Expressiveness is Effectiveness: Self-supervised Fashion-aware CLIP for Video-to-Shop Retrieval

Likai Tian[1,2], **Zhengwei Yang**[1,2], **Zechao Hu**[1,2], **Hao Li**[1,2], **Yifang Yin**[3] and **Zheng Wang**[1,2,*]

[1]National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, School of Computer Science, Wuhan University, China

[2]Hubei Key Laboratory of Multimedia and Network Communication Engineering

[3]Institute for Infocomm Research, A*STAR, Singapore

Abstract

The rise of online shopping and social media has spurred the Video-to-Shop Retrieval (VSR) task, which involves identifying fashion items (*e.g.*, clothing) in videos and matching them with identical products provided by stores. In real-world scenarios, human movement in dynamic video scenes can cause substantial morphological alterations of fashion items with aspects of occlusion, shifting viewpoints (parallax), and partial visibility (truncation). This results in those high-quality frames being overwhelmed by a vast of redundant ones, which makes the retrieval less effectiveness. To this end, this paper introduces a framework, named **S**elf-supervised **F**ashion-aware **CLIP** (SF-CLIP), for effective VSR. The SF-CLIP enables the discovery of salient frames with high fashion expressiveness via generating pseudo-labels from three key aspects of fashion expressiveness to assess occlusion, parallax, and truncation. With such pseudo-labels, the ability of CLIP is expanded to facilitate the discovery of salient frames. Furthermore, to encompass the comprehensive representations among salient frames, a dual-branch graph-based fusion module is proposed to extract and integrate inter-frame features. Extensive experiments demonstrate the superiority of SF-CLIP over the state-of-the-arts.

1 Introduction

With the emergence of e-commerce, online shopping has been progressively embraced and acclimated by the populace. The fashion trend led by Internet celebrities on video platforms like TikTok and Instagram has a vast influence in absorbing consumers for shopping. In this context, combining computer vision and clothing-related tasks [Yang *et al.*, 2023] can help identify clothing items, recognize patterns, and even suggest fashion choices. Therefore, the Video-to-Shop Retrieval (VSR) task emerges as the times require, aiming to match the fashion items (*e.g.*, clothing) showing in videos with identical items, which is also known as catalog imagery, provided by stores.

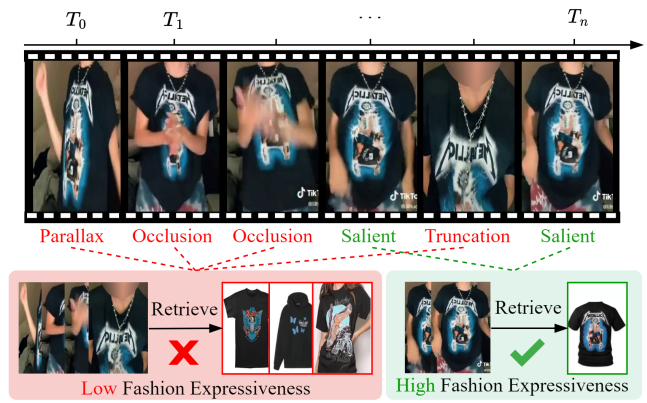

Figure 1: An example of the video in TikTok that contains frames with low fashion expressiveness (*e.g.*, occlusion, parallax, and truncation), which may lack discriminative fashion details like logos or clothing styles. In contrast, salient frames with high fashion expressiveness contribute to positive matches.

Several studies [Cheng *et al.*, 2017; Godi *et al.*, 2022] have investigated VSR by randomly sampling frames from videos and matching them within catalog imagery. However, dynamic video scenes inevitably introduce substantial noise, especially in VSR scenarios with human subjects exhibiting unpredictable movements. Such human movements bring morphological alterations with aspects of occlusion, shifting viewpoints (parallax), and partial visibility (truncation). Video frames with these alterations often act as noises, containing less vital information for retrieval. As shown in Figure 1, the query clothing suffers from being occluded by moving hands, experiences poor parallax by body twisting, and appears partially visible due to truncation.

To this end, this paper introduces the concept of **fashion expressiveness** and considers frames infected by such morphological alterations with low fashion expressiveness, which negatively misleads the retrieval (as shown in the red rectangle of Figure 1). Meanwhile, the salient frames with high quality are considered with high fashion expressiveness, which contributes to positive retrieval (as shown in the green rectangle of Figure 1). Since both types of frames coexist, randomly sampling frames poses a significant challenge in

*Corresponding author: wangzwhu@whu.edu.cn.

accurately targeting salient frames with high fashion expressiveness, thereby affecting the retrieval effectiveness.

Salient frame selection is not a fresh topic in the academic world. It has been widely investigated in video-based tasks like video-text retrieval [Wu *et al.*, 2021; Gorti *et al.*, 2022; Lu *et al.*, 2022] and video summarization [Ji *et al.*, 2019; Narasimhan *et al.*, 2021; Liu *et al.*, 2022]. The existing methods primarily focus on identifying salient frames by supervised interactions of multi- or single-modal features. However, the absence of fine-grained expressiveness labels poses a significant challenge in constructing direct interactions between frames and expressiveness descriptors.

Even the advanced FashionCLIP [Chia *et al.*, 2022], which fine-tunes CLIP [Radford *et al.*, 2021], the advanced Visual-Language Pertaining (VLP) model, on fashion datasets enables the cross-modal retrieval of fashion images, accurately capturing fashion expressiveness with absence labels remains an unaddressed area. To address the above challenges, we propose a **S**elf-supervised **F**ashion-aware **C**ontrastive **L**anguage-Image **P**re-training (SF-CLIP) framework. SF-CLIP builds pseudo-labels from three key fashion expressiveness aspects by assessing the occlusion, parallax, and truncation to broaden the capabilities of CLIP.

Specifically, the occlusion is assessed via inconsistencies across the outputs of multiple sub-networks with different activated neurons. This is based on the principle that a higher degree of occlusion, being less effective for retrieval, results in greater inconsistencies. For assessing parallax and truncation, SF-CLIP employs landmark analysis. The parallax of clothing is inferred from the relative positions of these landmarks, while the visibility of landmarks is used to gauge truncation levels. These simulations afford a detailed understanding of fashion expressiveness within video frames. Leveraging these pseudo-labels, SF-CLIP broadens the capabilities of the CLIP model to efficiently identify salient frames with high fashion expressiveness. Furthermore, to enhance the effectiveness of SF-CLIP and encompass the comprehensiveness of salient frames, a dual-branch graph-based fusion module is proposed to extract and integrate inter-frame features and employ automatic relation modeling to establish connections among them.

The SF-CLIP is delicately designed to highlight video frames with a strong sense of fashion expressiveness and comprehensively fuse them for effective retrieval. Frames with high fashion expressiveness are considered more informative and less susceptible to morphological alterations among a vast of video frames. SF-CLIP adeptly converts fashion expressiveness into improved effectiveness in VSR, notably enhancing retrieval performance. Furthermore, SF-CLIP also serves as an innovative tool for evaluating fashion expressiveness, demonstrating considerable ability in zero-shot scenarios. We hope this work could bring fundamental insights into related fields.

The contributions of this paper are threefold:

- We highlight the significance of salient frames in the Video-to-Shop Retrieval (VSR) task and introduce fashion expressiveness to determine the saliency of each video frame.
- We propose a Self-supervised Fashion-aware Contrastive Language-Image Pre-training (SF-CLIP) framework, which expands the ability of CLIP with a strong sense of fashion expressiveness for effective retrieval.
- The extensive experiments on two standard video-to-shop datasets, MovingFashion [Godi *et al.*, 2022] and DeepFashion2 [Ge *et al.*, 2019], demonstrate the superiority of the proposed SF-CLIP.

2 Related Work

Video-to-Shop Retrieval. Various methodologies have been developed for VSR. To fuse fashion features among individual frames, AsymNet [Cheng *et al.*, 2017] utilizes Long Short-Term Memory (LSTM) and a variable depth tree structure, while SEAM Match-RCNN [Godi *et al.*, 2022] employs a non-local attention mechanism. Despite the complexity and sophistication of their fusion strategies, they are inevitably limited by the expressiveness of the input frames, which are less likely to be informative when random sampling is performed in a complex and dynamic scenario. To cope with the above challenge in VSR, we focus on discovering salient frames with high fashion expressiveness. This approach aims at maximizing the extraction of valid fashion information within videos, thereby enhancing retrieval effectiveness.

Salient Frame Discovery. Real-world video scenarios often contain irrelevant or even disruptive information, which poses challenges in learning high-quality video representations. As a solution, salient frames have been introduced and proved to be effective in multiple video tasks. In video-text retrieval [Wu *et al.*, 2021; Gorti *et al.*, 2022; Lu *et al.*, 2022], salient frame strategies are employed to filter out query-irrelevant frames and achieve efficient cross-modal alignment. In video summarization [Ji *et al.*, 2019; Narasimhan *et al.*, 2021; Liu *et al.*, 2022], salient frames with rich information naturally align with the purpose of extracting representative frames for users to browse and quickly obtain the core information. In the context of the VSR task, we introduce the concept of fashion expressiveness and formulate several principles for discovering salient frames with high fashion expressiveness, thereby significantly enhancing retrieval effectiveness.

Visual-Language Pre-training. Visual-Language Pre-training (VLP) models [Narasimhan *et al.*, 2021; Li *et al.*, 2022b; Wang *et al.*, 2023] like CLIP [Radford *et al.*, 2021] in multimodal learning demonstrates exceptional portability in various downstream tasks, as evident from its success in both zero-shot [Sanghi *et al.*, 2022; Zhou *et al.*, 2023; Guo *et al.*, 2023] and fine-tuning [Rasheed *et al.*, 2023; Goyal *et al.*, 2023; Hegde *et al.*, 2023] manners. In fashion domain, FashionCLIP [Chia *et al.*, 2022] fine-tunes all CLIP parameters on new fashion data for classification and retrieval of fashion images in a fully supervised manner. In contrast, we propose the Self-supervised Fashion-aware CLIP (SF-CLIP) framework to perceive fashion expressiveness in video frames, thus filling the blank in fashion expressiveness studies and further expanding the ability of CLIP.

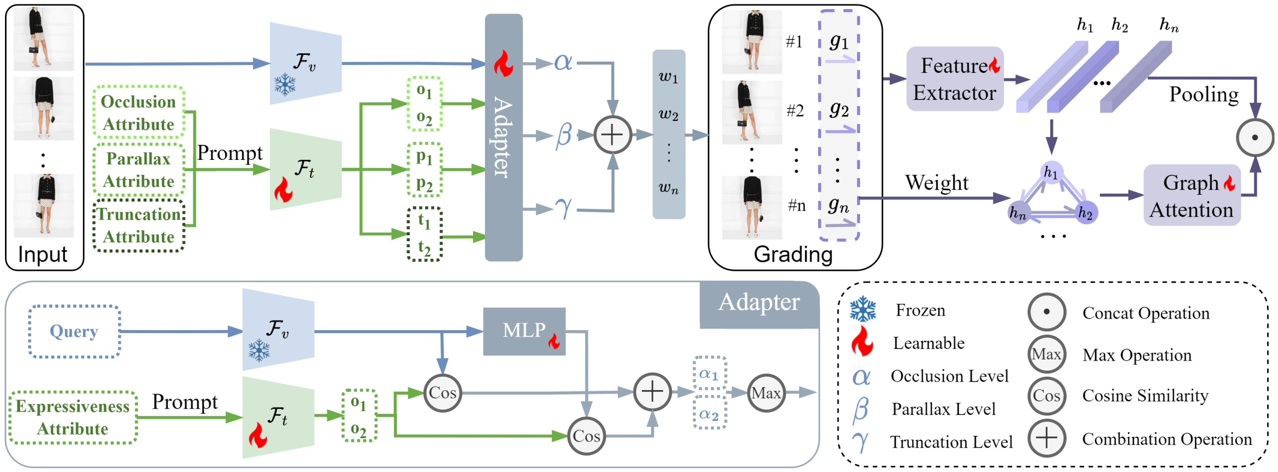

Figure 2: Structure of the proposed method. It consists of two modules, Fashion-aware CLIP and a dual-branch graph-based fusion module. Such structure is to select salient frames with high fashion expressiveness to enhance the effectiveness of VSR process.

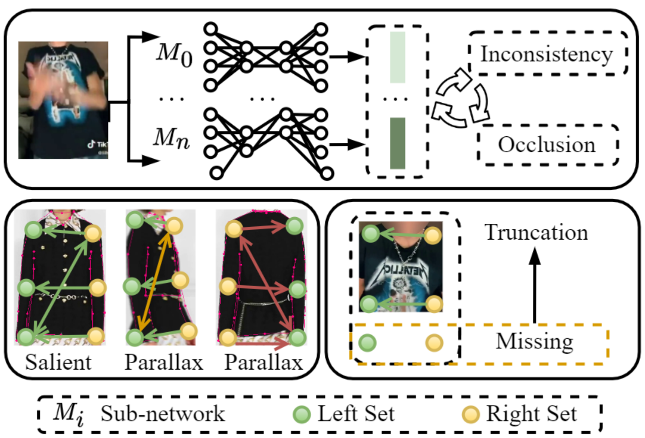

Figure 3: The process of generating pseudo-labels for three aspects of fashion expressiveness.

3 Method

In this section, details of the SF-CLIP framework for salient frame highlighting are outlined. Subsequently, a dual-branch graph-based fusion module is described, which is designed to fuse inter-frame features among salient frames for comprehensive fashion representation building. Figure 2 illustrates the global picture of the proposed framework. By default, as a standard preliminary step, a pre-trained clothing detection model [Li *et al.*, 2022a] is employed to identify clothing items associated with shop items in each frame.

3.1 Prompt Design

To effectively utilize the latent semantics of the pre-trained CLIP model, we devise antonym prompts accompanied by detailed context descriptions. To be more specific, fashion expressiveness is assessed from three aspects:

Occlusion. We consider two occlusion degrees: $o \in \mathcal{O} = \{$"partially", "heavily"$\}$ and design the corresponding textual prompt template "a photo of clothing $\{$o$\}$ covered by hair or arms".

Parallax. We define three different parallax degrees: $p \in \mathcal{P} = \{$"frontal", "side", "rear"$\}$ and design the corresponding fine-tuning prompt template "a photo of clothing taken from $\{$p$\}$ view".

Truncation. we distinguish two levels of truncation: $t \in \mathcal{T} = \{$"large", "small"$\}$ and develop corresponding fine-tuning prompt template "after cropping, $\{$t$\}$ part of clothing is retained".

3.2 Pseudo-label Generation for Occlusion

For the generation of pseudo-labels that annotate the severity of occlusion, we categorize the degree of clothing occlusion in the sample images into levels within $\mathcal{O}$, based on the consistency of the output feature under dropout. As illustrated in the top part of Figure 3, a pre-trained fashion retrieval model denoted as M, is applied to extract features. After dropout, it generates random sub-networks $M_1, M_2, ..., M_n$. We extract features from these sub-networks, acquiring a set of embeddings represented as $X = \{x_1, x_2, ..., x_n\}$. The consistency score is computed by the distances among these embeddings:

$$s(X) = \sigma\Big(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} D(x_i, x_j)\Big) \tag{1}$$

where $D(\cdot)$ is the Euclidean distance. We normalize the consistency score with Softmax to obtain the occlusion aspect score w_{occ}. Following this, we partition video frames into two degrees based on w_{occ} concerning a threshold score λ_{occ}.

3.3 Pseudo-label Generation for Parallax

To generate pseudo-labels for the parallax annotation, we classify the shooting viewpoint of clothing into three levels in $\mathcal{P}$ based on the relative locations of landmarks. As illustrated in Figure 3, we utilize the MMPose [Contributors, 2020] landmark detection model to detect landmarks of clothing in video frames. Subsequently, we propose a robust approach to assess the parallax of clothing based on landmark locations. In detail, we pre-define a set of six landmarks as K, including the top left k_{tl}, top right k_{tr}, left center k_{ml}, right center k_{mr}, bottom left k_{bl}, and bottom right k_{br}.

$$s(k_{ij}) = \begin{cases} 1, & k_i \in \{k_l\}, k_j \in \{k_r\} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$s(K) = \sum_{i=1}^{n} s(k_{ij}) + \sigma(\tan(k_{tr}, k_{bl})) \tag{3}$$

where k_l and k_r represent the set of left and right landmarks, respectively. The frontal extent is determined by whether a landmark k_i is positioned to the left of k_j. The side extent is indicated by the angles between the bottom left landmark k_{bl} and the top right landmark k_{tr}. We employ Softmax to normalize $s(K)$ and derive the parallax aspect score w_{par}. Video frames are partitioned into three degrees based on w_{par} using two threshold scores λ_{par_1} and λ_{par_2}.

3.4 Pseudo-label Generation for Truncation

To generate pseudo-labels for truncation annotation, we classify the truncation of clothing into two levels in $\mathcal{T}$ based on the confidence level of landmarks. As illustrated in Figure 3, if the confidence score of a landmark exceeds the threshold value λ, it indicates the presence of the corresponding patch; otherwise, it is considered absent. By calculating the ratio of the number of landmarks exceeding the confidence threshold to the total landmark count, we can estimate the truncation aspect score w_{tru}. Subsequently, video frames are partitioned into two groups based on w_{tru} using a threshold score λ_{tru}.

3.5 Framework Conduction

To balance the performance of Fashion-aware CLIP and the associated costs of self-supervised fine-tuning, we propose to embed an additional adapter $\mathcal{F}_a$ into the CLIP visual encoder. This layer functions collaboratively with the text encoder during the fine-tuning process while the original CLIP visual encoder is frozen. $\boldsymbol{f}_v = \mathcal{F}_v(I)$ refers to the visual embedding of the input video frame I, and $\boldsymbol{f}_t = \mathcal{F}_t(L_k)$ represent the text embedding corresponding to the k-th description template L_k. To prevent overfitting, we employ a linear combination of the Adapter's predictions and the original CLIP's predictions as the final output. Hence, the final prediction g_{clip} can be written as:

$$\boldsymbol{g}_{\text{clip}} = \boldsymbol{f}_v \boldsymbol{f}_t^{\mathsf{T}} + \mathcal{F}_a(\boldsymbol{f}_v) \boldsymbol{f}_t^{\mathsf{T}} \tag{4}$$

For the adapter layer, we employ standard learnable linear layer embeddings. The Fashion-aware CLIP framework minimizes the cross-entropy loss during the fine-tuning process.

3.6 Salient Frame Selection

The goal of salient frame selection is to eliminate non-salient frames and retain only those with high fashion expressiveness to enhance the effectiveness of VSR process. To accomplish this, we create image-text pairs by using prompt templates from the three aspects of fashion expressiveness: occlusion, parallax, and truncation, and then employ SF-CLIP to calculate the fashion expressiveness score for each video frame. Specifically, to determine the occlusion level of a frame, we compute the similarities between the text features o_1, o_2, which corresponds to the two occlusion levels, and the video frame features with Eq. 4. The occlusion level with the highest similarity denotes the occlusion level α. A similar process is applied to determine the parallax level β and truncation level γ of the video frame. Subsequently, by combining these three aspects, we obtain the fashion expressiveness score w_i, with which the top-N salient frames are picked out for effective retrieval.

3.7 Graph Fusion Module

To enhance the fashion representation of salient frames, we propose a dual-branch graph-based module to extract and fuse inter-frame features. This module consists of a global branch and a graph attention branch. Initially, we extract the feature embedding of salient frames donated as $\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_n$ and employ fashion expressiveness scores as the initial edge weights to model the relationships among salient frames.

Global Branch. We employ parameter-free average pooling to merge these embeddings into a global representation:

$$\boldsymbol{f}_g = avg(\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_n) \tag{5}$$

Graph Attention Branch. We dynamically update the importance of node $\boldsymbol{h}_j$ relative to node $\boldsymbol{h}_i$ using a self-attention mechanism ϕ [Veličković *et al.*, 2018], donated as α_{ij}:

$$\alpha_{ij} = \frac{\exp(\phi(W\boldsymbol{h}_i, W\boldsymbol{h}_j))}{\sum_{k \in N} \exp(\phi(W\boldsymbol{h}_i, W\boldsymbol{h}_k))} \tag{6}$$

where W is a linear mapping matrix. To enhance the generalization capability of the attention mechanism, we incorporate a multi-head attention layer for feature fusion:

$$\boldsymbol{h}_i^{l+1} = \|_{k=1}^{K} \, \sigma\left(\sum_{k \in N} \alpha_{ij}^k W^k \boldsymbol{h}_j^l\right) \tag{7}$$

where $||$ denotes the concatenation operation, K represents the number of heads, $\boldsymbol{h}_j^{(l)}$ corresponds to the hidden features of the j-th salient frame at layer l.

In the last layer of the network, the graph attention branch produces an output denoted as $\boldsymbol{f_s}$. To obtain the final fashion representation, we concatenate $\boldsymbol{f}_g$ and $\boldsymbol{f}_s$:

$$\boldsymbol{f}_{\text{final}} = \boldsymbol{f}_g \parallel \boldsymbol{f}_s \tag{8}$$

Loss Function. During the training process, we minimize the triplet loss $\mathcal{L}_{TR}$ for the CNN backbone while simultaneously minimizing the cross-entropy loss $\mathcal{L}_{CE}$ to classify street videos and shop images as positive or negative matches. The total loss $\mathcal{L}_{total}$ is calculated as the summation of both:

$$\mathcal{L}_{total} = \mathcal{L}_{TR} + \mathcal{L}_{CE} \tag{9}$$

4 Experiments

4.1 Datasets and Metrics

MovingFashion [Godi *et al.*, 2022] is a VSR dataset consisting of over 15,000 pairs of videos and corresponding online clothing items. It is categorized into two parts: Regular-MovingFashion, sourced from the Net-A-Porter website, and

Method	Venue	MovingFashion			Regular			Hard		
		R@1	R@5	Mean	R@1	R@5	Mean	R@1	R@5	Mean
Max Confidence [Ge *et al.*, 2019]	CVPR 19	0.29	0.59	0.44	0.31	0.63	0.47	0.21	0.46	0.33
Max Matching [Cheng *et al.*, 2017]	CVPR 17	0.26	0.60	0.43	0.29	0.65	0.47	0.17	0.44	0.30
NVAN [Liu *et al.*, 2019]	BMVC 19	0.38	0.62	0.50	0.47	0.73	0.60	0.11	0.28	0.19
VKD [Porrello *et al.*, 2020]	ECCV 20	0.40	0.49	0.44	0.49	0.59	0.54	0.13	0.20	0.16
MGH [Yan *et al.*, 2020]	CVPR 20	0.40	0.59	0.49	0.47	0.67	0.57	0.18	0.35	0.26
AsymNet [Cheng *et al.*, 2017]	CVPR 17	0.42	0.73	0.57	0.49	0.81	0.65	0.22	0.47	0.34
AsymNet [AVG]	CVPR 17	0.39	0.66	0.52	0.46	0.78	0.62	0.19	0.44	0.31
AsymNet [MAX]	CVPR 17	0.40	0.71	0.55	0.47	0.80	0.63	0.20	0.42	0.31
SEAM M-RCNN [Godi *et al.*, 2022]	WACV 22	0.49	0.80	0.64	0.55	0.86	0.70	0.30	**0.62**	0.46
Ours		**0.74**	**0.87**	**0.81**	**0.85**	**0.96**	**0.90**	**0.39**	0.60	**0.49**

Table 1: Evaluation of VSR performance in comparison with other state-of-the-art methods on MovingFashion [Godi *et al.*, 2022] dataset and its Regular and Hard subsets. Single-frame methods (top), Video-based person re-identification methods (middle-top), and VSR methods (middle-bottom) are compared. Bold numbers denote the best results.

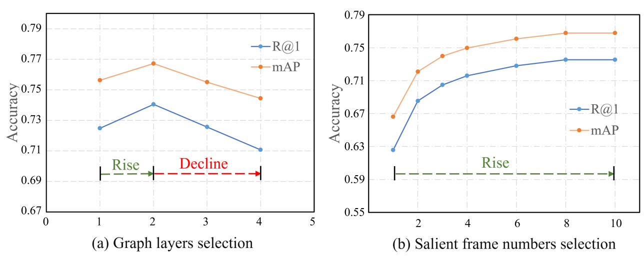

Figure 4: Ablation study of graph layers and salient frame numbers in MovingFashion, different color denotes different metrics.

Method	Multi-DeepFashion2		
	R@1	R@5	Mean
Max Confidence [Ge *et al.*, 2019]	0.19	0.44	0.31
Max Matching [Cheng *et al.*, 2017]	0.14	0.45	0.29
NVAN [Liu *et al.*, 2019]	0.22	0.37	0.29
VKD [Porrello *et al.*, 2020]	0.21	0.27	0.24
MGH [Yan *et al.*, 2020]	0.22	0.34	0.28
AsymNet [Cheng *et al.*, 2017]	0.21	0.50	0.35
AsymNet [AVG]	0.16	0.41	0.28
AsymNet [MAX]	0.15	0.42	0.28
SEAM M-RCNN [Godi *et al.*, 2022]	0.30	0.58	0.44
Ours	**0.58**	**0.80**	**0.69**

Table 2: VSR results on Multi-DeepFashion2.

Hard-MovingFashion, downloaded from social media platforms such as Instagram and TikTok.

Multi-DeepFashion2 [Godi *et al.*, 2022] The original DeepFashion2 dataset [Ge *et al.*, 2019] is primarily used for street-to-shop image retrieval. As in [Godi *et al.*, 2022], we pair shop images from the DeepFashion2 dataset with corresponding sequences of street images to simulate video scenarios, adapting it for the VSR task.

Evaluation Metric. According to previous studies [Cheng *et al.*, 2017; Godi *et al.*, 2022], we choose the standard top-K recall metric R@K to evaluate the performance, which is specifically compared on R@1, R@5, and the mean of them.

4.2 Implementation Details

During the pseudo-label generation process of occlusion aspect, we set the dropout rate to 0.5 and λ_{occ} to 0.9. As for truncation aspect, we set the confidence threshold λ to 0.8 and λ_{tru} to 0.6. The learning rate for Adapter is set to 3×10^{-4}, and TextEncoder is set to 5×10^{-7}, with a total of 30 epochs. To train Fashion-aware CLIP, we randomly select 1024 samples from MovingFashion dataset to generate pseudo-labels.

For the graph fusion module, we employ ResNet-50 as the backbone and apply two layers of graph convolution. The module is trained using Adam [Kingma and Ba, 2014] with a learning rate of 1×10^{-4} for a total of 60 epochs. The salient frame number is defined as 3 for training and 10 for testing.

4.3 Comparison with State-of-the-art Methods

We evaluate the effectiveness of the proposed method by comparing it with other state-of-the-art methods of different pipelines, as detailed in Table 1. It is unsurprising that single-frame methods relying on just one frame input fall short on R@1 compared to multi-frame methods. An interesting finding is that single-frame methods have better R@5 performance than some multi-frame ones on the Hard subset. As for video-based person re-identification, the main difference is VSR task faces limitations in available information due to the necessity of discarding facial features. Consequently, the evaluated re-identification methods have weaker performance than genuine VSR methods. Moreover, previous VSR methods, which randomly sample frames, are inevitably limited by numerous frames with low fashion expressiveness. As expected, these methods significantly lag behind the proposed method. Overall, the results demonstrate that our method outperforms other comparable methods across the entire MovingFashion dataset and its subsets. Additionally, we conduct experiments on Multi-DeepFashion2, where the proposed method consistently exhibits notable improvements over other comparable methods, as depicted in Table 2.

Figure 5: Qualitative results of basic method and the proposed SF-CLIP. Green rectangle represents the correct matching, respectively.

Basic		Expressiveness			MovingFashion		
B/L	G	Occ	Par	Tru	R@1	R@5	Mean
✓	-	-	-	-	71.1	85.9	78.5
✓	✓	-	-	-	71.9	86.6	79.2
✓	✓	✓	-	-	72.1	86.6	79.3
✓	✓	✓	✓	-	72.9	86.7	79.8
✓	✓	✓	✓	✓	**73.0**	**87.1**	**80.0**

Table 3: Ablation studies of each component of the proposed method on MovingFashion dataset, respectively.

4.4 Ablation Studies

Basic Component Analysis. The impact of each component within the proposed method is investigated in Table 3. Specifically, there are two basic components: the global branch indicated as "B/L", and the graph attention branch indicated as "G". "Occ", "Par" and "Tru" represent adding salient frame selection criteria concerning occlusion, parallax, and truncation, respectively. According to the first two rows of Table 3, it is evident that the graph attention branch has a positive effect, which enables comprehensive fashion representation building. To further demonstrate the effectiveness of pseudo labels generated for Fashion-aware CLIP fine-tuning, we utilize the three aspects as selection criteria of salient frames. In the last three rows of Table 3, the "Occ" criteria, designed for occlusion annotation, brings moderate improvement due to the use of sufficient, namely 10 salient frames, through which the occluded information is mutually compensated. "Par" increases the one-shot recall by a large margin, while "Tru" improves R@5. This progressive enhancement substantiates the accuracy of the generated pseudo-labels in multi-shot retrieval.

Incorporating CLIP Strategy Analysis. To investigate the impact of the proposed self-supervised Fashion-aware CLIP, we conduct experiments with different manners to incorporate CLIP, and the resulting performance is illustrated in Table 4. "No F" denotes the zero-shot operation, while "F" represents the fine-tuning operation. Upon comparing rows 1-3, the results demonstrate that although utilizing zero-shot CLIP can enhance the VSR process to some extent, fine-tuning the proposed Fashion-aware CLIP yields the optimal performance. One potential explanation is that the original CLIP encoder exhibits limitations in comprehending concepts such as occlusion, parallax, and truncation. Through self-supervised fine-tuning, the perception of fashion expressiveness is enhanced, leading to a more accurate selection of salient frames and VSR performance improvement.

Basic	CLIP		MovingFashion		
	No F	F	R@1	R@5	Mean
✓	-	-	71.9	86.6	79.2
✓	✓	-	72.8	86.8	79.8
✓	-	✓	**73.8**	**87.7**	**80.8**

Table 4: Ablation studies of incorporating CLIP strategy on MovingFashion dataset, respectively.

Graph Convolution Layers Analysis. To ascertain the optimal number of graph convolution layers for comprehensive feature fusion, we conduct an ablation study with varying layer numbers, specifically set to 1, 2, 3, and 4. As illustrated in Figure 4, the increment in the number of layers initially enhances VSR performance, followed by an obvious decrease. It is crucial to emphasize that graph convolution can be regarded as a form of Laplacian smoothing. However, a shallow graph convolution may fail to sufficiently propagate node information throughout the entire graph. Conversely, an excessively deep graph convolution raises concerns about potential over-smoothing issues.

Figure 6: Visualization of heatmaps of fashion representation.

Salient Frame Numbers Analysis. To investigate the optimal number of salient frames for the VSR task, we conduct experiments testing various salient frame numbers, ranging from 1 to 10. As depicted in Figure 4, a clear linear correlation between the number of salient frames and VSR performance is evident. Initially, accuracy improves with an increasing number of frames, but it eventually reaches a plateau. Our experiments reveal that the best performance, achieving an R@1 of 74%, is attained when using 10 frames. These findings underscore the critical role of salient frame quantity in VSR tasks, suggesting that 8 to 10 frames are sufficient to achieve satisfactory results.

4.5 Qualitative Results

Analysis of VSR Results. As mentioned, human movements in dynamic video scenarios can cause low fashion expressiveness with aspects of occlusion, parallax and trunca-

Figure 7: Visualization of assessment results of Fashion-aware CLIP and the corresponding grading rank, respectively.

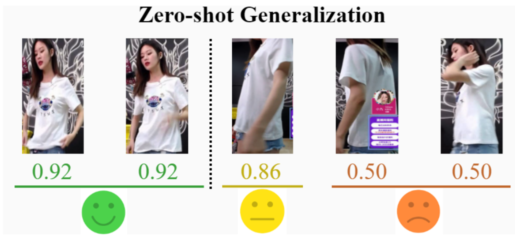

Figure 8: Visualization of zero-shot generation evaluation results. The scores indicate the level of fashion expressiveness, while smiley icons correspond to the results obtained from the user study.

tion, which poses a significant challenge for existing VSR methods. In response to this challenge, the proposed method aims to mitigate the impact of low fashion expressiveness by prioritizing salient frames. To showcase the effectiveness of the SF-CLIP, we compare it with the basic method that randomly samples frames. As depicted in Figure 5, we present ten candidate shop images with the highest similarity to the probe video. The results unequivocally demonstrate that the proposed method adeptly captures essential fashion details(e.g., texture), leading to accurate retrieval results. In contrast, the basic method exhibits heightened vulnerability to the challenges posed by low fashion expressiveness.

Analysis of Fashion Representation. To further investigate the effectiveness of the graph fusion module in capturing clothing knowledge, we specifically visualize the features of video frames. As illustrated in Figure 6, six heatmaps corresponding to their original frames are showcased. The results compellingly reveal that the graph fusion module adeptly identifies and precisely localizes significant patches(e.g., logos and texture), in accordance with our expectations.

Analysis of SF-CLIP. We visualize the fashion expressiveness scores of SF-CLIP on the MovingFashion dataset. Figure 7 showcases five video frames, along with their fashion expressiveness scores and grading ranks. The analysis highlights that video frames demonstrating high fashion expressiveness (columns 1 and 2) tend to receive higher scores, while clothing with issues such as severe occlusion (column 3), unfavorable parallax (column 4), or poor truncation (column 5) are attributed lower scores. Upon a comprehensive examination of the entire ranking, it becomes evident that the grading outcomes precisely align with expectations.

Analysis of Pseudo Labels. To evaluate the effectiveness of pseudo labels generated for fine-tuning Fashion-aware CLIP, we visualize scores of three aspects along with the

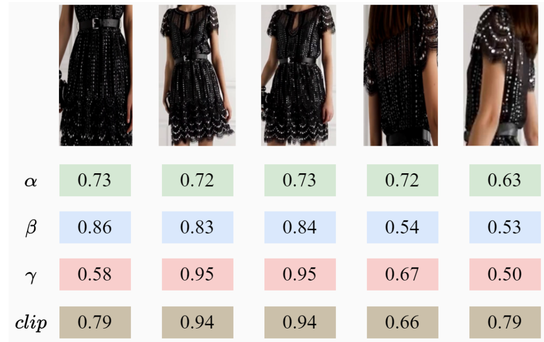

Figure 9: Visualization of three aspects of fashion expressiveness. α, β, and γ represent occlusion score, parallax score, and truncation score, while $clip$ donates the fashion expressiveness score evaluated by Fashion-aware CLIP, respectively.

fashion expressiveness score in Figure 9. Upon comparing columns 2 and 5, it is evident that the clothing in column 5 is significantly occluded by hair and arms, leading to a reduced score for the occlusion aspect. Similarly, in the comparison of columns 2 and 4, the clothing in column 4 displayed in a back view, receives a lower score for the parallax aspect, while the clothing in column 2, featuring a frontal view, achieves a higher score. Additionally, the analysis of columns 1 and 2 reveals that less truncated clothing receives a higher truncation score. These findings underscore the reliability of the three aspects in assessing fashion expressiveness.

Analysis of Zero-shot Generalization. Moreover, to evaluate the zero-shot generalization capability of SF-CLIP, we visualize its performance in assessing fashion expressiveness on online videos, as depicted in Figure 8. Through a user study, participants were given the opportunity to evaluate the fashion expressiveness within video frames. High expressiveness is represented by a green smiley, medium expressiveness by a yellow smiley, and low expressiveness by orange. The obtained evaluation results align with the outcomes of the user study, providing confirmation of the robust generalization ability inherent in the proposed method.

5 Conclusion

In this paper, we first introduce the concept of fashion expressiveness into the Video-to-Shop Retrieval (VSR) task for the measurement of video frame saliency. Subsequently, the Self-supervised Fashion-aware CLIP (SF-CLIP) framework is proposed to facilitate VSR task. The SF-CLIP is uniquely designed to discover salient frames with high fashion expressiveness, circumventing the need for manually annotating fine-grained expressiveness labels. The proposed method significantly improves the efficiency of extracting valid fashion information from video frames, thereby boosting the overall effectiveness of retrieval. Moreover, the proposed SF-CLIP fills the blank in fashion expressiveness studies and extends the capabilities of the CLIP model with respect to the assessment of fashion expressiveness. Extensive experiments demonstrate that SF-CLIP surpasses the state-of-the-art methods and sets a new record for the VSR task.

Acknowledgements

This research is supported by Hubei Key R&D Project (2022BAA033) and National Natural Science Foundation of China (62171325). The Supercomputing Center of Wuhan University supports the supercomputing resource.

Contribution Statement

The main contribution to this work is equally given by Likai Tian and Zhengwei Yang.

References

[Cheng *et al.*, 2017] Zhi-Qi Cheng, Xiao Wu, Yang Liu, and Xian-Sheng Hua. Video2shop: Exact matching clothes in videos to online shopping images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4048–4056, 2017.

[Chia *et al.*, 2022] Patrick John Chia, Giuseppe Attanasio, Federico Bianchi, Silvia Terragni, Ana Rita Magalhães, Diogo Goncalves, Ciro Greco, and Jacopo Tagliabue. Contrastive language and vision learning of general fashion concepts. *Scientific Reports*, 12(1):18958, 2022.

[Contributors, 2020] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020.

[Ge *et al.*, 2019] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5337–5345, 2019.

[Godi *et al.*, 2022] Marco Godi, Christian Joppi, Geri Skenderi, and Marco Cristani. Movingfashion: a benchmark for the video-to-shop challenge. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1678–1686, 2022.

[Gorti *et al.*, 2022] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5006–5015, 2022.

[Goyal *et al.*, 2023] Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19338–19347, 2023.

[Guo *et al.*, 2023] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 746–754, 2023.

[Hegde *et al.*, 2023] Deepti Hegde, Jeya Maria Jose Valanarasu, and Vishal Patel. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2028–2038, 2023.

[Ji *et al.*, 2019] Zhong Ji, Kailin Xiong, Yanwei Pang, and Xuelong Li. Video summarization with attention-based encoder–decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(6):1709–1717, 2019.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Li *et al.*, 2022a] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022.

[Li *et al.*, 2022b] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900, 2022.

[Liu *et al.*, 2019] Chih-Ting Liu, Chih-Wei Wu, Yu-Chiang Frank Wang, and Shao-Yi Chien. Spatially and temporally efficient non-local attention network for video-based person re-identification. *arXiv preprint arXiv:1908.01683*, 2019.

[Liu *et al.*, 2022] Tianrui Liu, Qingjie Meng, Jun-Jie Huang, Athanasios Vlontzos, Daniel Rueckert, and Bernhard Kainz. Video summarization through reinforcement learning with a 3d spatio-temporal u-net. *IEEE Transactions on Image Processing*, 31:1573–1586, 2022.

[Lu *et al.*, 2022] Haoyu Lu, Mingyu Ding, Nanyi Fei, Yuqi Huo, and Zhiwu Lu. Lgdn: Language-guided denoising network for video-language modeling. *Advances in Neural Information Processing Systems*, 35:25198–25211, 2022.

[Narasimhan *et al.*, 2021] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. *Advances in Neural Information Processing Systems*, 34:13988–14000, 2021.

[Porrello *et al.*, 2020] Angelo Porrello, Luca Bergamini, and Simone Calderara. Robust re-identification by multiple views knowledge distillation. In *European Conference on Computer Vision*, pages 93–110, 2020.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021.

[Rasheed *et al.*, 2023] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6545–6554, 2023.

[Sanghi *et al.*, 2022] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18603–18613, 2022.

[Veličković *et al.*, 2018] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[Wang *et al.*, 2023] Jinpeng Wang, Pan Zhou, Mike Zheng Shou, and Shuicheng Yan. Position-guided text prompt for vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23242–23251, 2023.

[Wu *et al.*, 2021] Dehao Wu, Yi Li, Yinghong Zhang, and Yuesheng Zhu. Multi-dimensional attentive hierarchical graph pooling network for video-text retrieval. In *2021 IEEE International Conference on Multimedia and Expo*, pages 1–6, 2021.

[Yan *et al.*, 2020] Yichao Yan, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Ying Tai, and Ling Shao. Learning multi-granular hypergraphs for video-based person reidentification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2899–2908, 2020.

[Yang *et al.*, 2023] Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, and Zheng Wang. Good is bad: Causality inspired cloth-debiasing for cloth-changing person reidentification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1472–1481, 2023.

[Zhou *et al.*, 2023] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185, 2023.