

CiteTracker: Correlating Image and Text for Visual Tracking

Supplementary Materials

Xin Li^{1,†}, Yuqing Huang^{1,2,†}, Zhenyu He^{2,*}, Yaowei Wang^{1,*}, Huchuan Lu³, and Ming-Hsuan Yang^{4,5}

¹Peng Cheng Laboratory ²Harbin Institute of Technology, Shenzhen

³Dalian University of Technology ⁴UC Merced ⁵Yonsei University

This document provides additional information on the experimental implementation and results.

1. Implementation Details

1.1. Vocabulary Construction

We use category information to describe tracking targets as it possesses a certain degree of generalization ability and is more robust to changes in targets. In addition, to provide a more specific and discriminative textual description, we carefully choose attributes that remain consistent over time. Specifically, we use 80 category labels from the MS COCO [3] dataset and three types of object attributes (color, texture, and material) from the OVAD [1] dataset to provide words for describing target states. Table 1 shows the values contained in every vocabulary. The used object categories are commonly seen in daily life. The attributes are selected based on the stability statistics of each attribute on the Got-10K [2] tracking dataset.

1.2. Training Settings.

We first train the image-text conversion model and then use it to generate descriptions to train the complete tracking model. The image-text conversion model is trained on a single GPU with the ViT-B/32 backbone from the CLIP [5] model using the CoCoOp [6] settings. The training data are obtained from the MS COCO [3] dataset and the annotation from the OVAD [1] dataset. The training is performed by minimizing the cross-entropy classification loss using the SGD optimizer with an initial learning rate of $2e-3$, which is decayed by the cosine annealing rule. The model was trained over 10 epochs with a batch size of 1, taking about 0.5 hours.

[†] Equal contribution, * corresponding author

Table 1. Values contained in the vocabularies of category, color, material, and texture.

Type	Label
Category vocabulary	person, bicycle, car, motorcycle, airplane, bus, train, truck, boat, traffic light, fire hydrant, stop sign, parking meter, bench, bird, cat, dog, horse, sheep, cow, elephant, bear, zebra, giraffe, backpack, umbrella, handbag, tie, suitcase, frisbee, skis, snowboard, sports ball, kite, baseball bat, baseball glove, skateboard, surfboard, tennis racket, bottle, wine glass, cup, fork, knife, spoon, bowl, banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake, chair, couch, potted plant, bed, dining table, toilet, tv, laptop, mouse, remote, keyboard, cell phone, microwave, oven, toaster, sink, refrigerator, book, clock, vase, scissors, teddy bear, hair drier and toothbrush.
Color vocabulary	black, blue, brown, gray, green, orange, pink, red, tan, violet, white, and yellow.
Material vocabulary	cement, ceramic, glass, leather, metal, paper, polymers, stone, textile, and wooden.
Texture vocabulary	rough, smooth, and soft.

For the complete tracking model, we use 4 GPUs to train it with a batch size of 64 using the AdamW optimizer [4]. The learning rates for the backbone and other components are $4e-5$ and $4e-4$, respectively. The total training process contains 300 epochs with 60k image pairs per epoch, which takes about 3 days.

2. More Detailed Experimental Results

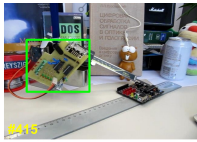



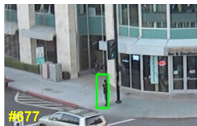
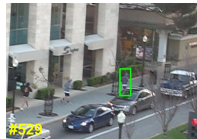

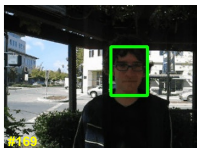
2.1. Visualized Results

Figure 1 shows the visualized results of the OStrack tracker and the proposed algorithm on six challenging sequences.



Figure 1. Visualized results of the proposed algorithm and the OStrack method on six challenging sequences with drastic changes. Our CiteTracker performs well with the aid of the generated text descriptions (shown above each row of pictures), while the OStrack method with solely visual cues struggles with these sequences.

Table 2. **Robustness evaluation in terms of temporal on several selected sequences.** It shows that the proposed algorithm significantly improves tracking performance in terms of the AUC score in the cases where the target appearance is with distractions in the initial image.

Sequence	Challenging initial frame	Interfering factor	OTrack (AUC %)	CiteTracker (AUC %)
Board		Rotation	67.92	89.31
Doll		Distractor	57.38	63.77
DragonBaby		Occlusion	27.76	72.48
Freeman3		Out-of-plane-rotation	54.93	79.4
Human3		Occlusion	57.15	65.2
Human4.2		Occlusion	17.64	48.08
Surfer		Out-of-plane-rotation	18.89	73.24
Trellis		dark light	52.07	70.95
Overall			57.5	61.1

2.2. Robustness Evaluation

Table 2 presents the detailed results and the initial frame of several sequences in terms of the robustness evaluation. It shows that our method has better robustness against interference in the initial frame.

References

- [1] María A. Bravo, Sudhanshu Mittal, Simon Ging, and Brox Thomas. Open-vocabulary attribute detection. In *arXiv*, 2022. 1

- [2] Lianghai Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE TPAMI*, 2019. [1](#)
- [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 *ECCV*, 2014. [1](#)
- [4] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. [2](#)
- [5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. [1](#)
- [6] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. [1](#)