Spatial-then-Temporal Self-Supervised Learning for Video Correspondence

Rui Li   Dong Liu

University of Science and Technology of China, Hefei, China

liruid@mail.ustc.edu.cn, dongeliu@ustc.edu.cn

Abstract

In low-level video analyses, effective representations are important to derive the correspondences between video frames. These representations have been learned in a self-supervised fashion from unlabeled images or videos, using carefully designed pretext tasks in some recent studies. However, the previous work concentrates on either spatial-discriminative features or temporal-repetitive features, with little attention to the synergy between spatial and temporal cues. To address this issue, we propose a spatial-then-temporal self-supervised learning method. Specifically, we firstly extract spatial features from unlabeled images via contrastive learning, and secondly enhance the features by exploiting the temporal cues in unlabeled videos via reconstructive learning. In the second step, we design a global correlation distillation loss to ensure the learning not to forget the spatial cues, and a local correlation distillation loss to combat the temporal discontinuity that harms the reconstruction. The proposed method outperforms the state-of-the-art self-supervised methods, as established by the experimental results on a series of correspondence-based video analysis tasks. Also, we performed ablation studies to verify the effectiveness of the two-step design as well as the distillation losses.

1. Introduction

Learning representations for video correspondence is a fundamental problem in computer vision, which is closely related to different downstream tasks, including optical flow estimation [10, 17], video object segmentation [3, 37], keypoint tracking [53], etc. However, supervising such a representation requires a large number of dense annotations, which is unaffordable. Thus, most approaches acquire information from simulations [10, 33] or limited annotations [39, 55], which result in poor generalization in dif-
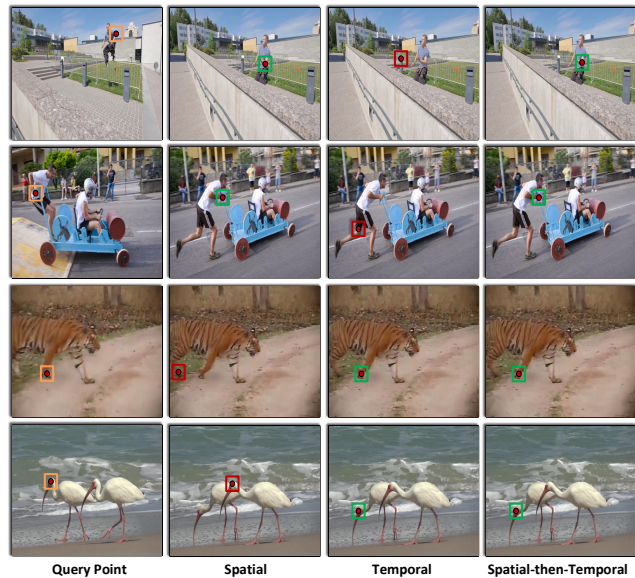
Figure 1. **Matching results given the query point.** While temporal-repetitive features fail to handle video correspondence with dramatic appearance changes and deformations (the first two rows), spatial-discriminative features are incompetent to recognize the temporal repetition and are misled by distractors with similar appearance (the last two rows). The green/red bounding boxes indicate correct/wrong matching. (**Zoom in for best view**)

ferent downstream tasks. Recently, self-supervised feature learning is gaining significant momentum. Several pretext tasks [15, 19, 24, 27, 48, 52] are designed, mostly concentrating on either spatial feature learning or temporal feature learning for space-time visual correspondence.

With the objective of learning the representations that are invariant to the appearance changes, **spatial feature learning** provides video correspondence with discriminative and robust appearance cues, especially when facing severe temporal discontinuity, i.e., occlusions, appearance changes, and deformations. Most recently, as mentioned in [50], the contrastive models [6, 15, 52] pre-trained on image data show competitive performance against dedicated methods for video correspondence. Thus, it is shown to be a better way for learning spatial representations in terms of quality and data efficiency, compared with the meth-

ods [19, 27, 40, 47, 54, 59] using large-scale video datasets for training. Though getting promising results, as shown in Figure 1, the learned spatial representations are misled by the distractors with similar appearance, which indicates the poor ability to recognize the temporal pattern.

In another line, **temporal feature learning** focuses on learning the temporal-repetitive features that occur consistently over time. With the temporal consistency assumption [2], the pixel repetition across video motivates recent studies to exploit the temporal cues via a reconstruction task [24, 30, 34, 58], where the query pixel in the target frame can be reconstructed by leveraging the information of adjacent reference frames within a local range. Then a reconstruction loss is applied to minimize the photometric error between the raw frame and its reconstruction. Nevertheless, temporal-repetitive features highly depend on the consistency of the pixels across the video. It thus can be easily influenced by the temporal discontinuity caused by the dramatic appearance changes, occlusions, and deformations (see Figure 1).

In light of the above observation, we believe the video correspondence relies on both spatial-discriminative and temporal-repetitive features. Thus, we propose a novel spatial-then-temporal pretext task to achieve synergy between spatial and temporal cues. Specifically, we firstly learn spatial features from unlabeled images via contrastive learning and secondly improve the features by exploiting the temporal repetition in unlabeled videos with frame reconstruction. While such an implementation brings together the advantages of spatial and temporal feature learning, there are still some problems. First of all, the previous studies [18, 19, 27, 47, 54] propose to learn coarse-grained features for video correspondence. With the objective of reconstructive learning, the video frames need to be down-sampled to align with the coarse-grained features [24], resulting in severe temporal discontinuity for the pixels to be reconstructed. Thus, the frame reconstruction loss becomes invalid. Second, in the context of sequential training, directly training with only new data and objective functions will degrade the discriminative features learned before.

To tackle the first problem, we firstly exploit temporal cues by frame reconstruction at different pyramid levels of the encoder. We observe that temporal repetition benefits from relatively smaller down-sampling rate. Hence, the model will learn better temporal-persistent features at the fine-grained pyramid level. To distillate the knowledge from it, we design a local correlation distillation loss that supports explicit learning of the final correlation map in the region with high uncertainty, which is achieved by taking the more fine-grained local correlation map as pseudo labels. This leads to better temporal representations on the coarse feature map. At the same time, we regard the model pre-trained in the first step as the teacher. Then a global

correlation distillation loss is proposed to retain the spatial cues. Eventually, we can obtain better temporal representations without losing the discriminative appearance cues acquired in the first step.

To sum up, our main contributions include: (i) We propose a spatial-then-temporal pretext task for self-supervised video correspondence, which achieves synergy between spatially discriminative and temporally repetitive features. (ii) We propose the local correlation distillation loss to facilitate the learning of temporal features in the second step while retaining the appearance sensitivity learned in the first step by the proposed global correlation distillation loss. (iii) We verify our approach in a series of correspondence-related tasks. Our approach consistently outperforms previous state-of-the-art self-supervised methods and is even comparable with task-specific fully-supervised algorithms.

2. Related Work

Self-supervised representation learning for video correspondence. Recent approaches focus on learning the dense representations from unlabeled videos in a self-supervised manner for video correspondence, which has proceeded along two different dimensions: reconstruction-based methods [24, 25, 27, 46, 47] and cycle-consistency-based methods [19, 48, 59]. In the first type, a query point is reconstructed from adjacent frames while the latter performs forward-backward tracking with the objective of minimizing the cycle inconsistency. Though getting promising results, the learned features still lack the capability of spatial discriminating. Meanwhile, VFS [54] try to learn the spatial and temporal representation through a frame-wise contrastive loss while the methods in [1, 47] try to realize it by exploiting the inter-video constraint, which may result in sub-optimal performance. The concurrent work SFC [18] proposes a two-stream network that learns the semantic and fine-grained features with two separate models. Compared to SFC [18] which does not consider the temporal pattern, our method is able to learn both spatial and temporal features with only a single model.

Self-supervised optical flow for video correspondence. Apart from learning powerful representations for video correspondence, another line of studies [10, 42] approaches the problem by directly regressing the optical flow produced by synthetic datasets, thus suffering from severe domain shifts. To address the problem, the study in [58] first introduces a method for learning optical flow without using any annotations by leveraging brightness constancy with photometric reconstruction loss, and Unflow [34] further improves it with occlusions reasoning. Though making great progress, it still does not generalize well in real scenarios. In [22, 29, 30], they improve performance by utilizing the optical flow predictions from the teacher model to guide the learning of the student model in the region with occlu-
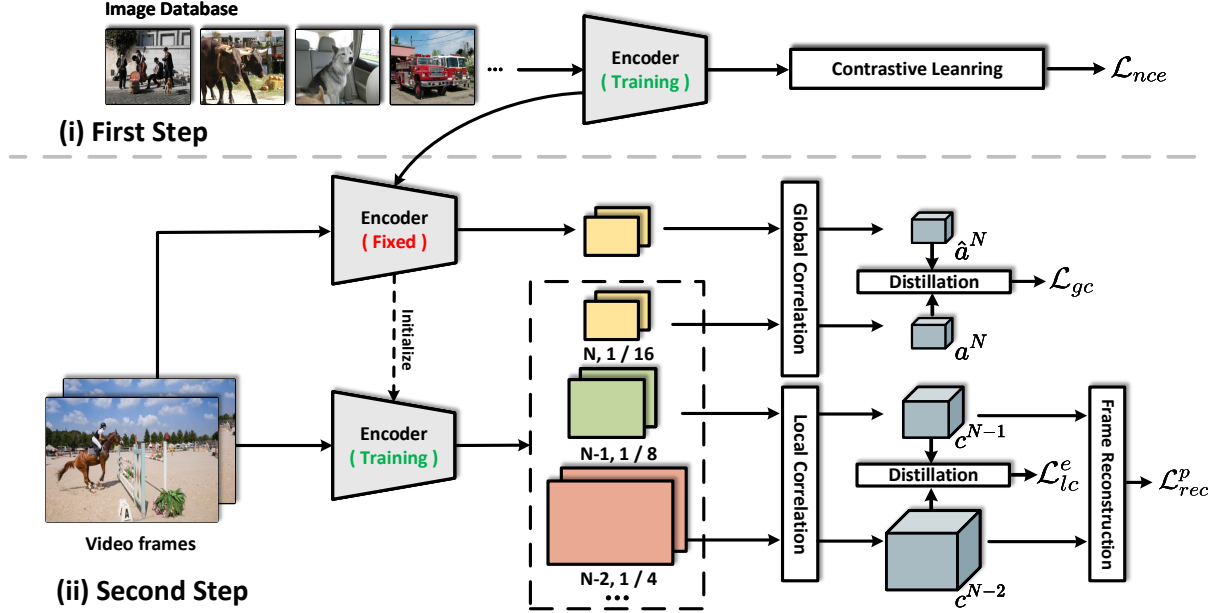
Figure 2. **Overview of the two-step design in our spatial-then-temporal pretext task. In the second step, we fix the encoder trained in the first step as the teacher.** Following prior studies [19, 47, 54], we focus on learning dense features at the level $N - 1$ (the level indicates the layer of the ResNet architecture). Based on discriminative appearance features already learned in the first step, we exploit temporal cues at each pyramid level of the encoder by reconstructive learning. Each query pixel in the target frame can be reconstructed by a weighted sum of pixels within a local window in the reference frame. Then, we distillate the knowledge from a more fine-grained pyramid level (i.e. $N - 2$) by aligning the local correlation maps in the region with high uncertainty, which leads to better temporal features at level $N - 1$. Meanwhile, we compute the global correlation to take account into the large motion between frames, and a novel global correlation distillation loss is devised between global correlation maps to retain spatially discriminant features learned in the first step.

sions. In this paper, we propose a novel self-distillation loss to combat temporal discontinuity in the context of self-supervised feature learning.

Self-supervised spatial representation learning. Self-supervised spatial feature learning recently gets the promising result with contrastive learning. In an early work [51], contrastive learning is formulated as an instance discrimination task, which requires the model to return low values for similar pairs and high values for dissimilar pairs. The performance is further improved by creating a dynamic memory-bank [15], introducing online clustering [4], and avoiding the use of negative pairs [8, 14]. Furthermore, the methods in [49,52,56] propose various pretext tasks to adapt the contrastive learning to dense prediction tasks. Even though showing superior performance for video correspondence [50], the contrastive model still struggles to figure out the temporal pattern between video frames.

3. Method

In this section, we firstly introduce spatial feature learning in the context of contrastive learning. Then, the temporal feature learning is improved with our pyramid frame reconstruction and local correlation distillation. In the end, we present our spatial-then-temporal pretext task along with the global correlation distillation used for retaining the

learned spatial cues in the first step.

3.1. Spatial Feature Learning

Spatial feature learning aims to learn the discriminative features of different objects, which provides video correspondence with robust appearance cues, especially when facing temporal discontinuity. Most recently, spatial feature learning is gaining significant momentum due to the advancements in contrastive learning. We begin by briefly reviewing instance discrimination objective in contrastive learning. Specifically, the query vector $q \in \mathbb{R}^d$ and a set of key vectors $\mathcal{K} = \left\{ k^+, k_1^-, k_2^-, \ldots, k_K^- \right\}$ which consists of one positive key $k^+ \in \mathbb{R}^d$ and K negative keys $\mathcal{K}^- = \left\{ k_j^- \right\}$, are encoded by the encoder ϕ plus an MLP head, and are further processed by l_2-normalization. The query and its positive key are generated from the same image with two different augmentations, while the negative keys refer to other instances. The objective of instance discrimination is to maximize the similarity between the query q and the positive key k^+ while the remaining query distinct to all negative keys $\mathcal{K}^-$. Thus, a contrastive loss is presented in InfoNCE [38] with a softmax formulation:

$$\mathcal{L}_{\text{nce}} = -\log \frac{\exp\left(q^T \cdot k^+ / \tau_c \right)}{\exp\left(q^T \cdot k^+ / \tau_c \right) + \sum_{i=1}^K \exp\left(q^T \cdot k_i^- / \tau_c \right)},$$

(1)

(b) Matching result at level (N-1)
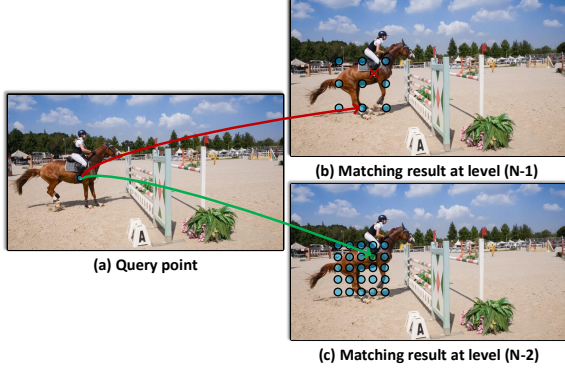
(c) Matching result at level (N-2)

Figure 3. **Visualization of the local correlation of the query point at different pyramid levels.** The red arrow indicates the wrong match given the query point. In (b), the frame reconstruction becomes invalid due to the absence of the query point in the reference frame. (**Zoom in for best view**)

where the similarity is measured via dot product, and τ_c is the temperature hyper-parameter. MoCo [15] builds a dynamic memory bank to maintain a large number of negative samples with a moving-averaged encoder. DetCo [52] further improves the contrastive loss $\mathcal{L}_{nce}$ by introducing global and local contrastive learning to enhance local representations. In this paper, we adopt DetCo to learn appearance features for most of our experiments.

3.2. Temporal Feature Learning

Temporal feature learning aims to learn temporally repetitive features between video frames, which is closely related to the recent studies of motion estimation [10, 30, 42]. Based on the temporal consistency assumption [2], the pixel repetition encourages the studies in [24, 30, 34, 48] to learn the video correspondence by reconstructive learning, where each query pixel in the target frame can be reconstructed by leveraging the information of adjacent reference frames with a limited range. More specifically, the target and reference frame $I_t, I_r \in \mathbb{R}^{H \times W \times 3}$ are projected into a pyramid pixel embedding space by the encoder ϕ. We denote these embeddings as $F_t^l, F_r^l \in \mathbb{R}^{h^l w^l \times d^l}$, where $l \in \{0, 1, \ldots, N\}$ is the index of each pyramid level and the bigger number represents the coarser pyramid level. The s^l denotes the stride up to each pyramid level and $h^l = H/s^l, w^l = W/s^l$. From this point of view, each feature point is aligned with the center pixel of strided convolution layers [24] and we can sample it to get the aligned target and reference frame $\hat{I}_t^l, \hat{I}_r^l \in \mathbb{R}^{h^l w^l \times 3}$. For each query pixel i in $\hat{I}_t^l$ at the pyramid level l, we can calculate the local correlation [25] $c^l(i, j)$ w.r.t. each key pixel j in the reference frame within a local window centered at i, considering the nature of temporal coherence in the video:

$$c^l(i,j) = \frac{\exp\left(F_t^l(i) \cdot F_r^l(j)/\tau\right)}{\sum_n \exp\left(F_t^l(i) \cdot F_r^l(n)/\tau\right)}, i \in \{1,.,h^l w^l\}, j, n \in \mathcal{N}(i) \ , \tag{2}$$

where $\mathcal{N}(i)$ is the index set in the reference frame with a limited range of R^l for the pixel i. Then the query pixel i in the target frame can be reconstructed by a weighted sum of pixels in $\mathcal{N}(i)$, according to the local correlation map $c^l \in \mathbb{R}^{h^l w^l \times (2R^l+1)^2}$:

$$\overline{I}_t^l(i) = \sum_{j \in \mathcal{N}(i)} c^l(i,j) \hat{I}_r^l(j) \ . \tag{3}$$

Then the reconstruction loss $\mathcal{L}_{rec}$ is defined as L_1 distance between $\hat{I}_t^l$ and $\overline{I}_t^l$ at the pyramid level l:

$$\mathcal{L}_{rec} = \left\| \hat{I}_t^l - \overline{I}_t^l \right\|_1 \ . \tag{4}$$

since the previous methods focus on learning dense representations at level $N-1$, we set l to $N-1$ by default, which is regarded as our baseline.

Pyramid frame reconstruction. We can further exploit more fine-grained temporal cues of the raw frames. As observed in Figure 2, we obtain a pair of pyramid features $\{F_t^l\}_{l=1}^{N-1}, \{F_r^l\}_{l=1}^{N-1}$, which is utilized to calculate the pyramid local correlation maps $\{c^l\}_{l=1}^{N-1}$ with Eq. (2). Then, we can reconstruct the target frame at each pyramid level using Eq. (3). We devise the pyramid reconstruction loss as:

$$\mathcal{L}_{rec}^p = \sum_l \left\| \hat{I}_t^l - \overline{I}_t^l \right\|_1 \ . \tag{5}$$

Local correlation distillation. For video correspondence, previous methods learn the coarse-grained features with large down-sampling rate, which would harm the pixel repetition between video frames. As observed in Figure 3, the query pixel in the target frame is absent in level $N - 1$, resulting in an obvious shortcut that the query pixel tries to match any other pixel with a similar color to reduce the frame reconstruction loss (red arrow). However, the missing pixel can be found at level $N - 2$ with higher resolution (see Figure 3 (c)). Hence, the temporal feature learning performed at this pyramid level would contribute to better temporal features. Here, we design a novel local correlation distillation loss to distillate the knowledge from more fine-grained pyramid levels. Specifically, we firstly leverage correlation down-sampling [42] on c^{N-2} to get pseudo labels $\hat{c}^{N-1}$ with the same size as c^{N-1}. Then we propose a local correlation distillation loss $\mathcal{L}_{lc}$ to minimize the mean squared error between them. By applying the novel self-distillation, we can eventually learn better temporal-related features at relatively coarse pyramid level with low computational costs of inference.

Entropy-based selection. The correlation of each query w.r.t. reference frame indicates more uncertainty when having smooth distribution, which has a higher probability of suffering the temporal discontinuity. Thus, it should be paid
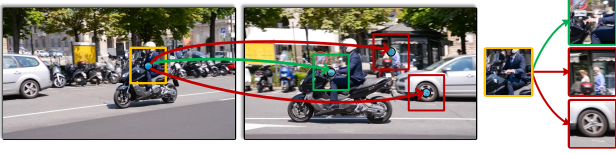
Figure 4. **Visualization of the global correlation of the query point for large motion.** The red arrow indicates the wrong match given the query point, and the box indicates the patch centered at each point. The model trained in the first step can handle this case with learned discriminative appearance features.

more attention to when applying local correlation distillation. We calculate the entropy for each query i:

$$\mathcal{H}(i) = \sum_j -\log c^{N-1}(i,j), i \in \{1,.,h^{N-1}w^{N-1}\}, j \in \mathcal{N}(i) \quad (6)$$

then we obtain a mask $m \in \{0,1\}^{h^{N-1}w^{N-1}}$ to filter out the region with lower entropy by setting a threshold T. The local correlation distillation loss with entropy selection is defined as:

$$\mathcal{L}_{\mathrm{lc}}^e = \sum_i m(i) \cdot \left\| c^{N-1}(i) - \hat{c}^{N-1}(i) \right\|_2^2 . \quad (7)$$

Eventually, our training loss of temporal feature learning is defined as $\mathcal{L}_{\mathrm{t}} = \mathcal{L}_{\mathrm{rec}}^p + \alpha \mathcal{L}_{\mathrm{lc}}^e$.

3.3. Spatial-then-Temporal Feature Learning

In this section, we will formally introduce our spatial-then-temporal self-supervised learning for video correspondence. Besides, we further devise some variants of training configurations for learning the spatiotemporal features for comparisons.

(i) **Spatial-then-Temporal**: We first train the encoder ϕ with unlabeled image data via contrastive learning. Next, we concentrate on learning temporal-related features with $\mathcal{L}_t$ on the video dataset. However, the model will not be able to capture spatial cues when training it with new object functions and data. Consequently, it may fail to deal with the dramatic appearance changes and deformations with large motion like before (see Figure 4). Thus, we introduce a global correlation distillation loss to tackle the problem. Precisely, we first fix the feature encoder ϕ trained in the first step as the teacher denoted as ϕ_t. For each query point $F_t^N(i)$ at pyramid level N, we compute the global correlation $a^N(i,j)$ using a softmax over similarities w.r.t. all keys in the reference frame, i.e.,

$$a^N(i,j) = \frac{\exp\left(F_t^N(i) \cdot F_r^N(j)/\tau\right)}{\sum_n \exp\left(F_t^N(i) \cdot F_r^N(n)/\tau\right)}, i,j,n \in \{1,\dots,h^N w^N\} , \quad (8)$$

we generate the pseudo labels $\hat{a}^N$ using Eq. (8) with ϕ_t. The global correlation distillation loss $\mathcal{L}_{\mathrm{gc}}$ is defined to minimize the mean squared error between a^N and $\hat{a}^N$:

$$\mathcal{L}_{\mathrm{gc}} = \left\| a^N - \hat{a}^N \right\|_2^2 , \quad (9)$$

by aligning the a^N with $\hat{a}^N$, the student tries to retain the appearance features learned in the first step. It is worth noting that compared with the local correlation that only considers the video correspondence in a local range, the global correlation compares the query with all instances in the reference frame. Thus, it also provides more valuable training samples for distilling the spatially discriminant features with $\mathcal{L}_{gc}$. The final loss of training on video data is $\mathcal{L} = \mathcal{L}_{\mathrm{t}} + \beta \mathcal{L}_{\mathrm{gc}}$

(ii) **Spatiotemporal**: We can also learn the spatial and temporal features jointly with both $\mathcal{L}_{nce}$ and $\mathcal{L}_t$. The training loss is summarized as $\mathcal{L} = \mathcal{L}_{\mathrm{nce}} + \gamma \mathcal{L}_{\mathrm{t}}$.

(iii) **Temporal-then-Spatial**: Then, we try to reverse the training order of (i). However, without learning the robust appearance features first, temporal feature learning may be misled by the distractors when facing severe temporal discontinuity.

The comparisons between different training configurations will be discussed in the experiments, and the spatial-then-temporal is used as the default training configuration.

4. Experiments

We verify the merit of our method in a series of correspondence-related tasks, including semi-supervised video object segmentation, human part propagation, and pose keypoint tracking. This section will firstly introduce our experiment settings, including implementation and evaluation details. Then detailed ablation studies are performed to explain how each component of our method works. We finally report the performance comparison with state-of-the-art methods to further verify the effectiveness of our method.

4.1. Implementation Details

Backbone. We exploit the encoder ϕ with both ResNet-18 and ResNet-50 [16] for self-supervised training. When training at the first step, the encoder plus an MLP head project each image to a vector. Then, for temporal feature learning, we reduce the stride up to layer res$_4$ from 16 to 8 (in Figure 2) to increase the spatial resolution of feature maps by a factor of 2. We further reduce the stride to 4 to compare with prior studies [24, 26, 35] which perform self-supervised training on high-resolution feature maps.

Training details. We first train our model using contrastive loss for 200 epochs on ImageNet [9] following most augmentations and hyper-parameters settings of [15, 52]. Then we perform temporal feature learning on the training set of YouTube-VOS [55] which consists of 3.5k videos. In this stage, the video frame is resized into 256×256, and channel-wise dropout with probability of 0.5 in Lab color space [24, 25] is adopted as the information bottleneck. We train the encoder with the stride of 8 for 90k/45k iterations

with a mini-batch of 128/64 for ResNet-18/ResNet-50, using Adam as our optimizer. The initial learning rate is set to 1e-4 with a cosine (half-period) learning rate schedule. The range R of computing the local correlation map is set to 12/6 on layer res_3/res_4. All experiments are performed with 4 GTX-3090 GPUs. We include more training details in the supplementary material.

Evaluation. We directly utilize the unsupervised pre-trained model as the feature extractor without any fine-tuning. Given the input frame with spatial resolution of $H \times W$, the evaluation is realized on the res_4 layer with a spatial resolution of $\frac{H}{8} \times \frac{W}{8}$ (stride=8) or $\frac{H}{4} \times \frac{W}{4}$ (stride=4). To propagate the semantic labels from the initial ground-truth annotation, the recurrent inference strategy is applied following recent studies [19, 24, 54]. More specifically, the semantical label of the first frame, as well as previous predictions, are propagated to the current frame with the help of affinity computed between the feature maps of video frames. We evaluate our method over three downstream tasks including semi-supervised video object segmentation on DAVIS-2017 [39], human part propagation on VIP [60], and pose keypoint tracking on JHMDB [21].

4.2. Ablation Studies

The ablation studies are performed with semi-supervised video object segmentation on the DAVIS-2017 validation set [39]. We use the mean of region similarity $\mathcal{J}_m$, mean of contour accuracy $\mathcal{F}_m$ and their average $\mathcal{J}\&\mathcal{F}_m$ as the evaluation metrics. We set the stride until layer res_4 to 8 for training and evaluation if not specified.

Temporal feature learning. We examine how each design in our temporal feature learning impacts the overall performance, which is shown in Table 1. From the table, we can see leveraging more temporal cues with per-level frame reconstruction leads to an improvement in the range of 0.7%. With the guidance of a better local correlation map at fine-grained pyramid level, $\mathcal{L}_{lc}$ boosts up the accuracy from 65.6% to 68.4%. Moreover, enforcing the local correlation distillation to focus on the region with higher entropy leads to a performance gain in the range of 0.9%. By fusing the above components, the performance reaches 69.3%.

Different training configurations. We then study the effect of three different ways as defined in Sec. 3.3, which is shown in Table 2.

(i) Spatiotemporal. We use data from both ImageNet and YouTube-VOS for training and find the Spatiotemporal will result in unsatisfactory performance. This is attributed to the conflicts of two object functions in terms of training hyper-parameters and augmentations. More specifically, the contrastive model benefits from complicated data augmentations [7] for learning invariant features within appearance changes. However, the pixel-level distortion may hinder the lower-level convolution features from finding the fine-

Frame Recons.	Pyramid Frame Recons.	Local Corr. Distillation	Entropy Selection	DAVIS $\mathcal{J}\&\mathcal{F}_m \uparrow$
✓				64.9
✓	✓			65.6
✓	✓	✓		68.4
✓	✓	✓	✓	**69.3**

Table 1. **Ablation study for temporal feature learning with ResNet-18**. All models are learnt on YouTube-VOS [55] and evaluated on DAVIS-2017 [39].

Configuration	Dataset	Loss Function	DAVIS $\mathcal{J}\&\mathcal{F}_m \uparrow$
Spatial only	I	$\mathcal{L}_{nce}$	66.5
Temporal only	YTV	$\mathcal{L}_t$	67.5
Spatiotemporal	I + YTV	$\mathcal{L}_{nce} + \mathcal{L}_t$	68.6
Spatial-then-Temporal	I→YTV	$\mathcal{L}_{nce} \rightarrow \mathcal{L}_t$	69.4
		$\mathcal{L}_{nce} \rightarrow \mathcal{L}_t + $ EWC	69.6
		$\mathcal{L}_{nce} \rightarrow \mathcal{L}_t + \mathcal{L}_{gc}$	**71.3**
Temporal-then-Spatial	YTV→I	$\mathcal{L}_t \rightarrow \mathcal{L}_{nce}$	66.9
		$\mathcal{L}_t \rightarrow \mathcal{L}_{nce} + $ EWC	67.0

Table 2. **Ablation study for different training configurations with ResNet-50**. "→" indicates the switch in our two-step training in terms of datasets and loss functions. I: ImageNet [9]. YTV: YouTube-VOS [55].

grained motion between video frames [54].

(ii) Spatial-then-Temporal. The above problem can be alleviated by our proposed Spatial-then-Temporal learning where each loss is optimized individually with its own hyper-parameters and augmentations. Our method without using $\mathcal{L}_{gc}$ can already reach 69.4%, and the $\mathcal{L}_{gc}$ further boosts up the performance from 69.4% to 71.3%. Here we also compare our method with the general continual model EWC [23] based on weight constraints. Our method still gets better results, which demonstrates the effectiveness of our proposed global correlation distillation.

(iii) Temporal-then-Spatial. Directly fine-tuning the temporal model with $\mathcal{L}_{nce}$ decreases the performance from 67.5% to 66.9%. Unexpectedly, using EWC as a regularization strategy does not provide much benefit.

Further analysis. We give a further analysis here based on the above experiments. On the one hand, temporal feature learning helps to recognize the temporal pattern between video frames, which is unable to accomplish by training an appearance model. As can be seen in the last two rows of Figure 1, the appearance model trained with $\mathcal{L}_{nce}$ is misled by two patches at different locations which have a similar appearance, while the model trained with $\mathcal{L}_t$ tends to learn better temporal representations (see Temporal and Spatial-then-Temporal). However, in the first two rows of Figure 1, the model trained only with $\mathcal{L}_t$ fails to get accurate correspondence when facing severe temporal discontinuity, e.g., occlusions, appearance changes, deformations, while the model trained with $\mathcal{L}_{nce}$ is able to correct the mistakes by tracking the points based on learned discriminative and

Method	Sup.	Backbone	Stride	Training Dataset		$\mathcal{J}\&\mathcal{F}_m \uparrow$	$\mathcal{J}_m \uparrow$	$\mathcal{F}_m \uparrow$
				Image	Video			
MoCo [15]		ResNet-18	8	ImageNet	-	60.8	58.6	63.1
SimSiam [8]		ResNet-18	8	ImageNet	-	62.0	60.0	64.0
Colorization [46]		ResNet-18	8	-	Kinetics	34.0	34.6	32.7
CorrFlow [25]		ResNet-18	8	-	OxUvA	50.3	48.4	52.2
MuG [31]		ResNet-18	8	-	OxUvA	54.3	52.6	56.1
UVC [27]		ResNet-18	8	COCO	Kinetics	59.5	57.7	61.3
ContrastCorr [47]		ResNet-18	8	COCO	TrackingNet	63.0	60.5	65.5
VFS [54]		ResNet-18	8	-	Kinetics	66.7	64.0	69.4
CRW [19]		ResNet-18	8	-	Kinetics	67.6	64.8	70.2
JSTG [59]		ResNet-18	8	-	Kinetics	68.7	65.8	71.6
CLSC [40]		ResNet-18	8	-	Kinetics	<u>70.5</u>	<u>67.4</u>	**73.6**
SFC [18]		ResNet-18	8	-	YTV	67.7	64.7	70.5
DUL [1]		ResNet-18	8	-	YTV	69.3	67.1	71.6
CLTC [20]		ResNet-18	8	-	YTV	70.3	67.9	72.6
Ours (Temporal)		ResNet-18	8	-	YTV	69.3	66.7	71.9
Ours		ResNet-18	8	ImageNet	YTV	**70.7**	**68.0**	<u>73.5</u>
MAST [24]		ResNet-18	4	-	YTV	65.5	63.3	67.6
MAMP [35]		ResNet-18	4	-	YTV	69.7	68.3	71.2
LIIR [26]		ResNet-18	4	-	YTV	72.1	69.7	74.5
Ours (Temporal)		ResNet-18	4	-	YTV	<u>72.5</u>	<u>69.9</u>	<u>75.1</u>
Ours		ResNet-18	4	ImageNet	YTV	**73.6**	**70.7**	**76.4**
MoCo [15]		ResNet-50	8	ImageNet	-	65.4	63.2	67.6
SimSiam [8]		ResNet-50	8	ImageNet	-	66.3	64.5	68.2
TimeCycle [48]		ResNet-50	8	-	VLOG	40.7	41.9	39.4
UVC [27]		ResNet-50	8	COCO	Kinetics	56.3	54.5	58.1
VINCE [13]		ResNet-50	8	-	Kinetics	65.6	63.4	67.8
VFS [54]		ResNet-50	8	-	Kinetics	<u>68.9</u>	<u>66.5</u>	<u>71.3</u>
Ours		ResNet-50	8	ImageNet	YTV	**71.3**	**68.5**	**74.0**
Ours		ResNet-50	4	ImageNet	YTV	**74.1**	**71.1**	**77.1**
SFC† [18]		ResNet-18 + ResNet-50	8	ImageNet	YTV	<u>71.2</u>	<u>68.3</u>	<u>74.0</u>
Ours†		ResNet-18 + ResNet-50	8	ImageNet	YTV	**72.1**	**69.5**	**74.7**
Supervised [16]	✓	ResNet-18	8	ImageNet	-	62.9	60.6	65.2
Supervised [16]	✓	ResNet-50	8	ImageNet	-	66.0	63.7	68.4
OnAVOS [45]	✓	ResNet-38	-	I + C + P	D	65.4	61.6	69.1
OSVOS-S [32]	✓	VGG-16	-	I + P	D	68.0	64.7	71.3
FEELVOS [44]	✓	Xception-65	-	I + C	D + YTV	71.5	69.1	74.0

Table 3. **Quantitative results for video object segmentation on validation set of DAVIS-2017** [39]. We show the results of state-of-the-art self-supervised methods and some supervised methods for comparison. Abbreviations for some datasets and their sizes – number of images or duration of videos – are: (h stands for hours). I:ImageNet [9] (1.28m). C:COCO [28] (30k). O:OxUvA [43] (14h). T:TrackingNet [36] (300h). K:Kinetics [5] (800h). V:VLOG [12] (344h). YTV:YouTube-VOS [55] (5h). D:DAVIS-2017 [39]. P:PASCAL-VOC [11]. † indicates using the late fusion [18] of two separate models for inference.

robust spatial cues (see Spatial and Spatial-then-Temporal).

4.3. Comparison with State-of-the-Art Methods

Results for video object segmentation. We compare our method against previous self-supervised methods in Table 3. For a fair comparison, we report both results of setting the stride until layer res_4 to 4 and 8. Our method achieves state-of-the-art performance using both ResNet-18 and ResNet-50. For ResNet-18, our method with a stride of 8 achieves 70.7%, surpassing all baselines using the same architecture. Benefiting from exploiting more fine-grained temporal cues for temporal feature learning by setting the stride to 4, the performance of our method reaches 73.6%, leading to a performance gain of 1.5% over LIIR [26]. For ResNet-50, our method still outperforms VFS [54] by 2.4%. It is worth noting that the studies in [19, 27, 40, 47, 54, 59]

are all pre-trained on large-scale video datasets, i.e., Kinetics [5], TrackingNet [36], while our method adopts a small video dataset plus an image dataset which has a smaller data size than video. Besides, our model trained with temporal feature learning reaches 69.3%/72.5% and still leads the performance over other methods with the same training data (YTV). Compared with SFC [18] which fuses learned semantic and fine-grained features with separate models, our single model with ResNet-50 already leads the performance, which demonstrates the importance of achieving synergy between spatial and temporal cues. Following the inference strategy of SFC, we fuse the spatial features (ResNet-50) learned with $\mathcal{L}_{nce}$ and temporal features (ResNet-18) learned with $\mathcal{L}_t$, the performance reaches 72.1% and surpasses SFC by 0.9%. More remarkably, Our method even outperforms some task-specific fully-supervised algorithms [32, 44, 45].

(a) Video Object Segmentation (1-4 objects)

(b) Human Part Propagation (20 parts)

(c) Pose Keypoint Tracking (15 key points)

Figure 5. Qualitative results for label propagation. Given the first frame (blue border) with different annotations, we propagate it to the current frame without fine-tuning our correspondence model. (a) Video object segmentation on DAVIS-2017 [39]. (b) Human part propagation on VIP [60]. (c) Pose keypoint tracking on JHMDB [21].

Methods	Sup.	VIP	JHMDB	
		mIoU ↑	PCK@0.1 ↑	PCK@0.2 ↑
TimeCycle [48]		28.9	57.3	78.1
UVC [27]		34.1	58.6	79.6
SFC [18]		34.0	59.3	80.8
CRW [19]		38.6	59.3	80.3
ContrastCorr [47]		37.4	61.1	80.8
VFS [54]		39.9	60.5	79.5
CLTC [20]		37.8	60.5	82.3
JSTG [59]		40.2	61.4	**85.3**
CLSC [40]		<u>40.8</u>	<u>61.7</u>	82.6
Ours		**41.0**	**63.1**	<u>82.9</u>
ResNet-18 [16]	✓	31.9	53.8	74.6
ATEN [60]	✓	37.9	-	-
Thin-Slicing Net [41]	✓	-	68.7	92.1

Table 4. **Quantitative results for human part propagation and pose keypoint tracking.** We show results of state-of-the-art self-supervised methods and some supervised methods for comparison.

Results for human part propagation. Next, we evaluate our method for human part tracking. Experiments are conducted on the validation set of VIP [60], which consists of 50 videos with 19 human semantic parts, requiring more precise matching than DAVIS-2017 [39]. Following [60], we adopt mean intersection-over-union (mIoU) as the evaluation metric and resize the video to 560×560. All models are set to ResNet-18 with a stride of 8 for fair comparisons. The results are shown in Table 4. Our method still surpasses previous state-of-the-art. Notably, our model outperforms ATEN [60] which is specifically designed for the task using human annotations. Figure 5 (b) depicts some visualization results on several representative videos. Our approach could output tight boundaries around the multiple targets.

Results for pose keypoint tracking. We then make a performance comparison on the downstream task of human pose tracking. We conduct the experiments on the validation of JHMDB [21] which has 268 videos. The annotations consist of 15 body joints for each person. The probability of correct keypoint [57] is utilized here to examine the accuracy with different thresholds. Following the evaluation protocol of [19, 27], we resize the video frames to 320×320. The results in Table 4 show a consistent performance gain over previous methods, which successfully demonstrates the transferability of our method to different downstream tasks. The visualization results in Figure 5 (c) show the robustness of our approach to various challenges.

5. Conclusions

In this paper, we present a new spatial-then-temporal pretext task for learning the representations of video correspondence. The key idea underpinning the proposed method is to achieve the synergy between spatial and temporal cues. Specifically, we firstly train the model using contrastive loss with still images and then perform reconstructive learning to exploit temporal cues with a video dataset. In the second step, we devise a global correlation distillation loss to retain the spatially-discriminative features learned in the first step. At the same time, we propose a local correlation distillation loss to alleviate the problem of temporal discontinuity. Extensive experiments on three kinds of downstream tasks validate the effectiveness of the proposed two-step task and loss functions. We hope that this work may provide a new perspective for video correspondence and brings us one step closer to the accurate correspondence for real-world videos.

References

[1] Nikita Araslanov, Simone Schaub-Meyer, and Stefan Roth. Dense unsupervised learning for video segmentation. In *NeurIPS*, pages 25308–25319, 2021. 2, 7

[2] Michael J Black and Padmanabhan Anandan. A framework for the robust estimation of optical flow. In *ICCV*, pages 231–236, 1993. 2, 4

[3] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *CVPR*, pages 221–230, 2017. 1

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, pages 9912–9924, 2020. 3

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 7

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 1

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 6

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021. 3, 7

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 5, 6, 7

[10] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, pages 2758–2766, 2015. 1, 2, 4

[11] Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 7

[12] David F Fouhey, Wei-cheng Kuo, Alexei A Efros, and Jitendra Malik. From lifestyle vlogs to everyday interactions. In *CVPR*, pages 4991–5000, 2018. 7

[13] Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. Watching the world go by: Representation learning from unlabeled videos. *arXiv preprint arXiv:2003.07990*, 2020. 7

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, pages 21271–21284, 2020. 3

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 1, 3, 4, 5, 7

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5, 7, 8

[17] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 1981. 1

[18] Yingdong Hu, Renhao Wang, Kaifeng Zhang, and Yang Gao. Semantic-aware fine-grained correspondence. In *ECCV*, pages 97–115, 2022. 2, 7, 8

[19] Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. In *NeurIPS*, pages 19545–19560, 2020. 1, 2, 3, 6, 7, 8

[20] Sangryul Jeon, Dongbo Min, Seungryong Kim, and Kwanghoon Sohn. Mining better samples for contrastive learning of temporal correspondence. In *CVPR*, pages 1034–1044, 2021. 7, 8

[21] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *ICCV*, pages 3192–3199, 2013. 6, 8

[22] Rico Jonschkowski, Austin Stone, Jonathan T Barron, Ariel Gordon, Kurt Konolige, and Anelia Angelova. What matters in unsupervised optical flow. In *ECCV*, pages 557–572, 2020. 2

[23] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 6

[24] Zihang Lai, Erika Lu, and Weidi Xie. Mast: A memory-augmented self-supervised tracker. In *CVPR*, pages 6479–6488, 2020. 1, 2, 4, 5, 6, 7

[25] Zihang Lai and Weidi Xie. Self-supervised learning for video correspondence flow. In *BMVC*, 2019. 2, 4, 5, 7

[26] Liulei Li, Tianfei Zhou, Wenguan Wang, Lu Yang, Jianwu Li, and Yi Yang. Locality-aware inter-and intra-video reconstruction for self-supervised correspondence learning. In *CVPR*, pages 8719–8730, 2022. 5, 7

[27] Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. In *NeurIPS*, pages 318–328, 2019. 1, 2, 7, 8

[28] 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*, pages 740–755, 2014. 7

[29] Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In *CVPR*, pages 6489–6498, 2020. 2

[30] Pengpeng Liu, Irwin King, Michael R Lyu, and Jia Xu. Ddflow: Learning optical flow with unlabeled data distillation. In *AAAI*, pages 8770–8777, 2019. 2, 4

[31] Xiankai Lu, Wenguan Wang, Jianbing Shen, Yu-Wing Tai, David J Crandall, and Steven CH Hoi. Learning video object segmentation from unlabeled videos. In *CVPR*, pages 8960–8970, 2020. 7

[32] K-K Maninis, Sergi Caelles, Yuhua Chen, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. Video object segmentation without temporal information. *IEEE transactions on PAMI*, 41(6):1515–1530, 2018. 7

[33] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, pages 4040–4048, 2016. 1

[34] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *AAAI*, 2018. 2, 4

[35] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Self-supervised video object segmentation by motion-aware mask propagation. In *ICME*, pages 1–6, 2022. 5, 7

[36] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *ECCV*, pages 300–317, 2018. 7

[37] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, pages 9226–9235, 2019. 1

[38] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[39] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 1, 6, 7, 8

[40] Jeany Son. Contrastive learning for space-time correspondence via self-cycle consistency. In *CVPR*, pages 14679–14688, 2022. 2, 7, 8

[41] Jie Song, Limin Wang, Luc Van Gool, and Otmar Hilliges. Thin-slicing network: A deep structured model for pose estimation in videos. In *CVPR*, pages 4220–4229, 2017. 8

[42] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419, 2020. 2, 4

[43] Jack Valmadre, Luca Bertinetto, Joao F Henriques, Ran Tao, Andrea Vedaldi, Arnold WM Smeulders, Philip HS Torr, and Efstratios Gavves. Long-term tracking in the wild: A benchmark. In *ECCV*, pages 670–685, 2018. 7

[44] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *CVPR*, pages 9481–9490, 2019. 7

[45] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. *arXiv preprint arXiv:1706.09364*, 2017. 7

[46] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *ECCV*, pages 391–408, 2018. 2, 7

[47] Ning Wang, Wengang Zhou, and Houqiang Li. Contrastive transformation for self-supervised correspondence learning. In *AAAI*, pages 10174–10182, 2020. 2, 3, 7, 8

[48] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *CVPR*, pages 2566–2576, 2019. 1, 2, 4, 7, 8

[49] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, pages 3024–3033, 2021. 3

[50] Zhongdao Wang, Hengshuang Zhao, Ya-Li Li, Shengjin Wang, Philip Torr, and Luca Bertinetto. Do different tracking tasks require different appearance models? In *NeurIPS*, pages 726–738, 2021. 1, 3

[51] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018. 3

[52] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *ICCV*, pages 8392–8401, 2021. 1, 3, 4, 5

[53] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose flow: Efficient online pose tracking. In *BMVC*, 2018. 1

[54] Jiarui Xu and Xiaolong Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. In *ICCV*, pages 10075–10085, 2021. 2, 3, 6, 7, 8

[55] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 1, 5, 6, 7

[56] Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. In *CVPR*, pages 3987–3996, 2021. 3

[57] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE transactions on PAMI*, 35(12):2878–2890, 2012. 8

[58] Jason J Yu, Adam W Harley, and Konstantinos G Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *ECCV*, pages 3–10, 2016. 2

[59] Zixu Zhao, Yueming Jin, and Pheng-Ann Heng. Modelling neighbor relation in joint space-time graph for video correspondence learning. In *ICCV*, pages 9960–9969, 2021. 2, 7, 8

[60] Qixian Zhou, Xiaodan Liang, Ke Gong, and Liang Lin. Adaptive temporal encoding network for video instance-level human parsing. In *ACM MM*, pages 1527–1535, 2018. 6, 8