

Liu, Y., Zhou, W., Liu, J., Qi, G.-J., Tian, Q., & Li, H. (2021). An end-to-end foreground-aware network for person re-identification. *IEEE Transactions on Image Processing*, *30*, 2060–2071. http://dx.doi.org/10.1109/TIP.2021.3050839.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the 2015 IEEE conference on computer vision and pattern recognition* (pp. 3431–3440). http://dx.doi.org/10.1109/CVPR.2015.7298965.

Lu, J.-M., Wang, M.-J. J., Chen, C.-W., & Wu, J.-H. (2010). The development of an intelligent system for customized clothing making. *Expert Systems with Applications*, *37*(1), 799–803. http://dx.doi.org/10.1016/j.eswa.2009.05.089.

Meta Platforms, Inc. (2022). Facebook. Retrieved from https://www.facebook.com/. Accessed May 20, 2022.

Meta Platforms, Inc. (2022). Instagram. Retrieved from https://www.instagram.com/. Accessed May 20, 2022.

Moon, G. E., Newman-Griffis, D., Kim, J., Sukumaran-Rajam, A., Fosler-Lussier, E., & Sadayappan, P. (2019). Parallel data-local training for optimizing Word2Vec embeddings for word and graph embeddings. In *Proceedings of the 2019 IEEE/ACM workshop on machine learning in high performance computing environments* (pp. 44–55). http://dx.doi.org/10.1109/MLHPC49564.2019.00010.

Park, S., Shin, M., Ham, S., Choe, S., & Kang, Y. (2019). Study on fashion image retrieval methods for efficient fashion visual search. In *Proceedings of the 2019 IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 316–319). http://dx.doi.org/10.1109/CVPRW.2019.00042.

Qiu, S. (2018). Global weighted average pooling bridges pixel-level localization and image-level classification. http://dx.doi.org/10.48550/arXiv.1809.08264, CoRR arXiv:1809.08264.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Vol. 1*, In *Proceedings of the 28th international conference on neural information processing systems* (pp. 91–99). https://dl.acm.org/doi/10.5555/2969239.2969250.

Ren, M., Kiros, R., & Zemel, R. S. (2015). Exploring models and data for image question answering. *Vol. 2*, In *Proceedings of the 28th international conference on neural information processing systems* (pp. 2953–2961). https://dl.acm.org/doi/10.5555/2969442.2969570.

Riahi, Y., Saikouk, T., Gunasekaran, A., & Badraoui, I. (2021). Artificial intelligence applications in supply chain: A descriptive bibliometric analysis and future research directions. *Expert Systems with Applications*, *173*, Article 114702. http://dx.doi.org/10.1016/j.eswa.2021.114702.

Saito, Y., Nakamura, T., Hachiya, H., & Fukumizu, K. (2020). Exchangeable deep neural networks for set-to-set matching and learning. In *Proceedings of the computer vision – ECCV 2020* (pp. 626–646). http://dx.doi.org/10.1007/978-3-030-58520-4_37.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the 2018 IEEE/CVF conference on computer vision and pattern recognition* (pp. 4510–4520). http://dx.doi.org/10.1109/CVPR.2018.00474.

Shimizu, R., Matsutani, M., & Goto, M. (2022). An explainable recommendation framework based on an improved knowledge graph attention network with massive volumes of side information. *Knowledge-Based Systems*, *239*, Article 107970. http://dx.doi.org/10.1016/j.knosys.2021.107970.

Shin, M., Park, S., & Kim, T. (2019). Semi-supervised feature-level attribute manipulation for fashion image retrieval. http://dx.doi.org/10.48550/arXiv.1907.05007, CoRR arXiv:1907.05007.

Snap Vision Ltd. (2022). Snap Fashion. Retrieved from https://www.snapfashion.com/. Accessed May 20, 2022.

Stergiou, S., Straznickas, Z., Wu, R., & Tsioutsiouliklis, K. (2017). Distributed negative sampling for word embeddings. *Vol. 31*, In *Proceedings of the 31st AAAI conference on artificial intelligence* (pp. 2569–2575). https://ojs.aaai.org/index.php/AAAI/article/view/10931.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the 2016 IEEE conference on computer vision and pattern recognition* (pp. 2818–2826). http://dx.doi.org/10.1109/CVPR.2016.308.

Tautkute, I., Trzcinski, T., Skorupa, A. P., Brocki, L., & Marasek, K. (2019). DeepStyle: Multimodal search engine for fashion and interior design. *IEEE Access*, *7*, 84613–84628. http://dx.doi.org/10.1109/ACCESS.2019.2923552.

Twiter, Inc. (2022). Twitter. Retrieved from https://twitter.com/. Accessed May 20, 2022.

Vittayakorn, S., Yamaguchi, K., Berg, A. C., & Berg, T. L. (2015). Runway to realway: Visual analysis of fashion. In *Proceedings of the 2015 IEEE winter conference on applications of computer vision* (pp. 951–958). http://dx.doi.org/10.1109/WACV.2015.131.

Wang, J., Cheng, X., Wang, R., & Liu, S. (2021). Learning outfit compatibility with graph attention network and visual-semantic embedding. In *Proceedings of the 2021 IEEE international conference on multimedia and expo* (pp. 1–6). http://dx.doi.org/10.1109/ICME51207.2021.9428401.

Wu, H., Mao, J., Zhang, Y., Jiang, Y., Li, L., Sun, W., & Ma, W.-Y. (2019). Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the 2019 IEEE/CVF conference on computer vision and pattern recognition* (pp. 6602–6611). http://dx.doi.org/10.1109/CVPR.2019.00677.

Xiang, X., Zhang, Y., Jin, L., Li, Z., & Tang, J. (2022). Sub-region localized hashing for fine-grained image retrieval. *IEEE Transactions on Image Processing*, *31*, 314–326. http://dx.doi.org/10.1109/TIP.2021.3131042.

Yan, C., Chen, Y., & Zhou, L. (2019). Differentiated fashion recommendation using knowledge graph and data augmentation. *IEEE Access*, *7*, 102239–102248. http://dx.doi.org/10.1109/ACCESS.2019.2928848.

Yang, Z., Ding, M., Zhou, C., Yang, H., Zhou, J., & Tang, J. (2020). Understanding negative sampling in graph representation learning. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1666–1676). http://dx.doi.org/10.1145/3394486.3403218.

Zhan, H., Lin, J., Ak, K. E., Shi, B., Duan, L.-Y., & C. Kot, A. (2022). A^3-FKG: Attentive attribute-aware fashion knowledge graph for outfit preference prediction. *IEEE Transactions on Multimedia*, *24*, 819–831. http://dx.doi.org/10.1109/TMM.2021.3059514.

Zhang, Y., & Chen, X. (2020). *Explainable recommendation: A survey and new perspectives*. Now Foundations and Trends, http://dx.doi.org/10.1561/1500000066.

Zhang, M., Palade, V., Wang, Y., & Ji, Z. (2021). Attention-based word embeddings using artificial bee colony algorithm for aspect-level sentiment classification. *Information Sciences*, *545*, 713–738. http://dx.doi.org/10.1016/j.ins.2020.09.038.

Zhang, J., & Peng, Y. (2019). SSDH: Semi-supervised deep hashing for large scale image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, *29*(1), 212–225. http://dx.doi.org/10.1109/TCSVT.2017.2771332.

Zhang, Y., Yao, Q., Shao, Y., & Chen, L. (2019). Nscaching: Simple and efficient negative sampling for knowledge graph embedding. In *Proceedings of the 2019 IEEE 35th international conference on data engineering* (pp. 614–625). http://dx.doi.org/10.1109/ICDE.2019.00061.

Zhang, Y., Zhang, Y., & Zhang, M. (2018). SIGIR 2018 workshop on ExplainAble recommendation and search (ears 2018). In *Proceedings of the 41st international ACM SIGIR conference on research & development in information retrieval* (pp. 1411–1413). http://dx.doi.org/10.1145/3209978.3210193.

Zhao, X., Qi, H., Luo, R., & Davis, L. (2019). A weakly supervised adaptive triplet loss for deep metric learning. In *Proceedings of the 2019 IEEE/CVF international conference on computer vision workshop* (pp. 3177–3180). http://dx.doi.org/10.1109/ICCVW.2019.00393.

ZOZO, Inc. (2022). WEAR. Retrieved from https://wear.jp/. Accessed May 20, 2022.

ZOZO, Inc. (2022). ZOZOTOWN. Retrieved from https://zozo.jp/. Accessed May 20, 2022.