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 different 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 downsampled 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-
Contrastive Learning

Encoder

(ii) Second Step

(i) First Step

the global correlation distillation used for retaining the reconstruction and local correlation distillation. In the end, poral feature learning is improved with our pyramid framing in the context of contrastive learning. Then, the tem-

3. Method

the temporal pattern between video frames.

dence [5], the contrastive model still struggles to figure out though showing superior performance for video correspon-
dances. 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.

3.1. 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 \( K = \{ k^+, k^-_1, k^-_2, \ldots, k^-_K \} \) which consists of one positive key \( k^+ \in \mathbb{R}^d \) and \( K \) negative keys \( K^- = \{ k^-_j \} \), are encoded by the encoder \( \phi \) 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 \( K^- \). Thus, a contrastive loss is presented in InfoNCE [38] with a softmax formulation:

\[
\mathcal{L}_{nce} = - \log \frac{\exp (q^T \cdot k^+/\tau_c)}{\exp (q^T \cdot k^+/\tau_c) + \sum_{i=1}^{K} \exp (q^T \cdot k^-_i/\tau_c)},
\]

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.
where the similarity is measured via dot product, and $\tau_c$ 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 $L_{nce}$ 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 $\phi$. We denote these embeddings as $F_{t, l}^i, F_{r, l}^i \in \mathbb{R}^{H_l \times W_l \times d_l}$, where $l \in \{0, 1, \ldots, N\}$ is the index of each pyramid level and the bigger number represents the coarser pyramid level. The $s_l^i$ denotes the stride up to each pyramid level and $H_l^i = H/s_l^i, W_l^i = W/s_l^i$. 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 $\hat{I}_t^l, \hat{I}_r^l \in \mathbb{R}^{H_l \times W_l \times 3}$. 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:

$$c^l(i, j) = \frac{\exp \left( \frac{F_{t, l}^i \cdot F_{r, l}^j}{\tau_c} \right)}{\sum_n \exp \left( \frac{F_{t, l}^i \cdot F_{r, l}^n}{\tau_c} \right)} , i \in \{1, \ldots, H_l^i, W_l^i\} , j, n \in \mathcal{N}(i) ,$$

where $\mathcal{N}(i)$ is the index set in the reference frame with a limited range of $H_l^i$ 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 $c^l \in \mathbb{R}^{H_l^i \times W_l^i \times (2N_l^i+1)^2}$:

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

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

$$L_{rec} = \left\| \hat{I}_t^l - \mathcal{T}_t^l \right\|_1 .$$

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 $\{f^{t, l}_{t, i}\}_{l=1}^{N-1}, \{f^{r, l}_{t, i}\}_{l=1}^{N-1}$, 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:

$$L_{rec}^p = \sum_{l} \left\| \hat{I}_t^l - \mathcal{T}_t^l \right\|_1 .$$

#### 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 $L_{lc}$ 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
more attention to when applying local correlation distillation. We calculate the entropy for each query \( i \):

\[
H(i) = \sum_j - \log c^{N-1}(i, j), \quad i \in \{1, \ldots, h_{N-1}w_{N-1}\}, \quad j \in \mathcal{N}(i),
\]

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:

\[
\mathcal{L}_i^c = \sum m(i) \cdot \| c^{N-1}(i) - \hat{c}^{N-1}(i) \|^2_{l_2}.
\]

Eventually, our training loss of temporal feature learning is defined as \( \mathcal{L}_t = \mathcal{L}_{rec} + \alpha \mathcal{L}_i^c \).

### 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 \( \phi \) 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 \( \phi \) trained in the first step as the teacher denoted as \( \phi_t \). 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.,

\[
a^N(i, j) = \frac{\exp(F_t^N(i) \cdot F_t^N(j)/\tau)}{\sum_n \exp(F_t^N(i) \cdot F_t^N(n)/\tau)}, \quad i, j, n \in \{1, \ldots, h_{N}w_{N}\},
\]

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

\[
\mathcal{L}_{gc} = \| a^N - \hat{a}^N \|^2_{l_2}.
\]

(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 \( \mathcal{L} = \mathcal{L}_{nce} + \gamma \mathcal{L}_t \).

(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 \( \phi \) 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 \( \text{res}_4 \) 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 \( 256 \times 256 \), 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 $H/8 \times W/8$ (stride=8) or $H/4 \times W/4$ (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 DAVIS-2017 validation set [39]. We use the mean of region similarity $J_m$, mean of contour accuracy $F_m$ and their average $J \& F_m$ as the evaluation metrics. We set the stride until layer res$_3$ 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, $L_{ge}$ 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 an 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-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 $L_{gc}$ can already reach 69.4%, and the $L_{gc}$ 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 $L_{nce}$ 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 $L_{nce}$ is misled by two patches at different locations which have a similar appearance, while the model trained with $L_t$ 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 $L_t$ fails to get accurate correspondence when facing severe temporal discontinuity, e.g., occlusions, appearance changes, deformations, while the model trained with $L_{nce}$ is able to correct the mistakes by tracking the points based on learned discriminative and

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>64.9</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>65.6</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>68.4</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>69.3</td>
</tr>
</tbody>
</table>

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].

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Dataset</th>
<th>Loss Function</th>
<th>DAVIS $J &amp; F_m$ ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial only</td>
<td>I</td>
<td>$L_{nce}$</td>
<td>66.5</td>
</tr>
<tr>
<td>Temporal only</td>
<td>YTV</td>
<td>$L_t$</td>
<td>67.5</td>
</tr>
<tr>
<td>Spatiotemporal</td>
<td>I+YTV</td>
<td>$L_{nce} \rightarrow L_t$</td>
<td>68.6</td>
</tr>
<tr>
<td>Spatial-then-Temporal</td>
<td>I→YTV</td>
<td>$L_{nce} \rightarrow L_t + EWC$</td>
<td>69.4</td>
</tr>
<tr>
<td>Temporal-then-Spatial</td>
<td>YTV→I</td>
<td>$L_t \rightarrow L_{nce}$</td>
<td>71.3</td>
</tr>
</tbody>
</table>


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].
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].

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

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 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 $320 \times 320$. 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


