Cross-view Semantic Alignment for Livestreaming Product Recognition

Wenjie Yang*, Yiyi Chen*, Yan Li, Yanhua Cheng, Xudong Liu, Quan Chen†, Han Li
Kuaishou Technology
wenjie.yang@nlpr.ia.ac.cn, {chenyiyi,liyan26,chengyanhua,liuxudong,chenquan06,lihan08}@kuaishou.com

Abstract

*Live commerce is the act of selling products online through live streaming. The customer's diverse demands for online products introduce more challenges to Livestreaming Product Recognition. Previous works have primarily focused on fashion clothing data or utilize single-modal input, which does not reflect the real-world scenario where multimodal data from various categories are present. In this paper, we present LPR4M, a large-scale multimodal dataset that covers 34 categories, comprises 3 modalities (image, video, and text), and is 50× larger than the largest publicly available dataset. LPR4M contains diverse videos and noise modality pairs while exhibiting a long-tailed distribution, resembling real-world problems. Moreover, a c**R**oss-v**I**ew semanti**C** alignm**E**nt (RICE) model is proposed to learn discriminative instance features from the image and video views of the products. This is achieved through instance-level contrastive learning and cross-view patch-level feature propagation. A novel Patch Feature Reconstruction loss is proposed to penalize the semantic misalignment between cross-view patches. Extensive experiments demonstrate the effectiveness of RICE and provide insights into the importance of dataset diversity and expressivity. The dataset and code are available at https://github.com/adxcreative/RICE.*

1. Introduction

Livestreaming Product Recognition (LPR) [3, 9, 11] is one of the significant machine learning application in the e-commerce industry. Its goal is to recognize products a salesperson presents in a live commerce clip through content-based video-to-image retrieval. The real-time and accurate recognition of livestreaming products can facilitate the online product recommendation, and thereby improve the purchasing efficiency of consumers.

The task of LPR involves two fundamental processes: multimodal-based intended product identification and shop product retrieval. This task poses significant challenges in real-world scenarios, including (1) the need to distinguish *intended products* from the cluttered background products in a livestreaming frame, exemplified in Fig. 1 (b), (2) the requirement for models to capture sufficient *fine-grained* features to match the ground-truth (GT) image accurately in the shop, where there are many images with subtle visual nuances, (3) the *heterogeneous* video-to-image and *cross-domain* livestream-to-shop problem, and (4) the *appearance changes* of products in the livestreaming domain due to articulated deformations, occlusions, diverse background clutters, and significant illumination variations, making it a highly intricate task to match the clip to the GT image in the shop. Various datasets have emerged in the computer vision community to study this task, including AsymNet [3], WAB[1], and MovingFashion [9]. However, AsymNet and MovingFashion lack crucial text modal, which provides essential auxiliary information for identifying intended products. Furthermore, the data scale of WAB is relatively small, with only 70K pairs, and only provides fashion clothing data, diverging from the real-world scenario.

In order to narrow the gap between existing datasets

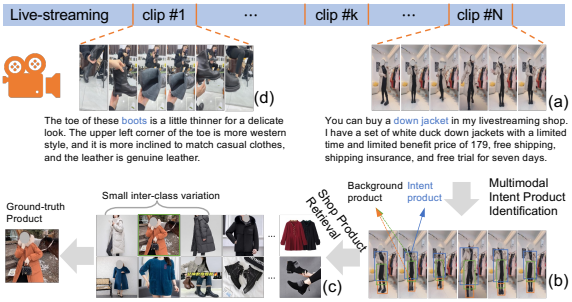
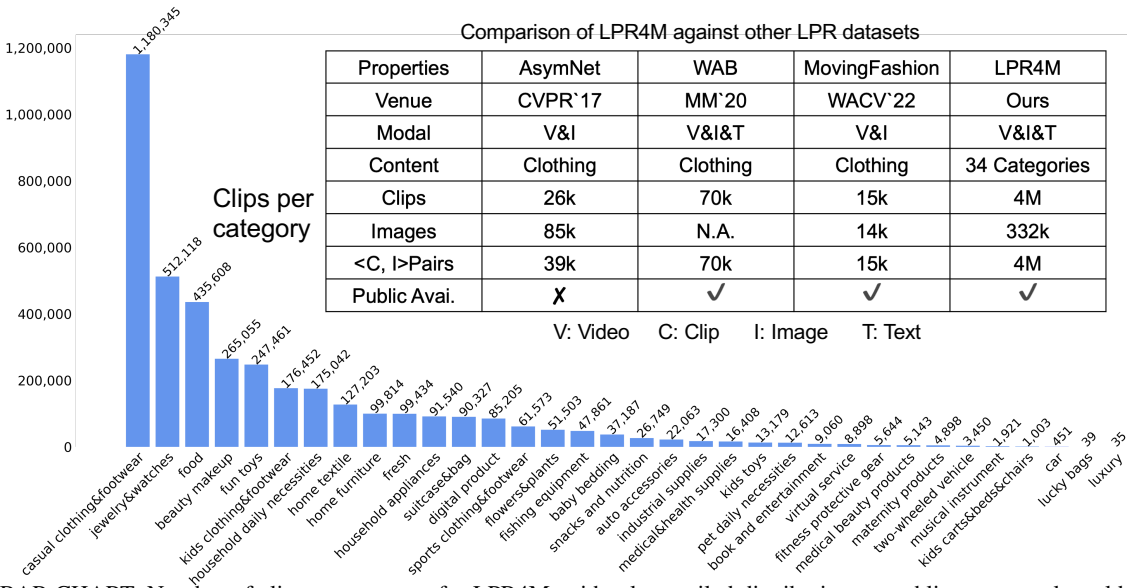


Figure 1. The pipeline of LPR. A livestreaming consists of many clips introducing different products. We show two clip examples with ASR text in (a) and (d). In (b), the intended product refers to the product the salesperson is introducing, and the other products on the screen are indicated as the distracted background products. (c) presents a shop with hundreds of images, some with subtle visual differences called small inter-class variations. The LPR aims to identify the clip's intended product using the ASR text prompt, then retrieve the ground-truth product from the shop images.

*Equal contribution
†Corresponding author

[1]https://tianchi.aliyun.com/competition/entrance/231772/information



Comparison of LPR4M against other LPR datasets

Properties	AsymNet	WAB	MovingFashion	LPR4M
Venue	CVPR`17	MM`20	WACV`22	Ours
Modal	V&I	V&I&T	V&I	V&I&T
Content	Clothing	Clothing	Clothing	34 Categories
Clips	26k	70k	15k	4M
Images	85k	N.A.	14k	332k
<C, I>Pairs	39k	70k	15k	4M
Public Avai.	✗	✓	✓	✓

V: Video C: Clip I: Image T: Text

Figure 2. BAR CHART: Number of clips per category for LPR4M, with a long-tailed distribution resembling most real-world problems. TABLE: Comparison of LPR4M against other LPR datasets in terms of modal, content, and scale. LPR4M offers significantly broader coverage of live commerce product categories and several orders of magnitude larger data scales.

and the real-world scenario and advance research in this challenging task, we present LPR4M, a large-scale multimodal live commerce dataset that includes extensive categories, diverse data modalities of clip, image, and text, as well as heterogeneous and cross-domain correspondences of $\langle clip, image \rangle$ pairs. This dataset offers several significant advantages. (1) *Large-Scale:* LPR4M contains over 4M pairs, significantly exceeding its precedents. (2) *Expressivity:* LPR4M draws data pairs from 34 commonly used live commerce categories rather than relying solely on clothing data. Additionally, LPR4M offers auxiliary clip ASR text and image title modalities, which are critical for intended product identification and product feature representation. (3) *Diversity:* LPR4M promotes clip diversity while preserving the real-world data distribution, with a focus on three components: product scale, visible duration, and the number of products in the clip, as depicted in Fig. 3. To the best of our knowledge, LPR4M is currently the largest dataset created explicitly for real-world multimodal LPR scenarios.

Our work based on LPR4M tackles a realistic problem: *how to achieve fine-grained LPR using large-scale multimodal pairwise data?* Given image and clip views, we first utilize Instance-level Contrastive Learning (ICL) to align global features. However, since instance features of these two views are extracted independently from the visual encoder, it can be challenging to differentiate between products with subtle visual differences without cross-view interactions. Consequently, we propose a patch-level semantic alignment approach to enable cross-view patch information propagation. We suggest measuring similarity via a cross-attention based Pairwise Matching Decoder (PMD), which treats image patches as *Query* and video patches as both *Key* and *Value*. In addition, we propose a novel Patch Feature Reconstruction (PFR) loss to provide patch-level supervision for pairwise matching, expecting to reconstruct each feature of an image patch from its paired video patches.

The main contributions of this paper can be summarized as follows. (1) A large-scale live commerce dataset is collected, offering a significantly broader coverage of categories and diverse modalities such as video, image, and text. This dataset is the most extensive one known to date, tailored explicitly for real-world multimodal LPR scenarios. (2) The RICE model is introduced to integrate instance-level contrastive representation learning and patch-level pairwise matching into a framework. (3) A novel Patch Feature Reconstruction loss is proposed to penalize the semantic misalignment between patches of video and image. (4) The benchmark dataset and evaluation protocols are carefully defined for LPR. Extensive experiments demonstrate the effectiveness of LPR4M and RICE.

2. Related Works

LPR datasets. As shown in the table of Fig. 2, we compare LPR4M with others in terms of modality, content, and scale. In particular, AsymNet [3], WAB, and MovingFashion [9] only provide fashion clothing data, and the text modality is absent in AsymNet and MovingFashion. Therefore, We collect LPR4M, which covers 34 widely used categories and provides visual and text modalities. LPR4M is 50× larger than WAB.

Video Object Detection (VOD). The main focus of recent VOD methods [38, 37, 4, 2, 26, 35] is exploiting temporal information to tackle the video variations, *e.g.*, oc-

clusion, motion blur and out of focus. The temporal relationship mining insights in VOD inspire this paper's design of the intended product detection module. However, unlike VOD, where most videos only contain a single object, the videos in LPR contain many cluttered background products. This situation suggests that in the absence of prompt text information, it is more challenging to identify the intended product by only relying on the visual inputs. Therefore, we explore the fusion of text and visual modalities and verify its effectiveness in this paper.

Fine-Grained Vision Recognition (FGVR). Similar to FGVR [6, 34, 5, 16, 7, 36, 30, 29], the LPR aims to learn discriminative instance feature to distinguish the subclasses with large intra-class and small inter-class variations. However, unlike traditional FGVR, each live commerce category contains an enormous amount of subclasses in LPR. Moreover, the number of subclasses will increase or decrease dynamically as a large number of products are newly added or taken off the shop every day. It makes it more challenging to handle the out-of-distribution subclasses.

Video-to-Shop Retrieval. Although fashion retrieval has made great progress [13, 19, 8, 15], there are few studies focus on retrieving products that are presented in e-commerce videos, referred to as *video-to-shop*. AsymNet is a *one-stage* method without detection. It employs LSTM to exploit temporal continuity in the video, then perform pair-wise matching by feeding the image and video feature into a similarity network. DPRNet [33] and SEAM Match-RCNN [9] adopt a *two-stage* pipelines. DPRNet first detects the products in the video and then performs image-to-image retrieval. SEAM Match-RCNN performs self-attention among the detected product boxes in a video to produce a video feature and uses the inner product between the video and image feature as a similarity. In this paper, we propose RICE integrates the *one-stage* and *two-stage* methods into a framework and study their advantages.

3. Dataset and Benchmark

In this section, we present the construction, characteristics, and benchmark of LPR4M.

Overview. Compared with other existing LPR datasets, LPR4M has several appealing properties, which are summarized in the following. (1) *Large-Scale.* As illustrated in the table of Fig. 2, LPR4M is the largest LPR dataset to date. It contains 4M exactly matched $\langle clip, image \rangle$ pairs of 4M live clips, and 332k shop images. Each image has 14.5 clips with different product variations, *e.g.*, viewpoint, scale, and occlusion. The example of image-to-clips and the number of clips per image are shown in Fig. 4 (d) and (b), respectively. Specifically, most of the images (80%) have ten matched clips and the number of clips per image range from 10 to 150. (2) *Expressivity.* The expressivity of LPR4M is mainly reflected in two aspects. Firstly, unlike other LPR datasets that only contain fashion clothing data, our data is more affluent, coming from 34 categories covering most of the daily necessities. This makes it closer to the real scenario. Secondly, the data of LPR4M is multimodal. We provide live clip ASR texts and shop image titles as auxiliary information to facilitate the intended product identification and form a full-scale characteristic of each product. (3) *Diversity.* Firstly, we collect clips according to the clip duration distribution of real livestreaming scenarios and obtain the clips with various durations, as shown in Fig. 4 (a). Secondly, the clips are further sampled by controlling the variation in terms of three properties, *i.e.*, product scale, intended product visible duration, and the number of products in the clip. It makes LPR4M a challenging benchmark. As illustrated in Fig. 3, we pick two categories to represent a variation. For each category row, the clip shows three different levels of difficulty progressively.

3.1. Data Collection and Cleaning

The basic unit of the dataset is a $\langle clip, image \rangle$ pair. All the clips are cut from hours of sequential livestreaming data crawled from Kuaishou[2]. A livesteaming has a unique online shop, which lists all the products to be introduced in this livestreaming. Firstly, we removed near- and exact-duplicate images in the shops by comparing the global average pooled *layer4* features after feeding them into ResNet [12]. Secondly, the human annotators cleaned the clips that contain the target product with short visible duration, small scales, and severe background clutters. Finally, given a clip, the matched product image is picked from the shop via the human annotator. In total, 4,033,696 clips and 398,796 images are kept to construct the training and test set of LPR4M.

Variations. The proportion of the number of clips in each variation is depicted in Fig. 4 (c). (1) *Scale.* According to the proportion (p) of the product box area to the entire frame area, the clips are classified into three subsets. The area is measured as the number of pixels in the product box. In LPR4M, there are more small products than large products. Specifically, approximately 54.5% of products are small ($p \leq 0.2$), 30.5% are medium ($0.2 < p \leq 0.4$), and 15% are large ($p > 0.4$). As shown in the first row of Fig. 3, the coat in the first clip is displayed in a zoom-in view, where the coat is large and overflows the frame. However, the physical size of the shoes in the third clip is relatively small. Because the considerable distance between the shoes and the camera, the scale is visually small. (2) *Visible duration.* Due to occlusions and changes in camera perspective, the target product is not always visible in the clip. Here, each clip is categorized by the proportion of visible duration to the entire clip duration, including 48.5% of long ($0.7 < p$), 29.6% of medium ($0.4 < p \leq 0.7$) and 21.9% of

[2]https://live.kuaishou.com

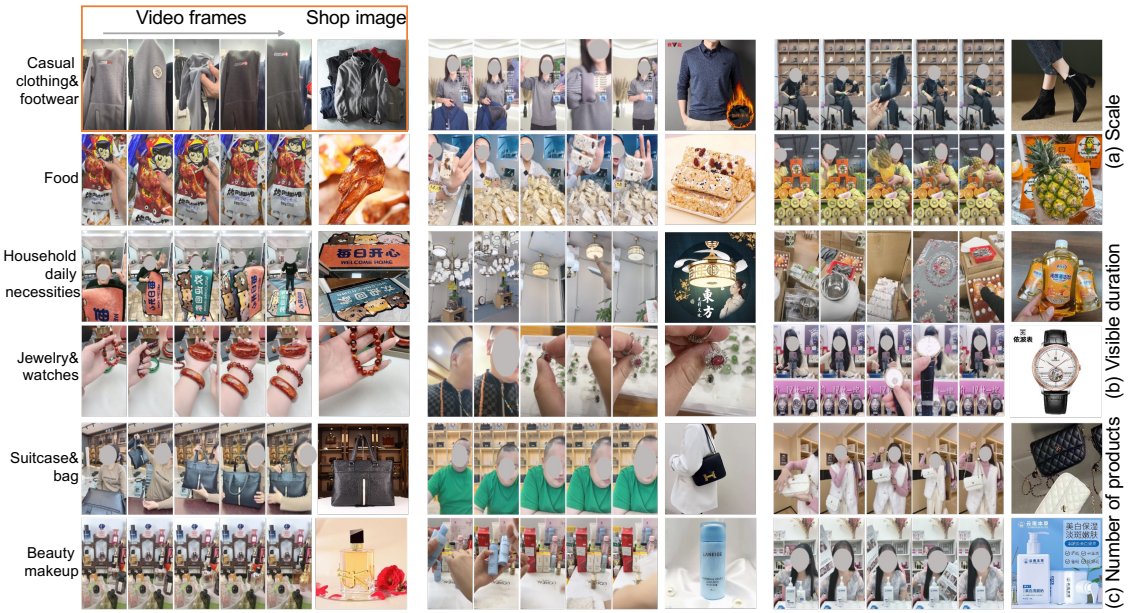


Figure 3. The $< clip, image >$ pairs of LPR4M. As shown on the left of the orange box, we extract five evenly spaced frames from the clip, with the shop image on the right. We choose two categories to illustrate one of the clip product variations of *scale*, *visible duration*, and *number of products*. Each row shows 3 data pairs for different degrees of difficulty of the corresponding variation, including (a) *large*, *medium* and *small* product scale, (b) *long*, *medium* and *short* visible duration, (c) *abundant*, *medium* and *few* products in the clip.

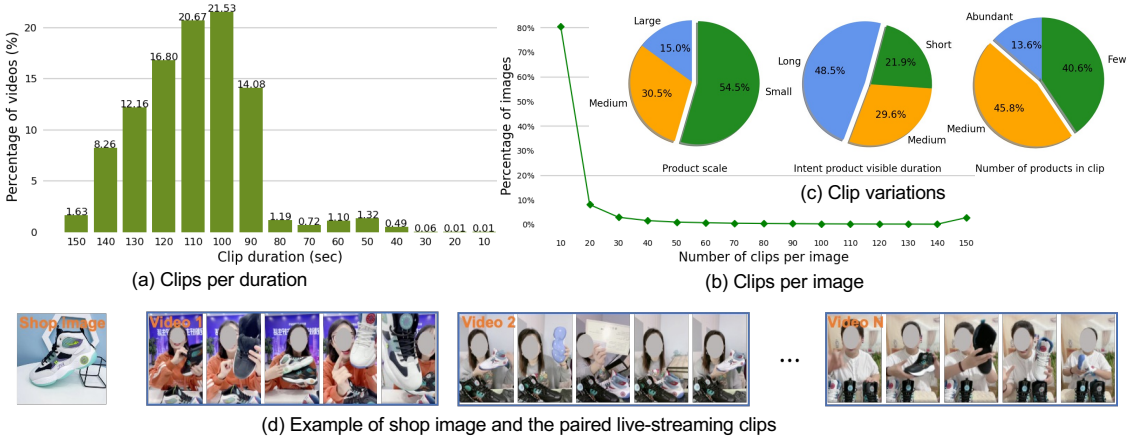


Figure 4. LPR4M statistics of clips. (a) The distribution of clip duration. (b) The number of clips per image. Approximately 80% of the images have ten paired clips. (c) The statistics of three clip variations. (d) The shop image, the paired clips, and the products in the clips suffer from different variations, *e.g.*, scale, viewpoint, and occlusion.

short ($p \leq 0.4$). For example, in the third clip of the fourth row in Fig. 3, the watch is occluded at the beginning and end of the clip, which significantly increases the difficulty of LPR. Note that the visible duration of the intended product is evaluated by the annotators. (3) *Background distractor.* In the livestreaming of beauty makeup, handbags, and jewelry, *etc.*, there are abundant products displayed on the screen. For example, the first clip in the last row of Fig. 3 contains more than two dozen perfumes. However, there is only one intended product in a clip, and it is challenging to distinguish the intended product from the distracted background products. Therefore, we asked the annotators to assess the number (n) of products in the clip (or background distractor) and accordingly classify the clips into three subsets, including 13.6% of abundant ($n > 7$), 45.8% of medium ($3 < n \leq 7$) and 40.6% of few ($n \leq 3$).

Clip Description and Image Title. In the case of a clip containing multiple products, it is ambiguous for the model to predict whether the clip and an image match based on visual information only. Therefore, we additionally provide clip descriptions and image titles to enrich the dataset. On the one hand, benefiting from promising results achieved by the Transformer and Convolution Neural Network based models in ASR [10, 32, 31], we adopt Conformer [10], a SOTA method on the widely used LibriSpeech benchmark [23], to extract text description from the clip voices.

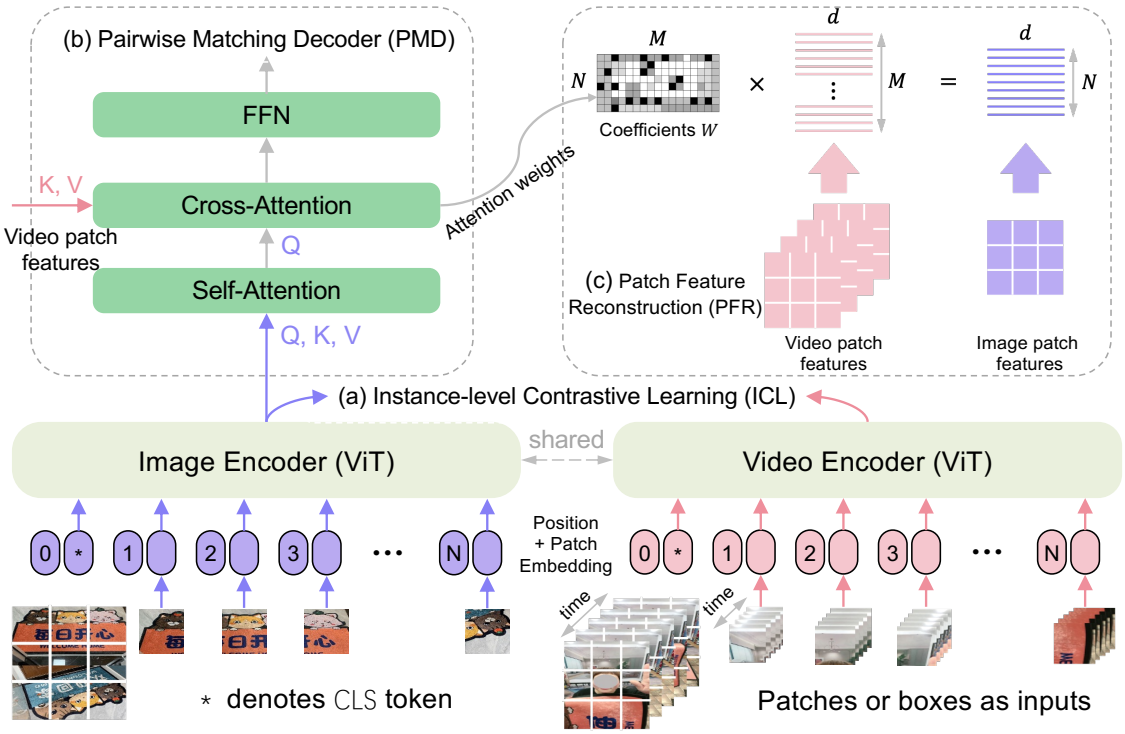


Figure 5. Overview of RICE framework. Given two views of the same product, *i.e.*, image and video, the patch features are extracted with the transformer encoder. In (a), RICE first performs contrastive loss on the global image and video features. Then, in (b), RICE employs transformer-based fusion model to perform patch-level feature interaction between the two views. The optimization of PMD aims to decrease the similarity of the two views from different products and increase that from the same product. Furthermore, in (c), RICE exploits reconstruction loss to penalize the misalignment between the semantic patches in the two views, which expects each patch feature of the image to be reconstructed from the patch features of the paired video. Best viewed in color.

On the other hand, the titles of product images are provided by the merchant and are available on the video website.

3.2. Livestreaming Product Recognition

As shown in Fig. 1, this task is to retrieve the GT images from the shop (gallery) for each livestreaming clip (query) and has been considered by several previous works [13, 11, 19, 8, 3, 9]. It emphasizes the retrieval performance and considers the impact of intended product identification. Specifically, a query clip is counted as missed if the intended product fails to be identified. Rank-k retrieval accuracy is used to measure retrieval performance, such that a successful retrieval is counted if the GT image has been retrieved in the rank-k results.

Splitting training and test set. In order to evaluate different methods, we split the training and test set and ensure the products in the training and test set are non-overlapping. The training and test sets contain 4,013,617/332,438 and 20,079/66,358 clips/images, respectively.

Intended product box annotation. To enable effective supervision of the detector training and evaluation of the detection accuracy, we annotate the intended product box for both the training and test set. For the training set, we sample 2% of the clips for intended box annotation and extract 10 frames at even intervals. For each test clip, we extract one frame every 3 seconds. The detection training/test set contains 1,120,410 /501,656 frames with 1,115,629/669,374 intended product boxes, respectively.

4. Method

This section presents the technical details of RICE. As shown in Fig. 5, the RICE first performs instance-level contrastive learning to learn discriminative feature for the product (Sec. 4.1). Then, we introduce PMD that pursues fine-grained similarity measurement by conducting patch-level feature propagation (Sec. 4.2). The PMD is further guided by the novel PFR loss to promote patch-level semantic alignment (Sec. 4.2). Finally, we study the impact of product location by replacing the input patches with the product boxes produced by intended product detector (Sec. 4.3).

4.1. Instance-level Contrastive Learning (ICL)

Let $\mathcal{V}$ be a set of livestreaming clips (or videos). Let $\mathcal{I}$ be a set of shop images. The objective of RICE is to learn a function to measure the similarity between the clip $\mathcal{V}_i \in \mathcal{V}$ and the image $\mathcal{I}_i \in \mathcal{I}$. Formally, taking $\mathcal{V}_i$ and $\mathcal{I}_i$ as input, the image encoder first splits the image into non-overlapping patches, which are projected into 1D tokens via

a linear projection. Then the transformer layers are used to extract the patch features, denoted as $\{i_{cls}, i_1, ..., i_N\}$. Likewise, the video encoder processes each video frame $\mathcal{V}_i^j$ independently and outputs a sequence of video patch features $\{v_{cls}, v_1, ..., v_M\}$, where j is the index of frame number $|\mathcal{V}_i|$ and $M = N \times |\mathcal{V}_i|$. Given an image with a resolution of 224×224 and patch size of 32×32, we have $N = 49$. Note that the image and video encoder share parameters. Following ViT and CLIP, we extract the global representation from the [CLS] token. In order to pull the clip and image of the same product while pushing away that of different products in the feature space, we perform InfoNCE loss [22] on the global representation, defined as:

$$\mathcal{L}_{nce} = -\mathbb{E}_{p(\mathcal{I},\mathcal{V})}\left[\log \frac{\exp(g_\theta(\mathcal{I}_i, \mathcal{V}_i))}{\sum_{\widetilde{\mathcal{V}}_k \in \widetilde{\mathcal{V}}} \exp(g_\theta(\mathcal{I}_i, \widetilde{\mathcal{V}}_k))}\right], \quad (1)$$

where $g_\theta(\mathcal{I}_i, \mathcal{V}_i) = g_\mathcal{I}(i_{cls})^\top g_\mathcal{V}(v_{cls})/\tau$ and $\widetilde{\mathcal{V}}$ consists of a positive sample $\mathcal{V}_i$ and $|\widetilde{\mathcal{V}}| - 1$ negative samples. $g_\mathcal{I}$ and $g_\mathcal{V}$ are transformations that map the [CLS] embedding of image and clip, *i.e.*, i_{cls} and v_{cls}, to the normalized lower-dimensional features. τ is a temperature parameter, and we use $\tau = 0.01$. The final contrastive loss between the image and clip is a symmetric version of $\mathcal{L}_{nce}$, given by:

$$\mathcal{L}_c = -\frac{1}{2}\mathbb{E}_{p(\mathcal{I},\mathcal{V})}[\log \frac{\exp(g_\theta(\mathcal{I}_i, \mathcal{V}_i))}{\sum_{\widetilde{\mathcal{V}}_k \in \widetilde{\mathcal{V}}} \exp(g_\theta(\mathcal{I}_i, \widetilde{\mathcal{V}}_k))} + \log \frac{\exp(g_\theta(\mathcal{I}_i, \mathcal{V}_i))}{\sum_{\widetilde{\mathcal{I}}_k \in \widetilde{\mathcal{I}}} \exp(g_\theta(\widetilde{\mathcal{I}}_k, \mathcal{V}_i))}], \quad (2)$$

where $|\widetilde{\mathcal{I}}| = |\widetilde{\mathcal{V}}|$ is the batch size.

4.2. Patch-level Semantic Alignment

Pairwise Matching Decoder (PMD). It is straightforward to exploit $g_\theta(\mathcal{I}_i, \mathcal{V}_i)$ in Eq. (1) as a measure of the similarity. However, the features of the clip and image are extracted independently from the visual encoder, without information propagation between $\mathcal{I}_i$ and $\mathcal{V}_i$. To this end, we perform patch-wise feature attention via a transformer decoder layer, named pairwise matching decoder, which consists of a self-attention and a cross-attention layer in this paper. As illustrated in Fig. 5 (b), the self-attention layer takes the image patch features as the *Query*, *Key* and *Value*, and the cross-attention layer takes the image patch features as *Query* while takes the video patch features as *Key* and *Value*. The matching loss of the similarity calculator is defined as follows:

$$\mathcal{L}_m = -\frac{1}{2}\mathbb{E}_{p(\mathcal{I},\mathcal{V})}[\log \frac{\exp(f_\theta(\mathcal{I}_i, \mathcal{V}_i))}{\sum_{\hat{\mathcal{V}}_k \in \hat{\mathcal{V}}} \exp(f_\theta(\mathcal{I}_i, \hat{\mathcal{V}}_k))} + \log \frac{\exp(f_\theta(\mathcal{I}_i, \mathcal{V}_i))}{\sum_{\hat{\mathcal{I}}_k \in \hat{\mathcal{I}}} \exp(f_\theta(\hat{\mathcal{I}}_k, \mathcal{V}_i))}], \quad (3)$$

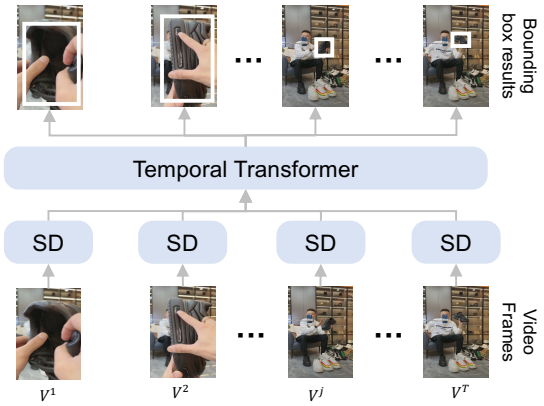


Figure 6. Illustration of the Single-Frame Detector (SD) and Multi-Frame Detector (MD). The MD exploits a temporal transformer to fuse the results from the SD.

where $f_\theta(\mathcal{I}_i, \mathcal{V}_i) = v^\top x_{cls}(\mathcal{I}_i, \mathcal{V}_i)$, $x_{cls}(\mathcal{I}_i, \mathcal{V}_i)$ is the [CLS] embedding of the decoder layer and v is a parametric vector. Here, we only sample N_{neg} negative instances for each GT $(\mathcal{I}_i, \mathcal{V}_i)$ pair, *i.e.*, $\hat{\mathcal{V}}$ consists of a positive sample $\mathcal{V}_i$ and N_{neg} negative samples.

Patch Feature Reconstruction (PFR). Furthermore, we pursue cross-view semantic alignment by searching similar patches in the clip to reconstruct the coupled image in the feature space. Here, we introduce how to perform patch feature reconstruction given a positive data point of two views of clip $\mathcal{V}_i$ and image $\mathcal{I}_i$. As shown in Fig. 5 (b), let

$$X = \{v_1, ..., v_M\} \in \mathbb{R}^{d \times M}$$

be the patch features of the clip, where $v_m \in \mathbb{R}^{d \times 1}$. Likewise, the patch features of the image are denoted as:

$$Y = \{i_1, ..., i_N\} \in \mathbb{R}^{d \times N}.$$

Then, the i_n can be represented by a linear combination of X. The insight behind this is that the image can be reconstructed from the clip if the clip contains the product in the image. Therefore, we solve for the coefficients $w_n \in \mathbb{R}^{M \times 1}$ of i_n with respect to X. Finally, the reconstruction loss is defined as:

$$\mathcal{L}_r = ||Y - XW||_F^2. \quad (4)$$

Since the attention weights a in the cross-attention layer indicate the correspondences between patches of the two views, it is intuitive to learn the reconstruction coefficients W from a. Specifically, the $a \in \mathbb{R}^{8 \times N \times M}$ is fed into two consecutive sets of convolution and ReLU layers to output the coefficients $W \in \mathbb{R}^{N \times M}$.

The final objective function for the RICE model is the weighted summation of $\mathcal{L}_c$, $\mathcal{L}_m$ and $\mathcal{L}_r$, given by:

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_m + \alpha\mathcal{L}_r, \quad (5)$$

where α is the trade-off weight and we use $\alpha = 0.1$ in the following experiments.

Methods	overall	scale			visible duration			number of product		
		small	medium	large	short	medium	long	abundant	medium	few
ICL_{patch}	27.1	23.9	**34.4**	30.0	23.6	28.8	**36.3**	20.2	25.0	**27.5**
ICL_{box}	30.0	30.1	**35.8**	31.1	26.6	30.1	**37.1**	17.2	29.2	**30.6**
RICE_{patch}	31.2	28.9	**37.0**	32.7	28.1	32.9	**39.6**	21.0	34.6	**31.5**
RICE_{box}	33.0	32.7	**39.0**	33.8	29.6	34.8	**42.0**	17.6	31.9	**33.4**

Table 1. The R1 performances of the ICL and RICE model. Results of evaluation on different input types, *i.e.*, patch and detected box, are shown in each row. The columns show the results on different test subsets split by video variations, *i.e.*, *product scale*, *visible duration*, *number of products*. The best performance for each subset is in bold.

4.3. Intended Product Detection (IPD)

In order to highlight the intended products in video and suppress the background products, we propose replacing patch inputs with the detected intended product boxes for the videos. As shown in Fig. 6, we adopt DAB-DETR [18] and TransVOD Lite [35] as the SD and MD, respectively. The SD detects products frame-by-frame. Given the T frames and one box label per frame, the SD is trained to predict an intended product box for each frame. However, detecting products with significant appearance changes using only a single frame can be challenging, as exemplified by the *shoe* with *small scale* in $\mathcal{V}^T$ in Fig. 6. Therefore, MD leverages a temporal transformer to capture product interactions in the temporal context and predict a more accurate box for each frame. For more details about the IPD, please refer to the supplementary material.

5. Experiment

5.1. Dataset and Evaluation Metrics

Experiments are performed on the LPR4M testset, which contains 20,079 livestreaming clips as query set and 66,358 shop images as gallery set, as described in Sec. 3.2. We adopt rank-*k* accuracy as the retrieval performance metrics.

5.2. Implementation Details

Model. The image and video encoder share parameters and are initialized with ViT-B/32 from CLIP [25], where the number of layers is 12 and the patch size is 32. Likewise, we initialize PMD with the similar parameters from CLIP. **Preprocessing.** We extract 10 evenly spaced frames from each clip as the video input. The images and video frames are resized to 224×224. For data augmentation, we randomly mask video frames with a percentage ranging from 0 to 0.9 and a probability of 0.5. **Optimization.** We use Pytorch [24] to implement the RICE model. The Adam [14] optimizer is used with a batch size of 256. For the learning rate, we decay it using a cosine schedule [21] following CLIP. The initial learning rate is 1e-7 for the image encoder and video encoder and 1e-4 for the newly introduced modules. All the experiments are carried out on 8 NVIDIA Tesla V100 GPUs, which takes about 90 hours for 3 epochs.

Dataset	Methods	R1	R5	R10
LPR4M	FashionNet [19]	13.4	33.8	50.4
	AsymNet [3]	22.0	46.7	63.8
	SEAM [9]	23.3	49.5	61.4
	NVAN [17]	21.4	45.2	62.7
	TimeSFormer [1]	28.6	56.8	69.0
	SwinB [20]	29.1	60.1	73.9
	RICE (Ours)	33.0	65.5	77.3
MF	NVAN [17]	38.0	62.0	70.0
	MGH [27]	40.0	59.0	66.0
	AsymNet [3]	42.0	73.0	86.0
	SEAM [9]	49.0	80.0	89.0
	RICE (Ours)	76.1	89.7	92.6

Table 2. The LPR4M and MovingFashion (MF) evaluation.

5.3. Impact of Video Variations

In this section, we carry out experiments to study the impact of video variations, *i.e.*, *product scale*, *visible duration*, *number of products*, as shown in Fig. 3. We evaluate two input types, *i.e.*, patch and detected box, for each model. The results are reported in Table. 1. As we can see, the performance declines when small scale, short visible duration, and abundant products are presented. (1) Compared to the patch input, the box input significantly improves the accuracy. For example, ICL_{box} outperforms ICL_{patch} by 6.2% on *small* split. Besides, ICL_{box} significantly reduces the performance gap between *small* and *medium* split. It indicates the IPD improves the robustness to scale variation. (2) As the performances on *abundant* split shows, the model with *box* input achieves lower accuracy than *patch* input, because it is challenging for the detector to distinguish the indent product from the abundant background products.

5.4. Comparison with state-of-the-art methods

In this section, we compare our RICE with state-of-the-art (SOTA) methods on LPR4M and MovingFashion (MF), except AsymNet [3] and WAB because AsymNet is not public available and WAB is a competition dataset with only Chinese introduction. The results are shown in Table 2. 1) On LPR4M, the FashionNet, AsymNet and SEAM are LPR methods and the others are video understanding methods. As we can see, our RICE surpasses not only the LPR meth-

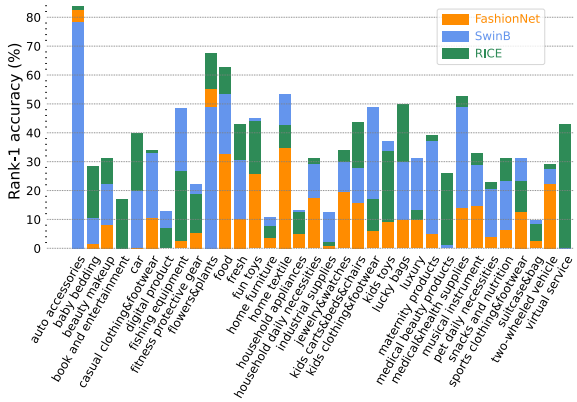


Figure 7. Per-category rank-1 performance on the 34 categories.

#	ICL	PMD	PFR	IPD	Txt	R1	R5	R10
a	✓					27.1	56.4	68.3
b	✓				✓	28.5	58.9	71.5
c	✓	✓				29.4	62.0	73.7
d	✓	✓	✓			30.3	62.7	74.0
e	✓	✓	✓	✓		31.3	63.2	74.3
f	✓	✓	✓	✓	✓	33.0	65.5	77.3

Table 3. Ablation study on the key components, *i.e.*, ICL: instance-level contrastive learning, PMD: pairwise matching decoder, PFR: patch feature reconstruction, IPD: intended product detection, Txt: text modality. The rank-k accuracy is reported.

ods but also the strong video understanding methods. 2) On MF, the NVAN and MGH are video understanding methods. Our approach achieves the best accuracy.

5.5. Ablation Study

In this section, we investigate the impact of each component of our approach by conducting ablation experiments. The results are reported in Table. 3. (c) Compared to the baseline ICL, the PMD obtains the R1 performance gains of 2.3% (29.4 to 27.1), which demonstrates the superiority of patch-level (local) over instance-level (global) similarity measurement. (d) The patch-level supervision provided by PFR facilitates semantic alignment and results in a considerable improvement of 0.9% for R1. (e) Our IPD replacing patch inputs with detected intended boxes significantly outperforms ICL by 1.0% R1 as it enables the model to focus on informative regions while suppressing distractions. In (b) and (f), the addition of text modality increases the R1 from 27.1% to 28.5% and 31.3% to 33.0%, respectively. It is because the text helps suppress the distracted background products. Here, the ChineseCLIP [28] is used to extract the embeddings of video ASR and image titles. The text similarity is computed as the dot product of normalized features. Then we combine the text and visual similarities to obtain the final $\langle clip, image \rangle$ similarity via addition.

5.6. Per-category performance

As shown in Fig. 7, we compare the rank-1 accuracy of FashionNet [19], SwinB [20] and our RICE on all 34 categories. Our RICE consistently outperforms FashionNet on all categories, and outperforms SwinB on most of the categories. Due to RICE of averaging the features of frames as video features, temporal information is not effectively utilized. But SwinB introduces 3D shifted windows to preserve temporal dynamics. As a result, SwinB performs well on certain categories with occlusions or view changes, *e.g.*, Suitcase&bag, as shown in the 5-th row of Fig. 3. The SwinB provides a promising way to enhance our model.

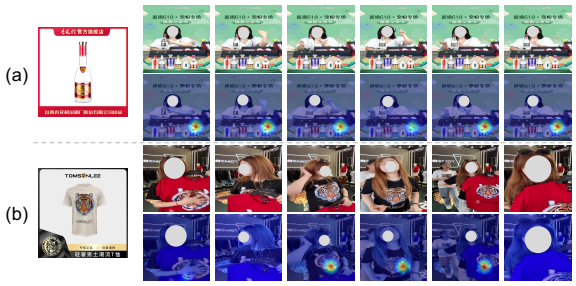


Figure 8. Attention map visualization of RICE on the LPR4M testset. The first column displays the shop images. We show the raw frames and the corresponding attention maps for each video.

5.7. Attention Region Visualization

To provide insight into PMD, we conduct further visualization. In Fig. 8, we show the attention map of RICE$_{patch}$ between shop image and video patches, where an image is regarded as the query, and attention weights on all spatial patches are visualized. We use the attention weights in the cross-attention layer of PMD for visualization. We make the following observations. (1) For the complex scenarios like (a) in Fig. 8, our approach can distinguish the target *Chinese liquor* from the nearby background *liquors*. (2) Interestingly, as shown in (b) of Fig. 8, even the target product is not always visible in the video, our approach still focuses on the corresponding regions accurately while pays less attention to the occluded regions.

6. Conclusions

In this paper, we present a large-scale dataset that offers broader coverage of categories and more sufficient data modalities named LPR4M. Moreover, the RICE model is proposed to integrate instance-level contrastive learning and patch-level cross-view semantic alignment mechanism into a framework. The extensive experiments demonstrate the effectiveness of the proposals clearly and show that additional performance gains can be achieved via integrating intended product detection and text modality. In this work, we show that it is a promising way to enhance the LPR model from the aspect of large-scale multimodal training. We hope the proposed LPR4M and the RICE baseline can spur further investigation into the LPR task.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proc. ICML*, volume 2, page 4, 2021.

[2] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *Proc. CVPR*, pages 10337–10346, 2020.

[3] Zhi-Qi Cheng, Xiao Wu, Yang Liu, and Xian-Sheng Hua. Video2shop: Exact matching clothes in videos to online shopping images. In *Proc. CVPR*, pages 4048–4056, 2017.

[4] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. Relation distillation networks for video object detection. In *Proc. ICCV*, pages 7023–7032, 2019.

[5] Yao Ding, Yanzhao Zhou, Yi Zhu, Qixiang Ye, and Jianbin Jiao. Selective sparse sampling for fine-grained image recognition. In *Proc. ICCV*, pages 6599–6608, 2019.

[6] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proc. CVPR*, pages 4438–4446, 2017.

[7] Weifeng Ge, Xiangru Lin, and Yizhou Yu. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In *Proc. CVPR*, pages 3034–3043, 2019.

[8] 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 *Proc. CVPR*, pages 5337–5345, 2019.

[9] Marco Godi, Christian Joppi, Geri Skenderi, and Marco Cristani. Movingfashion: a benchmark for the video-to-shop challenge. In *Proc. WACV*, pages 1678–1686, 2022.

[10] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *Proc. Interspeech*, pages 5036–5040, 2020.

[11] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. Where to buy it: Matching street clothing photos in online shops. In *Proc. ICCV*, pages 3343–3351, 2015.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016.

[13] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proc. ICCV*, pages 1062–1070, 2015.

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

[15] Zhanghui Kuang, Yiming Gao, Guanbin Li, Ping Luo, Yimin Chen, Liang Lin, and Wayne Zhang. Fashion retrieval via graph reasoning networks on a similarity pyramid. In *Proc. ICCV*, pages 3066–3075, 2019.

[16] Yan Li, Junge Zhang, Jianguo Zhang, and Kaiqi Huang. Discriminative learning of latent features for zero-shot recognition. In *Proc. CVPR*, pages 7463–7471, 2018.

[17] 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.

[18] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. In *Proc. ICLR*, 2021.

[19] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proc. CVPR*, pages 1096–1104, 2016.

[20] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proc. CVPR*, pages 3202–3211, 2022.

[21] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[23] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210. IEEE, 2015.

[24] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. *Proc. NIPS Workshops*, 2017.

[25] 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 *Proc. ICML*, pages 8748–8763. PMLR, 2021.

[26] Guanxiong Sun, Yang Hua, Guosheng Hu, and Neil Robertson. Mamba: Multi-level aggregation via memory bank for video object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2620–2627, 2021.

[27] Yichao Yan, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Ying Tai, and Ling Shao. Learning multi-granular hypergraphs for video-based person re-identification. In *Proc. CVPR*, pages 2899–2908, 2020.

[28] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*, 2022.

[29] Wenjie Yang, Houjing Huang, Xiaotang Chen, and Kaiqi Huang. Bottom-up foreground-aware feature fusion for practical person search. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):262–274, 2022.

[30] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *Proc. CVPR*, pages 1389–1398, 2019.

[31] Ching-Feng Yeh, Jay Mahadeokar, Kaustubh Kalgaonkar, Yongqiang Wang, Duc Le, Mahaveer Jain, Kjell Schubert, Christian Fuegen, and Michael L Seltzer. Transformer-transducer: End-to-end speech recognition with self-attention. *arXiv preprint arXiv:1910.12977*, 2019.

[32] Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar. Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7829–7833. IEEE, 2020.

[33] Hongrui Zhao, Jin Yu, Yanan Li, Donghui Wang, Jie Liu, Hongxia Yang, and Fei Wu. Dress like an internet celebrity: Fashion retrieval in videos. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 1054–1060, 2021.

[34] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *Proc. CVPR*, pages 5012–5021, 2019.

[35] Qianyu Zhou, Xiangtai Li, Lu He, Yibo Yang, Guangliang Cheng, Yunhai Tong, Lizhuang Ma, and Dacheng Tao. Transvod: end-to-end video object detection with spatial-temporal transformers. *TPAMI*, 2022.

[36] Haowei Zhu, Wenjing Ke, Dong Li, Ji Liu, Lu Tian, and Yi Shan. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *Proc. CVPR*, pages 4692–4702, 2022.

[37] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *Proc. ICCV*, pages 408–417, 2017.

[38] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *Proc. CVPR*, pages 2349–2358, 2017.