Adapting Video Instance Segmentation for Instance Search

An Thi Nguyen
an.t.nguyen@ntnu.no
Norwegian University of Science and Technology
Gjøvik, Norway

ABSTRACT

Modern applications of instance search face many challenges. Real-life applications are often complex, containing a large variety of object classes. Furthermore, the same object can be subject to large appearance variations due to viewpoint changes, partial occlusion, and different geometric transformations. Much of the research up to now has focused on narrow instance search applications for rigid and near-planar objects with a small number of queries, such as searching for landmarks or brand logos. In this paper, we adapt YouTube-VIS, a large dataset for video instance segmentation, for the instance search task. The dataset contains 8,430 unique queries and 40 different object classes, where many objects inhibit viewpoint variations and occlusion. We present a two-stage architecture to account for the difference between the two tasks of instance search and video instance segmentation. The first stage performs instance segmentation to separate the instances from the background, while the second stage encodes the isolated instances using a fine-tuned CLIP image encoder. Despite our relatively simple architecture, we achieve promising results without performing any post-processing steps such as query expansion or re-ranking.

CCS CONCEPTS

• **Computing methodologies** → **Visual content-based indexing and retrieval**; *Neural networks*; • **Information systems** → *Information retrieval.*

KEYWORDS

instance search, instance-level image retrieval, CLIP, video instance segmentation, deep metric learning

ACM Reference Format:
An Thi Nguyen. 2023. Adapting Video Instance Segmentation for Instance Search. In *20th International Conference on Content-based Multimedia Indexing (CBMI 2023), September 20–22, 2023, Orléans, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3617233.3617249

1 INTRODUCTION

In recent years, significant efforts in the field of instance search have been dedicated to investigating various methods of processing convolutional neural network features for use as embeddings, mainly focusing on landmark datasets [2, 5, 7, 15, 17–19]. However, instance search methods on landmark datasets may not generalize well to other real-life applications. The landmarks in commonly used datasets for instance search are often large, rigid and near-planar, while object instances in many real-life applications can be small and inhibit different appearance variations. Most of these benchmark datasets are also very small, the Oxford105k and Paris106k datasets only containing 11 and 12 unique landmark instances, respectively [10, 11], or 26 and 25 query groups for the revisited versions of these datasets [13]. While there exists large-scale datasets such as the Google Landmarks Dataset [21], the Met dataset [24], Product1M [25], and Stanford Online Products [9], these datasets mainly contain rigid objects that take up a large portion of the images. However, constructing large-scale multi-class datasets for instance search can be challenging, as one has to collect several images of the same object from different angles and in different contexts.

More success has been achieved at constructing large-scale datasets in other tasks within computer vision, such as in video object tracking (VOT) and video instance segmentation (VIS). In the VIS task, the objective is to simultaneously detect, segment and track object instances in videos [23]. Object tracking and instance search can be viewed as similar problems, but with some differences. Object tracking concerns itself with localizing an object across frames within a narrow time period in one specific setting. In contrast, instance search aims to localize an object within a collection of frames, where the frames may belong to different settings and be months apart. The background context of the frames can differ greatly in instance search. There can therefore be generalization problems to training and evaluating instance search models on datasets for visual object tracking.

In the absence of large datasets for instance search, building embeddings based on features from pixel-level segmentations can aid in generalization for an instance search model that is trained on object tracking datasets, given that the segmentation is sufficiently accurate. As many objects in current benchmark datasets for instance search are captured in fixed background contexts (i.e. landmarks and artwork), applying only moderately accurate segmentation can also serve as an improvement to current practices. Furthermore, the appearance and viewpoint variations of instances in object tracking datasets, as well as the different object classes (of which many are non-rigid), provide challenges that are closer to real-life applications.

In this paper, we aim to investigate the discriminability of embeddings based on segmentations for instance search. We adapt YouTube-VIS [23], a large dataset for video instance segmentation, for the instance search task. The dataset contains 8,430 unique instances and 40 different object classes, where many objects inhibit viewpoint variations and occlusion. We present a two-stage architecture to account for the difference between the two tasks of

CBMI 2023, September 20–22, 2023, Orléans, France

ACM ISBN 979-8-4007-0912-8/23/09.
https://doi.org/10.1145/3617233.3617249

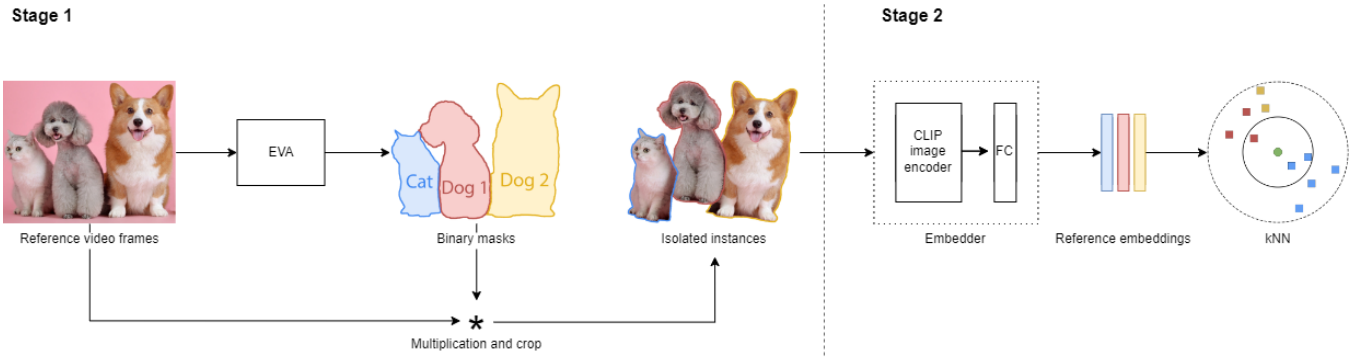

Figure 1: The proposed two-stage transformer architecture

instance search and video instance segmentation. The first stage performs instance segmentation to separate the instances from the background, while the second stage uses a CLIP image encoder to generate embeddings from the isolated instances. We evaluate our architecture on the YouTube-VIS and OVIS datasets, and add distractors from the COCO dataset to assess the discriminability of the embeddings for a large dataset. Despite our relatively simple architecture, we achieve promising results without performing any post-processing steps such as query expansion or re-ranking.

2 RELATED WORK

2.1 Segmentation

To our knowledge, only one previous study has attempted to leverage segmentation masks for the instance search task. Zhan et al. [26] extract object features from segmentations obtained by fully convolutional instance-aware semantic segmentation. The authors claim that their method ignores background context in the feature extraction, thus providing a workaround for using large object tracking datasets for evaluation. However, they extract features from the deeper convolutional layers (e.g. conv3 and conv4 of ResNet-101), where large regions of the image have been pooled together, including background context. It is therefore unclear how well their proposed method would perform on real-life instance search applications with objects in varying background contexts. Instead of extracting the features directly from the network layers, we first perform an element-wise multiplication between the original image and the binary segmentation mask to remove background context.

2.2 CLIP

Although the zero-shot Contrastive Language-Image Pre-Training (CLIP) model has achieved very good results on a variety of benchmarks without task-specific training data, its ability to perform fine-grained classification tasks, such as distinguishing between different variations of aircraft or flower species, is limited [14]. CLIP-art [3] aims to address the challenge of fine-grained classification and artwork retrieval by fine-tuning CLIP on the Met dataset, which contains descriptive text for each artwork. In [1], the CLIP text encoder is fine-tuned on a combination of image and text features for conditioned fashion image retrieval. A combiner network is then used to fuse the image and text features from CLIP. Due to the lack of descriptive text labels for the instances, except for the object categories which may be inadequate for the large intra-class variations in our task, we choose to fine-tune only the CLIP image encoder using a pair-wise multi-similarity loss.

3 METHODOLOGY

3.1 Two-stage Transformer Architecture

Figure 1 shows our two-stage transformer-based architecture for instance search. In the first stage, we pass all reference video frames through the instance segmentation model to obtain the binary masks. We use EVA [4] for instance segmentation. Then, we obtain the isolated instances by performing element-wise multiplications between the original video frames and the corresponding binary masks, and cropping the result. In the second stage, we pass the isolated instances through an embedder consisting of a ViT-B/16 384x384 CLIP model [14], followed by a fully-connected layer which outputs 64-dimensional embeddings. Finally, we compute k-nearest neighbors (kNN) on the reference embeddings.

3.2 Dataset

3.2.1 YouTube-VIS. Our pipeline is trained and evaluated on the 2022 YouTube-VIS (YVIS) dataset, comprising 2,985 high-resolution YouTube videos for training, 492 for validation, and 503 for testing. There are 8,430 unique video instances, 241k manual annotations, and 40 different object categories. As the YVIS validation and test annotations are kept secret for the Video Object Segmentation Challenge, we create our own train/validation/test split from the YVIS training set. We randomly split the YVIS training set into 2,388 training videos, 298 validation videos, and 299 test videos.

Stage 1: The YVIS training set, with our own train/validation/test split, is used for fine-tuning and evaluating the instance segmentation model in Stage 1. The mask annotations and class labels constitute the ground truth annotations.

Stage 2: We construct a separate dataset for fine-tuning and evaluating the embedder in Stage 2. This dataset is obtained from the YVIS training set by performing element-wise multiplications between the video frames and corresponding binary mask annotations, and cropping the result so that we end up with the isolated instances only. The same train/validation/test split is used as in Stage 1.

Table 1: Instance search results on perfect and actual segmentation. We evaluate our approach on the YVIS dataset (first column); YVIS with the OVIS training set as distractors (second column); YVIS with both the OVIS and COCO datasets as distractors (third column); OVIS training set alone (fourth column); and on actual YVIS segmentations output from Stage 1 (fifth column).

	Perfect segmentation				Actual segmentation
	YVIS	YVIS+OVIS	YVIS+OVIS+COCO	OVIS	YVIS
mAP	0.866	0.829	0.809	0.596	0.764
mAP@r	0.831	0.797	0.778	0.516	0.747

3.2.2 OVIS. The OVIS dataset [12] is a VIS dataset focused on occlusion, consisting of 5,223 unique instances from 901 videos with severe object occlusions. The 25 object categories in the dataset are a subset of the YVIS categories. We use the OVIS training set as distractors when performing instance search on the YVIS dataset. To assess the accuracy of the YVIS-trained embedder on occluded objects, we also calculate the instance search performance on the OVIS training set.

3.2.3 COCO. We use the COCO training set [6], comprising over 800k instances, as additional distractors to further evaluate the discriminability of the embeddings for a large dataset.

3.3 Implementation Details

3.3.1 Stage 1: We fine-tune all layers of EVA for 2 epochs using the AdamW optimizer, a batch size of 1, and a learning rate of 2.5e-7 with 0.1 weight decay. The fine-tuned model achieves 0.725 box AP and 0.646 mask AP on YVIS.

3.3.2 Stage 2: We use the ViT-B/16 384x384 CLIP implementation from the PyTorch Image Models library [22], trained on LAION-2B [16] and fine-tuned on ImageNet. We set the batch size to 32, and the optimizer to Adam with a learning rate of 0.00001 and 0.0001 for the backbone and fully-connected layer, respectively. The weight decay is set to 0.0001. We construct batches by randomly sampling C unique instances, and then randomly sampling M images belonging to each of the C instances. We set C = 8, M = 4, and k in kNN to be the maximum of the number of occurrences of each instance. We train the embedder on the isolated instances of YVIS using a multi-similarity loss and multi-similarity miner [20], until the validation accuracy plateaus.

3.4 Evaluation

We use mean Average Precision (mAP) and mean Average Precision at r (mAP@r) [8] as accuracy metrics. We evaluate two types of embeddings: one is derived by using the ground truth masks as inputs to Stage 2 (perfect segmentation), while the other is derived by using masks generated by Stage 1 (actual segmentation). When evaluating the whole pipeline using actual segmentations output from Stage 1, we use an IoU threshold of 0.6 between a ground truth bounding box in a frame and detected bounding box in the same frame, and a confidence score threshold of 0.3.

4 RESULTS

4.1 Quantitative

Table 1 shows the instance search results on YVIS, YVIS with the OVIS training set as distractors, YVIS with both the OVIS and COCO datasets as distractors, and the OVIS training set alone. Given perfect segmentation, we achieve good results on YVIS with and without distractors added, despite our simple architecture. On OVIS, we achieve a lower mAP and mAP@r of 0.596 and 0.516, respectively, suggesting the need for more sophisticated techniques to handle heavily occluded objects. Note that we do not perform any fine-tuning on the OVIS dataset.

Using actual segmentations output from the instance segmentation model as the reference set, we achieve a mAP and mAP@r of 0.764 and 0.747, respectively. The lower instance search accuracy on the actual segmentations can be attributed to some reference objects not being detected in challenging frames, such as when objects are significantly distorted or largely out of the frame. Inaccurate masks may also cause confusion when the resulting shapes of the detected reference objects differ from those of the query object. Some work remains to fill the gap between the perfect and actual segmentation. We expect that our proposed method will become more useful with the advance of more accurate instance segmentation techniques.

4.2 Qualitative

Figure 2 shows some qualitative results of our instance search approach. These examples have been selected to showcase various strengths and limitations of the embeddings. In the rows denoted with an asterisk, we performed instance search on YVIS with OVIS and COCO distractors, using the embeddings generated from perfect segmentation. For the remaining rows, we performed the search on YVIS using embeddings generated from actual segmentations. The leftmost image on each row depict the query. Given sufficiently differentiable features, the embeddings are able to discriminate between very fine-grained categories, i.e. two different goldfish (a, b). We also see that the embeddings work well for non-rigid objects (c) and textured objects that are small (d), have moderate occlusion (e), viewpoint variation (f), and partly out-of-frame queries that are textured (g, h). Furthermore, similar instances in very different positions are grouped together (i). These results suggest that the embeddings are discriminative for small objects and moderately difficult appearance variations, given sufficiently differentiable texture and shape.

Figure 3 shows some failure cases of our approach. We find that the embeddings do not work as well for less textured queries that are extremely small (a) or partly out of frame (b). In some cases

Figure 2: 19 nearest neighbors, where the leftmost image on each row depict the query image. In the rows denoted with an asterisk, we performed instance search on YVIS with OVIS and COCO distractors, using the embeddings generated from perfect segmentation. For the remaining rows, we performed the search on YVIS using embeddings generated from actual segmentations.

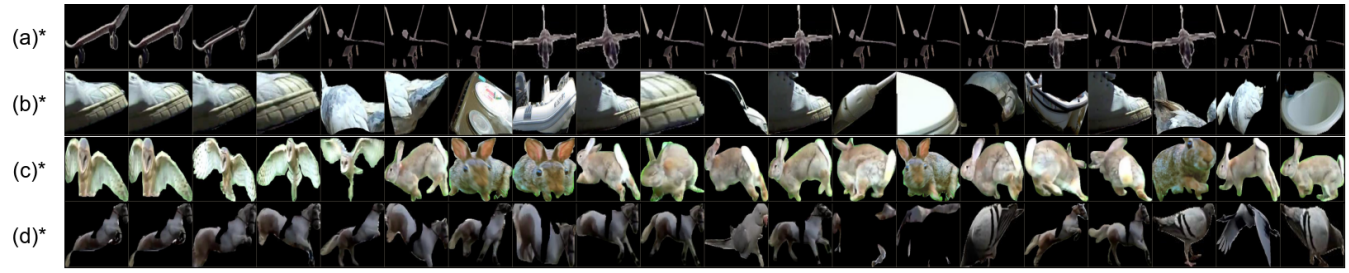

Figure 3: Failure cases of instance search on YVIS with OVIS and COCO distractors, all using the embeddings generated from perfect segmentation. The model confuses a skateboard with airplanes (a); a shoe with various objects such as bird feathers and basin (b); an owl with rabbits (c); and a horse with birds (d).

the embeddings cause confusion between different coarse-grained categories (c, d).

5 CONCLUSION

In this paper, we adapt large VIS datasets for training and evaluating instance search techniques. We propose a two-stage architecture to account for the difference between the two tasks, comprised of instance segmentation and a CLIP image encoder. Despite our relatively simple architecture, we achieve promising results without performing any post-processing steps such as query expansion or re-ranking. Our results suggest that the embeddings are discriminative for small and textured, non-rigid, partly out-of-frame and moderately occluded objects. However, some work remains to fill the gap between the perfect and actual segmentations, as well as techniques to handle heavily occluded objects. Future work may explore incorporating VOT and VIS techniques to better handle frames where objects are highly distorted, and where large parts of the objects are out of frame. We hope our work may inspire further research into adapting large visual object tracking datasets for the instance search task.

ACKNOWLEDGMENTS

This work was supported by industry partners and the Research Council of Norway with funding to project number 320783. We thank SFI NORCICS for providing computational resources, and Dr. Eric Arazo Sánchez for providing feedback on the draft.

REFERENCES

[1] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Conditioned and composed image retrieval combining and partially fine-tuning CLIP-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4959–4968.

[2] Bingyi Cao, Andre Araujo, and Jack Sim. 2020. Unifying deep local and global features for image search. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 726–743.

[3] Marcos V Conde and Kerem Turgutlu. 2021. CLIP-Art: Contrastive pre-training for fine-grained art classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3956–3960.

[4] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19358–19369.

[5] Albert Jimenez, Jose Alvarez, and Xavier Giro-i Nieto. 2017. Class Weighted Convolutional Features for Visual Instance Search. In *Procedings of the British Machine Vision Conference 2017*. British Machine Vision Association, London, UK, 144. https://doi.org/10.5244/C.31.144

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[7] Eva Mohedano, Kevin McGuinness, Noel E. O'Connor, Amaia Salvador, Ferran Marques, and Xavier Giro-i Nieto. 2016. Bags of Local Convolutional Features for Scalable Instance Search. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval (ICMR '16)*. Association for Computing Machinery, New York, NY, USA, 327–331. https://doi.org/10.1145/2911996.2912061

[8] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. 2020. A metric learning reality check. In *European Conference on Computer Vision*. Springer, 681–699.

[9] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. 2016. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4004–4012.

[10] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2007. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE conference on computer vision and pattern recognition*. IEEE, 1–8.

[11] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2008. Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 1–8.

[12] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip Torr, and Song Bai. 2022. Occluded Video Instance Segmentation: A Benchmark. *International Journal of Computer Vision* (2022).

[13] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. 2018. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5706–5715.

[14] 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.

[15] Ali S. Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. 2016. [Paper] Visual Instance Retrieval with Deep Convolutional Networks. *ITE Transactions on Media Technology and Applications* 4, 3 (2016), 251–258. https://doi.org/10.3169/mta.4.251

[16] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* 35 (2022), 25278–25294.

[17] Jingkuan Song, Tao He, Lianli Gao, Xing Xu, and Heng Tao Shen. 2018. Deep Region Hashing for Generic Instance Search from Images. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (April 2018). https://ojs.aaai.org/index.php/AAAI/article/view/11277 Number: 1.

[18] Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. 2021. Instance-level image retrieval using reranking transformers. In *proceedings of the IEEE/CVF international conference on computer vision*. 12105–12115.

[19] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. 2016. Particular Object Retrieval With Integral Max-Pooling of CNN Activations. In *ICLR 2016 - International Conference on Learning Representations (International Conference on Learning Representations)*. San Juan, Puerto Rico, 1–12. https://hal.inria.fr/hal-01842218

[20] 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.

[21] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. 2020. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2575–2584.

[22] Ross Wightman. 2019. PyTorch Image Models. https://github.com/huggingface/pytorch-image-models. https://doi.org/10.5281/zenodo.4414861

[23] Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5188–5197.

[24] Nikolaos-Antonios Ypsilantis, Noa Garcia, Guangxing Han, Sarah Ibrahimi, Nanne Van Noord, and Giorgos Tolias. 2021. The met dataset: Instance-level recognition for artworks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[25] Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. 2021. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11782–11791.

[26] Yu Zhan and Wan-Lei Zhao. 2021. Instance search via instance level segmentation and feature representation. *Journal of Visual Communication and Image Representation* 79 (2021), 103253.