Rethinking Self-supervised Correspondence Learning: A Video Frame-level Similarity Perspective

Jiarui Xu Xiaolong Wang
UC San Diego

Abstract

Learning a good representation for space-time correspondence is the key for various computer vision tasks, including tracking object bounding boxes and performing video object pixel segmentation. To learn generalizable representation for correspondence in large-scale, a variety of self-supervised pretext tasks are proposed to explicitly perform object-level or patch-level similarity learning. Instead of following the previous literature, we propose to learn correspondence using Video Frame-level Similarity (VFS) learning, i.e, simply learning from comparing video frames. Our work is inspired by the recent success in image-level contrastive learning and similarity learning for visual recognition. Our hypothesis is that if the representation is good for recognition, it requires the convolutional features to find correspondence between similar objects or parts. Our experiments show surprising results that VFS surpasses state-of-the-art self-supervised approaches for both OTB visual object tracking and DAVIS video object segmentation. We perform detailed analysis on what matters in VFS and reveals new properties on image and frame level similarity learning.

1. Introduction

Learning visual correspondence across space and time is one of the most fundamental problems in computer vision. It is widely applied in 3D reconstruction, scene understanding, and modeling object dynamics. The research of learning correspondence in videos can be cast into two categories: the first one is learning object-level correspondence for visual object tracking [59, 57, 66], relocalizing the object with bounding boxes along the video; the other one is learning fine-grained correspondence, which is commonly applied in optical flow estimation [30, 20] and video object segmentation [9, 77]. While both lines of research have been extensively explored, most approaches acquire training supervision from simulations or limited human annotations, which increases the difficulty for generalization across different data and tasks.

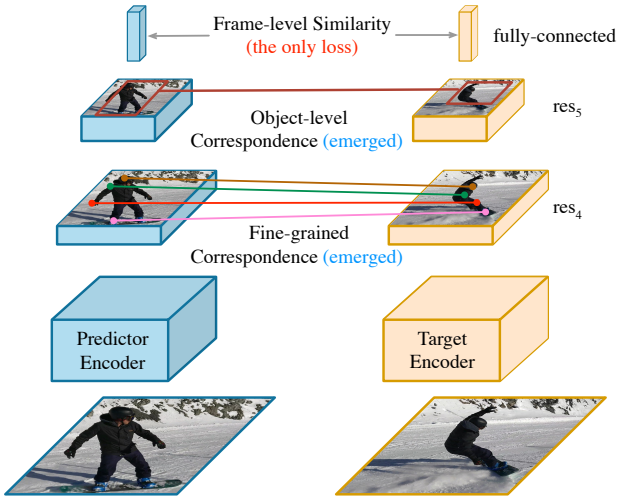

Figure 1. Video Frame-level Similarity (VFS) learning. It compares the fully-connected layer embeddings of frames from the same video for learning. By minimizing the frame-level feature distance, the fine-grained and object-level correspondence can automatically emerge in res_4 and res_5 blocks in the ResNet architecture, without using any explicit tracking-based pretext task.

One way to tackle this problem is to learn representations for correspondence using free temporal supervision signals in videos. Recently, a lot of efforts have been made in self-supervised learning of space-time visual correspondence [67, 71, 33], where different pretext tasks are designed to learn to track pixels or objects in videos. For example, Wang et al. [71] propose to use the cycle-consistency of time (i.e., forward-backward object tracking) as a supervisory signal for learning. Building on this, Jabri et al. [33] combine the cycle-consistency of time with patch-level similarity learning and achieve a significant improvement in learning correspondence. Given this encouraging result, we take a step back and ask the questions: Do we really need to design self-supervised object (or patch) tracking task explicitly to learn correspondence? Can image-level similarity learning alone learn the correspondence?

Recent development in image-level similarity learning (e.g., contrastive learning) has shown the self-supervised