# Cross-view Semantic Alignment for Livestreaming Product Recognition Supplementary Material

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

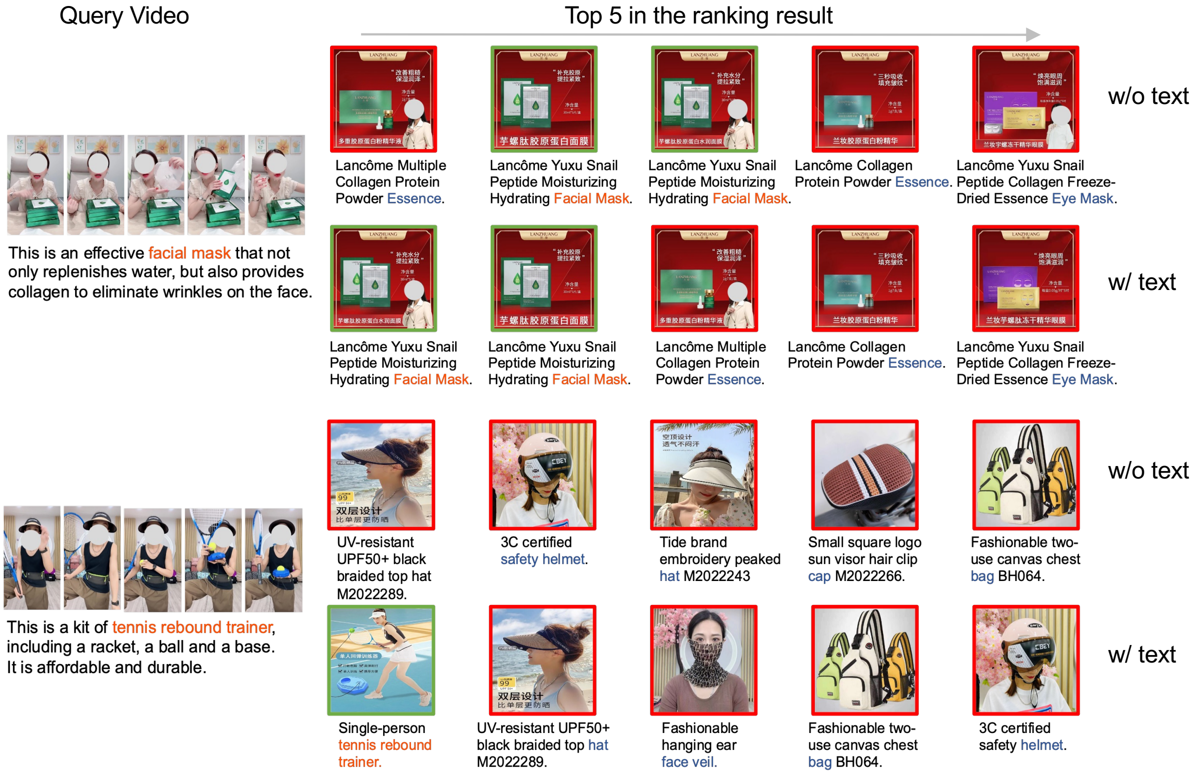

Query Video
Top 5 in the ranking result
w/o text
w/ text
This is an effective facial mask that not only replenishes water, but also provides collagen to eliminate wrinkles on the face.
Lancôme Multiple Collagen Protein Powder Essence.
Lancôme Yuxu Snail Peptide Moisturizing Hydrating Facial Mask.
Lancôme Yuxu Snail Peptide Moisturizing Hydrating Facial Mask.
Lancôme Collagen Protein Powder Essence.
Lancôme Yuxu Snail Peptide Collagen Freeze-Dried Essence Eye Mask.
Lancôme Yuxu Snail Peptide Moisturizing Hydrating Facial Mask.
Lancôme Yuxu Snail Peptide Moisturizing Hydrating Facial Mask.
Lancôme Multiple Collagen Protein Powder Essence.
Lancôme Collagen Protein Powder Essence.
Lancôme Yuxu Snail Peptide Collagen Freeze-Dried Essence Eye Mask.
This is a kit of tennis rebound trainer, including a racket, a ball and a base. It is affordable and durable.
UV-resistant UPF50+ black braided top hat M2022289.
3C certified safety helmet.
Tide brand embroidery peaked hat M2022243
Small square logo sun visor hair clip cap M2022266.
Fashionable two-use canvas chest bag BH064.
Single-person tennis rebound trainer.
UV-resistant UPF50+ black braided top hat M2022289.
Fashionable hanging ear face veil.
Fashionable two-use canvas chest bag BH064.
3C certified safety helmet.

Figure 1: Effectiveness of text modality. For each query video, we show the top 5 shop images of two models, *i.e.*, RICE w/o text and RICE w/ text. The texts below the video and image are ASR text and title, respectively. Images with green and red boundary denote true positive and false positive. Fusing text modality can help the model tackle the challenges posed by small inter-class variation (the first example) and ambiguous intended product (the second example).

## 1. Qualitative analysis

We perform qualitative experiments to analyze the impact of text modality and the ranking results of the proposed model. The qualitative analysis are conducted with RICE on the LPR4M test set.

### 1.1. Effectiveness of text modality

As shown in the first example in Fig. 1, due to the view change of the product in the video, the model (*i.e.*, RICE w/o text) mistakenly retrieved an Essence as the top-1 result. By introducing the text modality, the model (*i.e.*, RICE w/ text) correctly

[*] Equal contribution
[†] Corresponding author

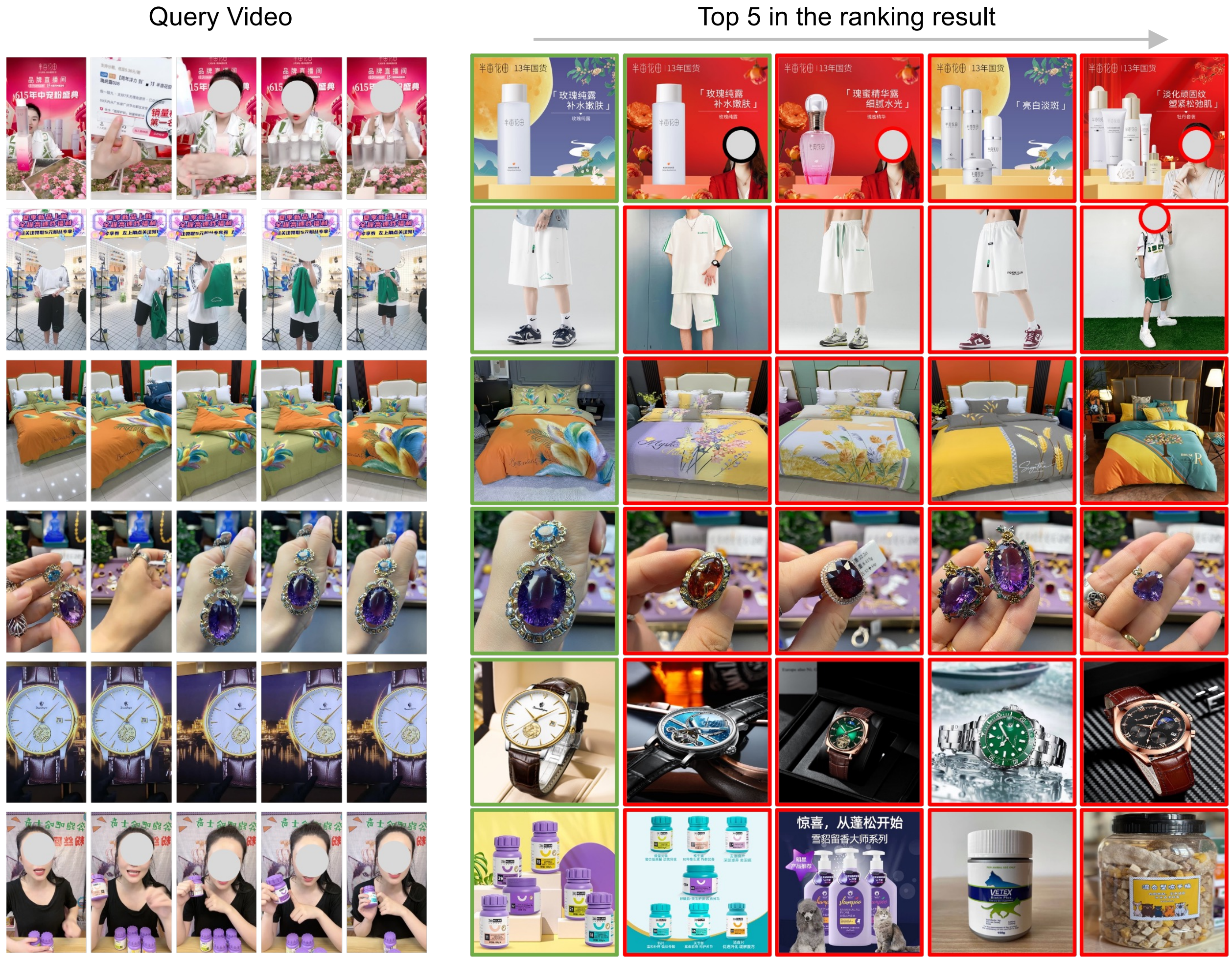


Figure 2: Top 5 ranking results of RICE. We show examples from 6 categories, *i.e.*, beauty, fashion, home goods, jewelry, watches, and nutrition products. As we can see, the proposed RICE is able to perceive subtle differences between products and provide accurate top-1 retrieval results.

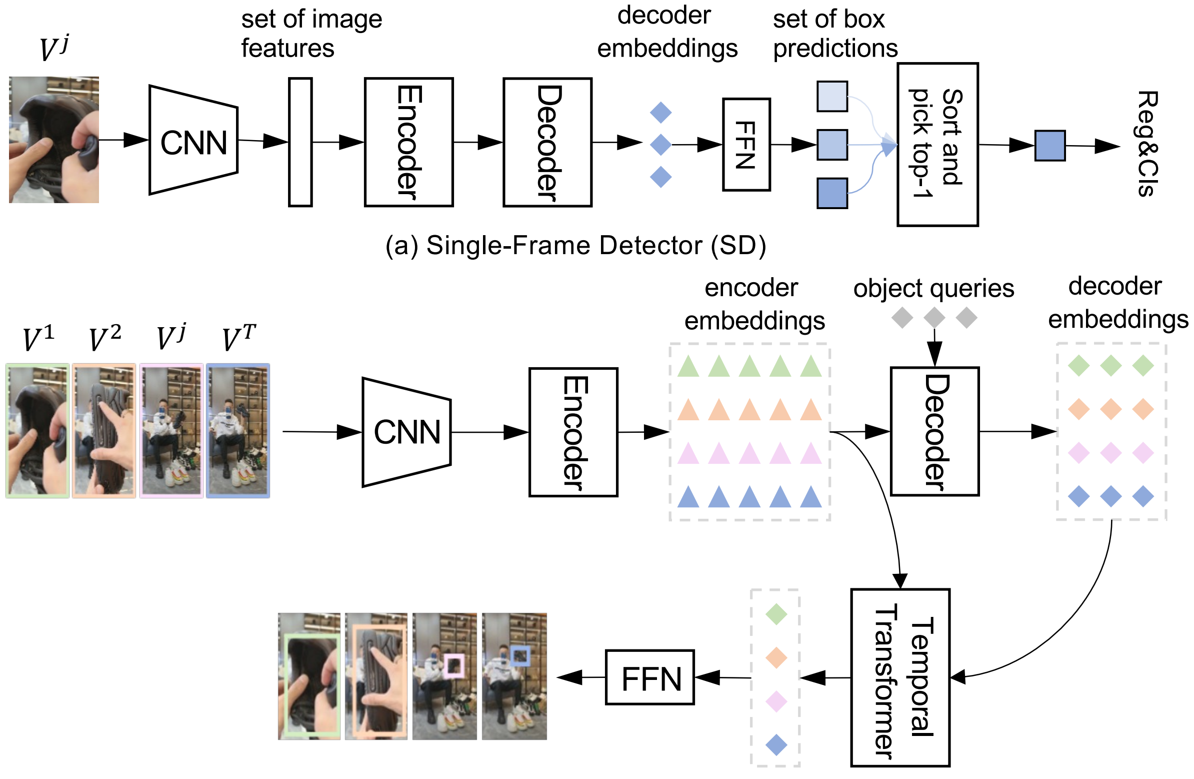
retrieved a facial mask, as the ASR text and image title both clearly indicated that the product was a facial mask. Moreover, as the second example in Fig. 1 shows, in the absence of textual mode and contextual information, the intended product in the video is ambiguous, it could be the hat, tennis rebound trainer or waist bag, so the model (RICE w/o text) retrieves hats and bags from the shop. By introducing the text modality, the model (RICE w/ text) accurately retrieve the tennis rebound trainer.

### 1.2. Ranking results

The ranking results in Fig. 2 demonstrate that the proposed RICE is able to perceive subtle differences between products. For example, in the fourth row, the true positive (top-1) and false positive (top-4 and top-2) products have very similar appearances, with only subtle differences in their outlines, our RICE model can still distinguish them and provide accurate retrieval results.

## 2. Intended Product Detection (IPD)

In livestreaming videos, there are often numerous background products in addition to the intended ones. However, the video encoder (as shown in Fig. 5 in the main paper) takes image patches as input, where both the intended and background products hold equal importance. In order to highlight the intended products in video and suppress the background products,



(a) Single-Frame Detector (SD)

(b) Multi-Frame Detector (MD)

Figure 3: Illustration of Single-Frame and Multi-Frame Detector. We adopt DAB-DETR [4] and TransVOD Lite [8] as the SD and MD, respectively. In (a), DAB-DETR takes frame $\mathcal{V}^j$ as input and produces a set of box predictions, then the box with maximum class score is selected to perform regression and classification. In (b), TransVOD Lite exploits a temporal transformer module (TTM) to fuse the encoder and decoder embeddings of SD and outputs a decoder embedding for each frame. This final decoder embedding is used for the intended box prediction. The TTM is the same as the Sequential Hard Query Mining (SeqHQM) module in TransVOD Lite (Fig. 5 in [8]). Different colors indicate different frames.

we propose replacing patch inputs with detected intended product boxes for the videos. Specifically, the intended product detector takes the video frames as input, and aims to predict one box of the intended product for each frame (*i.e.*, predicts one box for one frame). This section first present the technical detail of IPD (Sec. 2.1). Then, we introduce the training data and protocol (Sec. 2.2). Finally, we conduct extensive experiments to evaluate the performances of IPD (Sec. 2.3).

### 2.1. Methods

Let $\{\mathcal{V}_i^1, \mathcal{V}_i^j, \dots, \mathcal{V}_i^T\}$ be the frames of the $i$-th video, where $T = |\mathcal{V}_i|$ is the frame number and $j$ is the index of $T$. There are some misuses, and for conciseness, we omit the index $i$ of the video in the following.

**Single-Frame Detector (SD).** The SD performs frame-by-frame intended product detection, without any information propagation across frames. It can be a two-stage [7], single-stage [2] or transformer-based [1] detector. Here, we adopt DAB-DETR [4] as the SD. As shown in Fig. 3 (a), given a video frame $\mathcal{V}^j$ with the intended product box label $\hat{b}^j \in \mathbb{R}^4$ and class label $\hat{p}^j \in \mathbb{R}^1$ (we annotate a box for each frame), the SD takes $\mathcal{V}^j$ as input and predicts a set of $N$ boxes. The predicted box locations and class scores are denoted as $\mathcal{B}^j = \{b_1^j, b_2^j, \dots, b_N^j\}$ and $\mathcal{P}^j = \{p_1^j, p_2^j, \dots, p_N^j\}$, respectively. Then we select the box with the maximum class score as the final intended product box and perform classification for 34 categories along with bounding box regression. We follow the overall setting of DAB-DETR, *e.g.*, architecture, loss function, optimization strategy, to train the SD.

**Multi-Frame Detector (MD).** The MD aims to exploit temporal information to tackle the video variations, *e.g.*, occlusion, motion blur, small scale and out of focus. We adopt TransVOD Lite [8] as the MD in our paper. For a more intuitive understanding, we provide a revised illustration of TransVOD Lite, as depicted in Fig. 3 (b). The MD builds upon the SD by incorporating a temporal transformer module (TTM). The TTM here is the same as the Sequential Hard Query Mining (SeqHQM) module in TransVOD Lite (Fig. 5 in [8]). The TTM takes the encoder and decoder embeddings of video frames as input. It introduces self-attention to refine the decoder embeddings and cross-attention to use the encoder embeddings as context. By employing a multi-step filtering approach, at each step, the decoder embeddings predict class scores. Subsequently, based on these scores, the decoder embeddings with lower class scores are gradually discarded until only the last remaining decoder embedding is retained for each frame. This final decoder embedding is then used for predicting the intended bounding box. For more details about the MD, please refer to the original paper of TransVOD Lite [8].

## 2.2. Training details

**Architecture.** The SD, *i.e.*, DAB-DETR, includes an ImageNet-pretrained ResNet-50 backbone, 6 transformer encoders and 6 transformer decoders, 300 learnable queries, and 2 prediction heads for box locations and classes. The TTM in MD is the Sequential Hard Query Mining (SeqHQM) module proposed by TransVOD Lite, the code of which is available at `https://github.com/qianyuzqy/TransVOD_Lite/tree/main`.

**Dataset.** In LPR4M, we annotate a training subset and the whole test set for training and evaluating detector. The detection training/test set contains 1,120,410/501,656 frames with 1,115,629/669,374 intended product boxes, respectively.
**Optimization.** We use Pytorch to implement the detectors. For data augmentation, we resize the input images such that the shortest side is at least 480 and at most 800 pixels while the longest at most 1333.

For the training of SD, we first load the COCO [3] pre-trained weights of DAB-DETR[1], then finetune on LPR4M. We finetune SD with AdamW [6] setting the initial transformer's learning rate to $10^{-4}$, the backbone's to $10^{-5}$ , and weight decay to $10^{-4}$. Training the SD for 20 epochs on 32 Nvidia V100 GPUs takes 3 days, with 2 images per GPU (hence a total batch size of 64). The learning rate is dropped by 0.1 after 15 epochs.

For the training of MD, the parameters of SD are frozen, we only train TTM and FFN. We extract 10 frames at even intervals for each video. Similar to the training of SD, we train MD with AdamW setting the initial transformers learning rate to $2 \times 10^{-4}$, the backbones to $2 \times 10^{-5}$, and weight decay to $10^{-4}$. Training the MD for 20 epochs on 32 Nvidia V100 GPUs takes 2.5 days, with 20 images (or 2 videos with 10 images for each) per GPU (hence a total batch size of 640). The learning rate is dropped by 0.1 after 15 epochs.

## 2.3. Experiments

In this section, we first conduct experiments to evaluate the performances of the SD and MD.

| Methods | overall | | | scale | | | visible duration | | | number of product | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AP | $AP_{50}$ | $AP_{75}$ | small | medium | large | short | medium | long | abundant | medium | few |
| SD | 36.6 | 46.6 | 39.6 | 27.5 | 44.8 | **69.0** | 23.6 | 45.8 | **65.4** | 20.7 | 21.1 | **39.8** |
| MD | 37.0 | 47.1 | 40.2 | 27.7 | 45.1 | **72.7** | 24.8 | 47.5 | **65.7** | 25.1 | 21.2 | **40.3** |

Table 1: The performances of SD and MD are reported. The columns show the AP of the total test set and the $AP_{50}$ of different splits. The best performance is in bold.

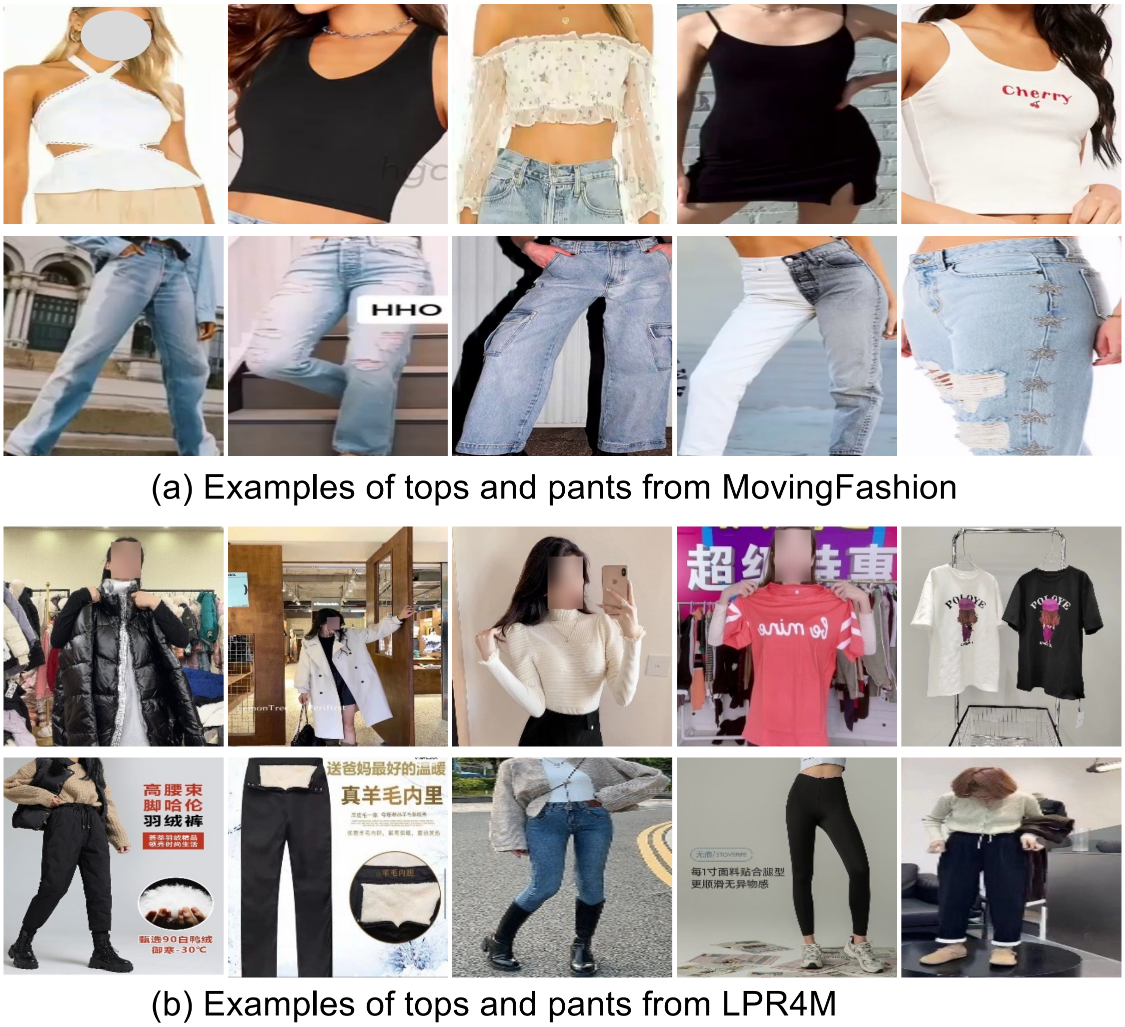
**Performance of Detector.** As shown in Table. 1, we compare the AP performance of SD and MD. The MD outperforms SD on the whole test set and all the splits, which indicates the effectiveness of temporal contexts. In addition, we notice that the highest AP can be achieved on the simple splits with large-scale products, long visible duration, and few product candidates.

| Methods | R1 | R5 | R10 |
|---|---|---|---|
| SD | 30.6 | 62.6 | 73.5 |
| MD | 31.3 | 63.2 | 74.3 |

Table 2: Comparison of single-frame and multi-frame detection.

**Single-frame *vs*. Multi-frame Detection.** The Table. 2 details the rank-*k* accuracy of the proposed model, which takes as input the detected product boxes produced by the single-frame and multi-frame detector, respectively. Directly performing

[1] https://github.com/IDEA-Research/DAB-DETR



(a) Examples of tops and pants from MovingFashion

(b) Examples of tops and pants from LPR4M

Figure 4: In (a) and (b), we compared the dressing styles of MovingFashion and LPR4M. Vests and jeans are distinctive characteristics of MovingFashion, while coats and leggings are unique features of LPR4M when compared to MovingFashion.

| Methods | R1 | R5 | R10 |
|---|---|---|---|
| PMD+Rand | 27.6 | 57.3 | 69.7 |
| PMD+Hard | 29.4 | 62.0 | 73.7 |

Table 3: Comparison of random and hard negative sampling for training PMD.

detection on a single frame achieves 30.6% of R1. The mining of the relation between the current and reference frames leads to a boost of 0.7%. The results verify the idea of exploring temporal product relation for improving intended product detection in video.

## 3. Effect of Hard Negative Sampling

A negative $\langle clip, image\rangle$ pair is hard if the products within them share similar semantics but differ in fine-grained details. The contrastive similarity from the ICL is used for the in-batch hard negative sampling, where the negative clips are more similar to the image have a higher chance of being sampled. As shown in Table. 3, the hard negatives achieve a considerable R1 margin of 1.8% over the random negatives.

## 4. Impact of negative sample number

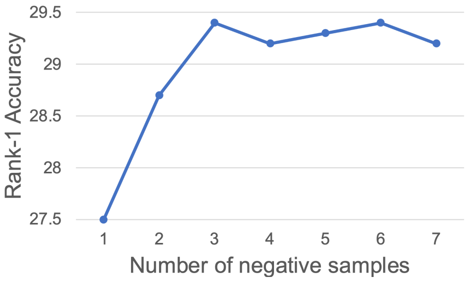


Figure 5: Impact of negative sample number.

We also study the effect of the number of hard negative samples $N_{neg}$ for training PMD. According to Fig. 5, the model achieves best performance when $N_{neg} = 3$ and does not always perform better with the increase of $N_{neg}$.

## 5. Dataset Quality and Diversity

1) Quality checks. Firstly, for ASR, we randomly sample 2k livestreaming and extract 2 clips from each, with a duration of about 10 seconds per clip. Each second contains 3.5 words on average, and we obtain $2k$*2*10*3.5=$140k$ words in total. The word recognition accuracy is 90.16% finally. Secondly, we uniformly sample 1k pairs from each category to obtain 34k pairs for match quality check. The final match accuracy is 92.10%.

2) Social diversity. The samples of WAB are derived from China. The MF did not provide country labels of samples. Regarding upper wear, the MF is often characterized by Euro-American style vests, while LPR4M features mostly summer T-shirts and winter coats. We make the visual comparison in Fig 4.

3) Product diversity. Following [5], we select a very large threshold (0.9) for feature cosine similarity to remove the near-duplicate images that caused by product shift or zoom-in. However, the dataset still presents a significant number of identical products with varying views, also non-identical products that look very similar. We believe such removal will not affect the diversity of the data.

4) In online setting, we cache one livestream frame every $k$ second, *e.g.*, $k$=10. These adjacent frames can be assembled into videos for LPR. This bears little gap from the trimmed videos in the dataset.

## References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proc. ECCV*, pages 213–229. Springer, 2020.

[2] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proc. ICCV*, pages 2980–2988, 2017.

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. ECCV*, pages 740–755. Springer, 2014.

[4] 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.

[5] 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.

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. ICLR*, 2018.

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Proc. NIPS*, 28, 2015.

[8] 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.