Weighted Retriever Ensembles for Video-to-Product Ads Curation

Faizan Ahemad
fahemad3@gmail.com
Intl-ML, Amazon India
Bengaluru, Karnataka, India

Mohammed Abdulla
gmohm@amazon.com
Intl-ML, Amazon India
Bengaluru, Karnataka, India

Sachin Farfade
sfarfade@amazon.com
Intl-ML, Amazon India
Bengaluru, Karnataka, India

Figure 1: Each row of results represents results from one model type, the first image of the row is the query image for retrieval. Green bordered rows represent correct results. Our proposed ensemble model can predict correctly even if individual models fail. The ensemble is able to improve beyond any single model's capabilities. All the ASINs, product images (referred as catalog) and review images are taken from publicly available data shown on Amazon.in website.

ABSTRACT

We present Video-to-Product Ads curation system for MiniTV to identify visually relevant products ads corresponding to objects of interest in video. This retrieval task is significantly challenging due to domain gap and peculiarity in images extracted from videos. Traditionally, images to product retrieval problems are solved using contrastive models with extensive labelled image data. In this paper, we present a framework that enhances traditional retrieval systems using pre-trained CLIP models. Specifically, we present three retrieval paradigms: attributes prediction model, image to image matching model and self supervision model which are built on top of CLIP and can address the challenges without manual labelled data. We analyze the strengths and weakness of each retrieval paradigm. Additionally, we present a ensemble model that combines all three models using a novel post scoring weighing method. The ensemble model outperforms individual models, even without domain specific training data. Our ensemble model provides 25.7% gain over best individual model (37.85 vs 30.66 for Precision@5). Also, the ensemble without any task-specific training achieves close to 90% of Precision@5 of the model with task-specific training.

AIMLSystems, October 8–11, 2024, Baton Rouge, USA

ACM ISBN 979-8-4007-1161-9…$XX.XX
https://doi.org/10.1145/nnnnnnn.nnnnnnn

CCS CONCEPTS

• **Computing methodologies → Visual content-based indexing and retrieval**; *Ensemble methods*; Image representations.

KEYWORDS

Video-to-Product, Ensemble, Image to Image Retrieval, Vision Transformer

ACM Reference Format:
Faizan Ahemad, Mohammed Abdulla, and Sachin Farfade. 2024. Weighted Retriever Ensembles for Video-to-Product Ads Curation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email. Under Review. (AIMLSystems).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

MiniTV is a video streaming service introduced in India, offering a curated collection of short videos to all customers without the need for a subscription. Differing from Prime Video, which is exclusive to Prime customers, MiniTV is accessible to a broader audience. We propose to show product suggestions in video based on products appeared in the video, Figure 2 shows the sample user interface (UI). We identify objects of interest within the videos and retrieve relevant products from the publicly available amazon catalog to display as ads. This paper focuses on relevant product retrieval via image to image product search and presents Video-to-Product ads retrieval system for MiniTV.

Customers can click the shop button to open the interface to interact with ads as shown in Figure 2. Our overall system is a human-in-the-loop approach as shown in Figure 4, where human annotators identify objects of interest from videos and crop a still

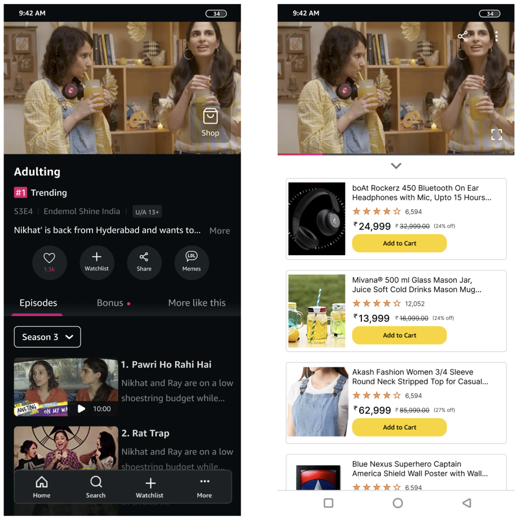

Figure 2: User interface of MiniTV Ads in MiniTV Player.

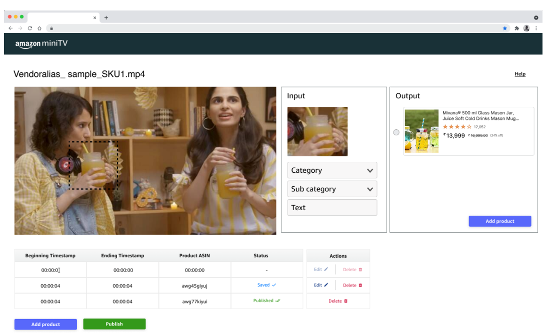

Figure 3: Annotator interface for searching relevant products using a cropped image.

image of the object to query our retrieval system. The annotator interface shown in Figure 3 is used for this purpose by the annotators. The retrieval system (described in this paper) then performs Maximum Inner Product search (MIPS) over the embeddings of product images from the sponsored product catalog, returning top matching products based on visual similarity. Human annotators then vet these products for relevance, taking into account both the video and image context. Human annotators are required for video retrieval as there are several factors like saliency, customer preference, and contextual understanding, that are difficult for algorithms to fully grasp without human intervention. This paper in particular describes our retrieval system used in the annotation process.

Prior works from industry have proposed scalable solutions for Image-based product retrieval. Notable works include Amazon's "Shop the Look" system [7] and others like Facebook's Groknet [1], eBay [13], Alibaba [14]. presents a web-scale fashion and home product visual search system These systems consist of three key components: 1) an image representation model tailored for product retrieval, 2) a retrieval system capable of efficiently searching through millions of product images, and 3) parallel data between the query and catalog domains for training purposes. Similarly, Video-to-Shop aims to match clothing items present in social videos or single images with an e-commerce database. Notable systems include MovingFashion [8], video2shop [5], and DPRNet [15] which rely on videos with labeled product data for training.

Our objective was to tackle the challenges of domain gap between catalog images from public amazon website and videos without requiring manual data labeling. Advancements in self-supervised learning, by methods like SimCLR [4] and DINO [2], enable learning image embeddings using unlabeled images. We developed three distinct image encoders starting from CLIP [10] pretrained image model using ViT [6] architecture: (i) **Product Attribute Prediction**: multi task model that uses cosine similarity on CLIP backbone for stable learning, (ii) **Image-Image model**: learnt using triplet margin loss and online hard mining, and (iii) **Self-Supervised model** that adopts DINO with modifications to improve generalization to new domains when only images from the new domain are available without any labels. We experimented with multiple methods of combining the three image encoders and discovered that training instabilities lead to difficulty in combining these models during training. To combat this, we introduce an **inference-time ensemble approach** that combines our three proposed models, enhancing retrieval accuracy beyond individual model capabilities. Our ensembling method provides 25.7% gain over best individual model (37.85 vs 30.66 for Precision@5). It can match 90% of performance compared to model trained with task specific data.

2 APPROACH & ENSEMBLE STRATEGY

Given a query image, denoted as q_i and belonging to the query domain Q, the objective is to find a catalog image, represented as c_j from the catalog domain C, such that the corresponding similarity function $S(q_i, c_j)$ is maximized. Here, S represents the actual real-world visual similarity function that bridges the gap between the two distinct domains, and our goal is to approximate this function using a trained model. We utilize the following data sources to approximate the similarity function S: (1) Catalog Domain (C^A): entire set of catalog images sourced from Amazon. The superscript A signifies the availability of parallel mapping data between catalog images and ASINs [1]. (2) Review Domain (R^A): images uploaded by customers as part of their reviews. Each of these images are associated with a ASIN and (3) Metadata Domain ($M_1, M_2, M_3, ..., M_k$): ASINs in the Amazon catalog possess metadata attributes such as GL, category, title, bullet points, and item description encompassing numerical values, categorical values, and natural language text.

2.1 Classification Model

Each catalog image c_j is mapped to ASIN metadata attributes $M_{1...k}$. We employ the task of predicting the catalog attribute given the catalog image c_j to help our model learn similar features for ASINs which have similar attributes. Recent work [1, 9] has shown that attributes with more granularity help in better pre-training and learning of representations. We include highly granular attributes such as product titles, bullet points, sub-category and browsenode in our task. We use ViT image backbone as our image encoder g with

[1]An ASIN is a unique identifier (Product Id) assigned to each product on the Amazon catalog.

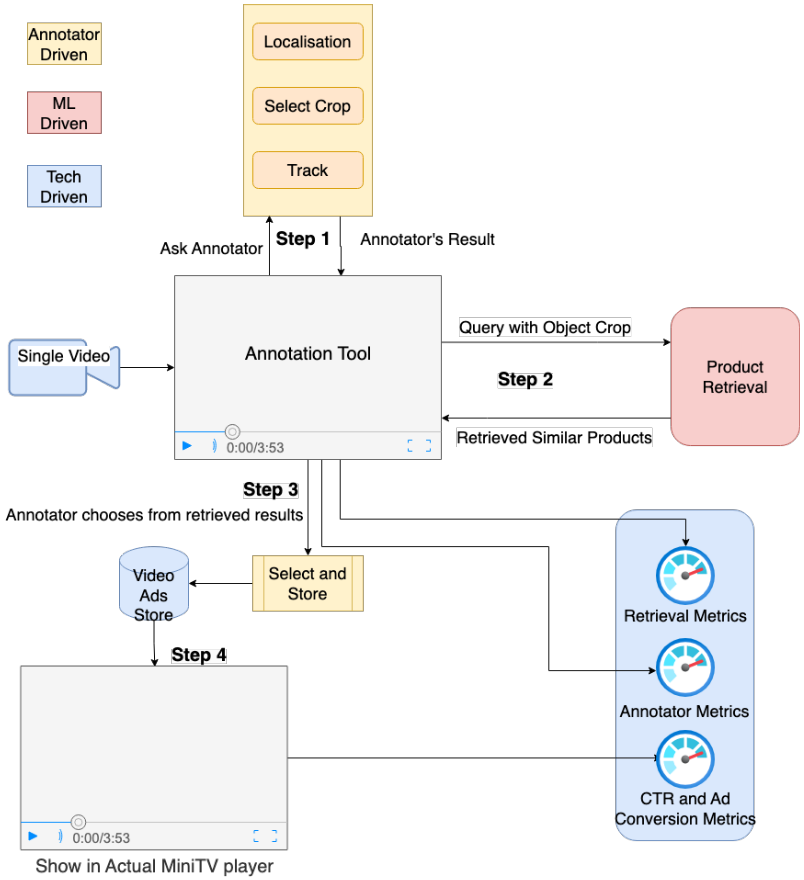

Figure 4: Overall system design.

k attribute heads $H_{1...k}$ made of multi-layer fully connected layers, each attribute head H_i takes the image features encoded by g and predicts the attribute M_i. Since the attributes $M_{1...k}$ have different types, including text, numeric and categorical types, instead of predicting the exact attribute, we minimise the cosine distance between the output of attribute heads and a fixed text encoder t from CLIP text backbone which encodes the raw metadata attributes $M_{1...k}$ into vectors. The loss is given by

$$L_{classification} = \sum_{i}^{K}(1 - \frac{H_i(t(\mathbf{M}_i)) \cdot g_\theta(c_j)}{\|H_i(t(\mathbf{M}_i))\| \|g_\theta(c_j)\|}) \quad (1)$$

2.2 Deep Metric Learning

We use Triplet margin loss as our deep metric learning method to train our retrieval model. We form triplets of data points, consisting of an anchor image x^a, a positive image x^p, and a negative image x^n, and train the model to ensure that the positive image is closer to the anchor image than the negative image. The loss function for training the model is given by

$$L_{triplet} = max(0, d(f(x^a), f(x^p)) - d(f(x^a), f(x^p) + \alpha) \quad (2)$$

where $d(a, b)$ is the euclidean distance between embedding a and b and α is the triplet margin. We use ViT image backbone as the function f. We leverage Catalog C^A and Reviews R^A to train the triplet model. The anchor and positive images are sampled from different views of catalog images and review images of the same ASIN and negative images are sampled using Multi Similarity Hard negative mining [12].

2.3 Self Supervised Model

Self-supervised learning (SSL) has emerged as a promising approach to train deep neural networks using unlabeled image data. We use images from Catalog C^A and Reviews R^A to train the SSL model. Of all the SSL methods, we have chosen DINO [3] for our use case as these models can work on small batch sizes and produce effective representation unlike contrastive methods. DINO is trained using a teacher-student model to match distinct views of same image.

Modified DINO: In the traditional DINO training, each image in the training data is treated as an individual class. In our case, where catalog images C^A and review images R^A are associated with an ASIN (Product Id) tag, we leverage this information to create the global and local views of the DINO training. Specifically, we generate the global and local views using different images that share the same ASIN. We need to differentiate between different color images of the same ASIN, hence we exclude color jitter and gray-scale transformations and use only random flip, resize and random perspective transformations. These modifications enables us to train DINO models that are better suited for our retrieval task and improves their performance in capturing and leveraging the ASIN-based image variations.

2.4 Score Based Weighted Ensemble Strategy

Our novel score based weighted ensembling strategy combines the outputs of multiple modeling approaches without any additional training. Let $F_1, F_2, F_3, \ldots, F_N$ represent the functions that compute similarity measures between two images from N different modeling approaches. Given a query image q and a universe of retrieval images U, our goal is to select the image u from the retrieval universe U that maximizes the following ensemble score:

$$\underset{u}{\operatorname{argmax}} \sum_{i=1}^{N} \lambda_i * F_i(q, u) \quad \forall u \in U \quad \ni \sum_{i=1}^{N} \lambda_i = 1 \tag{3}$$

Here, λ_i determines the importance or contribution of each modeling approach to the final retrieval. By adjusting the hyperparameters λ_i as shown in Figure 7, based on the characteristics of the retrieval task, we can fine-tune the ensemble and optimize its performance.

We tried two other ensembling strategies. Firstly we took equal samples from each individual model, this method doesn't correct retrieval errors when all the models fail individually, which we saw our method does as shown in Figure 1. The second method of ensembling was to ensemble (by mean pooling and weighted mean pooling) the image vector representation from each model's last layer $[CLS]$ token before the KNN step for both indexing and searching, this also yielded subpar results (about 10% lower in absolute metrics) and also did not correct retrieval errors when all the models fail individually. For weighted mean pooling in second method we optimised individual model's image vector representation weights in overall ensemble vector by performing hyperparam search where all weights sum to 1 and individual weights varied from 0 to 1 and finally choose the best pooling contribution weights for all three models in the second method of ensemble.

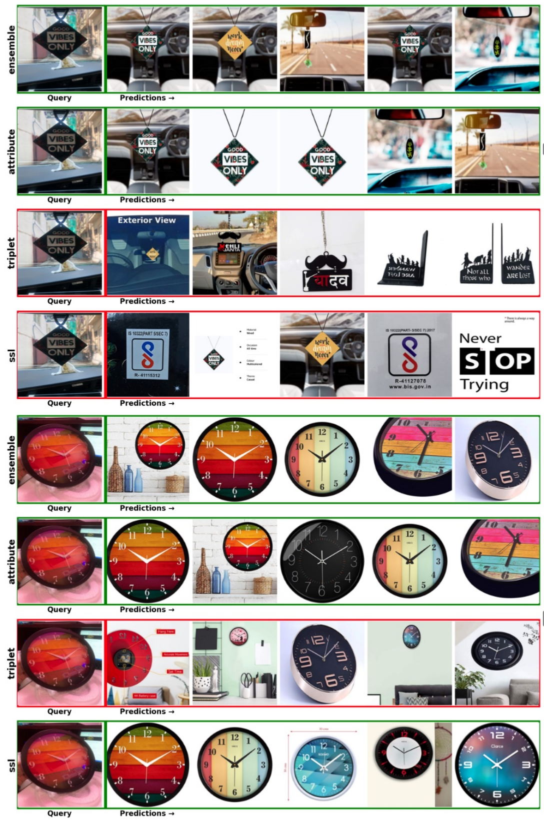

Figure 5: Instances where ensemble model predicts correctly along with one or more individual models also predicting with success.

3 RESULTS & ANALYSIS

We use publicly available data of amazon catalog images, customer review images and amazon ASINs meta data. We use 170K ASINs with 1M images from the product catalog and 1.3M review images. We keep 7200 images from catalog and 17K images from review images for testing. All the ASINs, images (referred as catalog) and review images are taken from publicly available data shown on Amazon.in website. All ViT image backbones are CLIP pre-trained.

Retrieval Method: We compute inner product between query image embedding and embedding of all the images in retrieval set. We then choose top k images using maximum inner product (MIPS). We use precision@5 as the main metric for comparison of different architectures. For a given query image, precision@k is 1 if ground truth ASIN image is among the k results retrieved else it is 0.

Our main results are summarised in Table 1. To evaluate the performance of our model we report Precision@5 for query images belonging to our product catalog as well as wild images taken by customers from customer reviews. The models are trained either using Catalog images (Cat) and/or Review images (Rev) as specified in the "Training Data" column. For the baseline model we have used the CLIP model [11]. The CLIP model has a strong performance on catalog test data and moderate result of 22.65 of precision@5 for test data of review images. All the subsequent models use CLIP pre-trained weights as initialisation. Models trained with classification paradigm are reported as "Attributes" Model in table Table 1. We use the embeddings from ViT backbone as the embeddings for retrieval. Models trained on "Cat" data and "Cat + Rev" data perform better compared to the baseline. Interestingly results from this model on "Review" test data trained only using "Cat" data vs trained with "Cat + Rev" data is just 98bps less. This suggest that the classification paradigm when trained on top of CLIP weights generalises to unseen data set. We observed that RMSE loss or softmax loss did not result in a stable model and cosine distance works the best for models initialised to clip weights, as this loss function is similar to CLIP's loss function. Second set of models (Triplet Model) is trained with triplet loss paradigm (Sec. 2.2). Triplet loss based supervised models require positive and negative data samples with respect to anchor image for training and is the most popular method for retrieval used in existing systems from Amazon, eBay, Alibaba and Facebook. Triplet model trained on "Cat + Rev" data, provides the highest increment of 703 bps compared to any individual model.

Figure 6: Ours trained model capture color, shape and even pattern information within images. When attribute based matching is not sufficient our individual models from Section 2.2, 2.3 enable our ensemble to identify the correct match.

When trained on catalog data alone, it improves catalog performance but degrades unseen review images performance compared to baseline. This suggests that Triplet model do not generalise well to unseen data but when labeled data is available they are still the best model for retrieval. Third set of models are the SSL models, Modified DINO. (Sec 2.3. When only images from new domain are given without labels, SSL models generalise well on the new domain.

The results at the bottom of Table Table 1 shows the results for ensemble models. As an ablation study, we tried testing different combination of models and have reported their precision@5. All the ensemble models have better precision@5 compared to individual model. Except for "Weighted Ensemble" model, every other ensemble model uses equal weights for combining the individual models. Best Weighted ensemble model assigns weight of 0.2 for "Attributes" model, 0.4 for "Triplet" and 0.4 for "DINO" and these weights were identified by using grid search using separate held out data. The grid search results are plotted in figure 7 and a combination of Attributes, Triplet and SSL model have the highest precision for multiple combinations, showing that choice of weights for ensemble are relatively stable. Weighted Ensemble models performs the best compared to every other model when trained on both the publicly available catalog and review images, but interestingly the ensemble model trained only on catalog images has 90% precision@5 compared to the former model. This shows that ensemble models can generalise to unseen datasets.

Table 1: Precision@5 for different models. The columns "Catalog" and "Review" report the metric on held out catalog test data and review test data respectively. TTA reports the metric with Test Time Augmentation. Models are trained with catalog data (Cat) and customer reviews data (Rev) as specified in "Training data" column.

Model	Training Data	Catalog	Review	+ TTA
Zero Shot Clip	-	85.92	22.65	22.07
Attributes	Cat	87	25.87	26.56
	+ Rev	91.89	26.85	28.02
Triplet Model	Cat	93.16	20.60	21.09
	+ Rev	91.3	29.68	30.66
Modified DINO	Cat	89.05	25.78	25.29
Attributes	Cat	93.75	29.78	30.56
+ Triplet + DINO	+ Rev	91.08	34.08	34.57
Attributes + Triplet	Cat	93.58	28.22	27.53
+ DINO + CLIP	+ Rev	92.41	33.30	31.73
Weighted Ensemble	Cat	96.25	**33.49**	<u>33.76</u>
	+ Rev	95.20	**37.30**	**37.85**

Table 2: GL wise Precision@5 for our models.

Model	Home	Kitchen	furniture	Apparel
Baseline CLIP ViT	31.84	68.05	16.31	12.98
Weighted Ensemble	37.3	90.27	22.75	20.89
+ Review Images	40.43	88.88	30.17	28.51

We also perform Test time augmentation, and reports its results in Table 1 under "Review (TTA)" column. During the inference stage we take the query image and apply multiple image augmentations, one at a time on the query image thereby generating multiple query images for the same query image. Then all these query images representing the same original image is used as a separate query image for retrieval and the scores are combined to select the best results. For most models test time augmentation improves precision by $0.5 - 1.0\%$.

We plot the precision@k for different values of K in Figure 7. Models having higher precison@5, consistently have a higher precision at all other k values. This allows us to choose the "k" to be shown to customers based on business requirement. To check how are models are performing on different categories, we test our models performance on different GLs in Table 2. For GL kitchen our models perform the best and for other Gls we still have room to improve the precision@k.

Finally, We asked domain experts to judge Precision@5 of our model vs current image similarity model used for image search in search page for 1000 actual MiniTV video extracted frames. Our model improves considerably in Precision@5 compared to the current image similarity model used in Amazon image search by 1.27x or 27%.

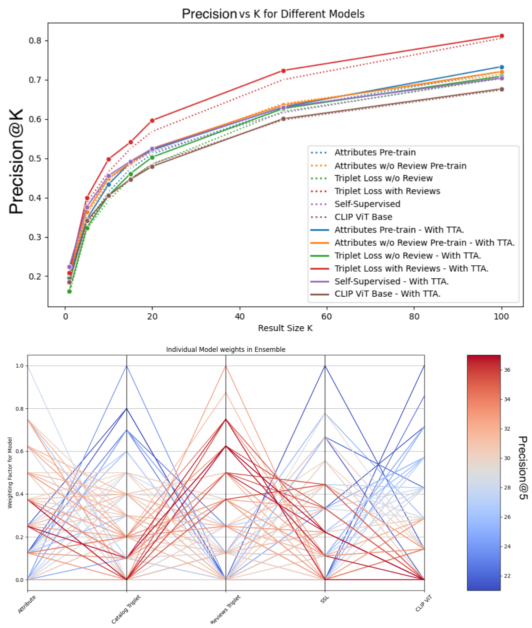

Figure 7: Top: Precision @ K for models using Test time augmentation vs not using (shown in dotted). Test time augmentation enhances performance in all model classes. Bottom: Precision @ K for various ensemble model weight combinations. No single model is best but rather a weighted combination of all models provides the best results.

Table 3: Precision@5 for our method vs current model.

Model	Num Samples	Precision@5
Amazon Image Search	1000	1x
Weighted Ensemble (Ours)	1000	1.27x

Results shown in Table 3 show our model at 66% Precision@5 improves considerably over 52% Precision@5 from current image similarity model on our domain of MiniTV cropped image to image product retrieval search.

We show noteworthy visual results in Figure 1, 5 and 6. In Figure 1 we show examples **where ensemble succeeds while remaining models fail**. This is surprising yet expected since ensemble combines each model by using our scoring method described in Section 2.4.

4 CONCLUSIONS

We demonstrate that our ensemble retrieval method effectively combines the strengths of all three models while mitigating their weaknesses, ultimately providing an enhanced retrieval accuracy beyond the capabilities of individual models. Moreover, the ensemble model without task-specific training data achieves 90% of the performance compared to the model trained with task-specific data, highlighting its adaptability to unseen datasets. Test Time Augmentation further improves precision for most models. Additionally, the GL-wise performance reveals that certain categories, such as Kitchen, outperform others, suggesting room for improvement in some areas.

REFERENCES

[1] Sean Bell, Yiqun Liu, Sami Alsheikh, Yina Tang, Edward Pizzi, M Henning, Karun Singh, Omkar Parkhi, and Fedor Borisyuk. 2020. GrokNet: Unified computer vision model trunk and embeddings for commerce. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2608–2616.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9650–9660.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 9630–9640. https://doi.org/10.1109/ICCV48922.2021.00951

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 1597–1607. https://proceedings.mlr.press/v119/chen20j.html

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

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs.CV]

[7] Ming Du, Arnau Ramisa, Amit Kumar K C, Sampath Chanda, Mengjiao Wang, Neelakandan Rajesh, Shasha Li, Yingchuan Hu, Tao Zhou, Nagashri Lakshminarayana, Son Tran, and Doug Gray. 2022. Amazon Shop the Look: A Visual Search System for Fashion and Home. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Washington DC, USA) *(KDD '22)*. Association for Computing Machinery, New York, NY, USA, 2822–2830. https://doi.org/10.1145/3534678.3539071

[8] Marco Godi, Christian Joppi, Geri Skenderi, and Marco Cristani. 2022. MovingFashion: a Benchmark for the Video-to-Shop Challenge. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1678–1686.

[9] Guan Zhe Hong, Yin Cui, Ariel Fuxman, Stanley H. Chan, and Enming Luo. 2023. Towards Understanding the Effect of Pretraining Label Granularity. arXiv:arXiv:2303.16887

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

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

[12] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. 2019. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5022–5030.

[13] Fan Yang, Ajinkya Kale, Yury Bubnov, Leon Stein, Qiaosong Wang, Hadi Kiapour, and Robinson Piramuthu. 2017. Visual Search at EBay. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) *(KDD '17)*. Association for Computing Machinery, New York, NY, USA, 2101–2110. https://doi.org/10.1145/3097983.3098162

[14] Yanhao Zhang, Pan Pan, Yun Zheng, Kang Zhao, Yingya Zhang, Xiaofeng Ren, and Rong Jin. 2018. Visual Search at Alibaba. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 993–1001. https://doi.org/10.1145/3219819.3219820

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

Received 13 May 2024