
Self-Supervised Spatial Temporal Feature Learning for Video Correspondence

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 This paper proposes to learn reliable dense correspondence from videos in a
2 self-supervised manner. Our learning process integrates two highly related tasks:
3 tracking large image regions and establishing fine-grained pixel-level associa-
4 tions between consecutive video frames. We exploit the synergy between both
5 tasks through a shared inter-frame affinity matrix, which simultaneously mod-
6 els transitions between video frames at both the region- and pixel-levels. While
7 region-level localization helps reduce ambiguities in fine-grained matching by
8 narrowing down search regions; fine-grained matching provides bottom-up features
9 to facilitate region-level localization. Our method outperforms the state-of-the-art
10 self-supervised methods on a variety of visual correspondence tasks, including
11 video-object and part-segmentation propagation, keypoint tracking, and object
12 tracking. Our self-supervised method even surpasses the fully-supervised affinity
13 feature representation obtained from a ResNet-18 pre-trained on the ImageNet.

14 1 Introduction

15 Learning representations for video correspondence is a fundamental problem in computer vision,
16 which is closely related to different video applications, including optical flow estimation [4][6], video
17 object segmentation [1][10], keypoint tracking [15], etc. However, supervising such a representation
18 requires a large number of dense annotations, which is unaffordable. Thus most approaches acquire
19 supervision from simulations or limited annotations, which result in poor generalization in different
20 downstream tasks. Recently, self-supervised feature learning is gaining significant momentum, and
21 several pretext tasks are designed for space-time visual correspondence using abundant unlabeled
22 videos.

23 The key to this task lies in two different perspectives. The first one is **temporal feature learning**,
24 which aims to learn the fine-grained correspondence of pixel/object between frames. With the nature
25 of temporal coherence in the video, the temporal feature learning can be formed as a reconstruction
26 task, where the query pixel in the target frame can be reconstructed by leveraging the information of
27 adjacent reference frames within a local region. Then a reconstruction loss is applied to minimize
28 the photometric error between the raw frame and its reconstruction. However, in real videos, the
29 temporal discontinuity occurs frequently due to the occlusions, dramatic appearance changes, and
30 deformations, especially for pixels in each frame with severe down-sampling. In such scenarios,
31 the frame reconstruction loss apparently becomes invalid, which result in inferior performance.
32 MAST [8] proposes to apply frame reconstruction with a higher feature resolution by decreasing
33 the stride of the backbone, which requires a larger memory and computation cost. Besides, the
34 frame reconstruction is achieved under the assumption of temporal continuity, thus fail to learn
35 representation that is invariant to appearance change in time. Another way to exploit the free temporal
36 supervision is by exploiting temporal cycle-consistency. [7][12] track object forward and backward

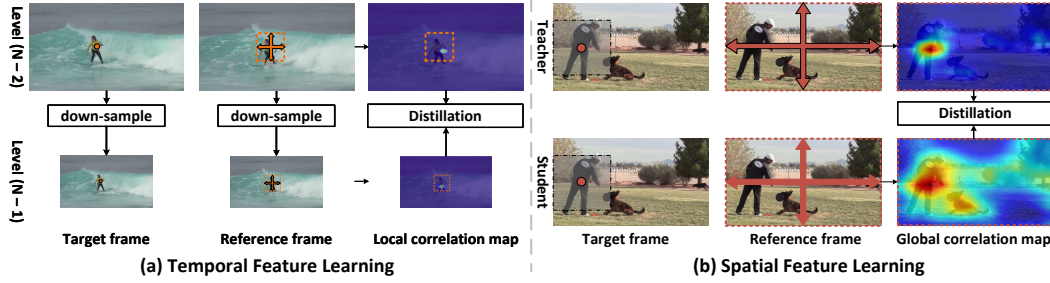


Figure 1: Visualization of our key idea. Given a query point randomly sampled in target frame, we visualize the result of computing the local correlation and global correlation map *w.r.t.* reference frame. The dashed line in red represents the range of computing correlation map *w.r.t.* query point. The reference frame is randomly sampled in the memory bank of inference strategy [11].

with the objective of maximizing the temporal cycle correspondence consistency with reconstruction and contrastive loss. However, compared to the correspondence which is realized at object-level, the frame reconstruction is conducted on raw image space, which provides more accurate supervision for learning fine-grained correspondence.

The second one is **spatial feature learning**, which pays more attention to learning the object appearance that is invariant to viewpoint and deformation changes. [11] adopts a novel intra-inter consistency loss to learn inter-video discriminative feature while [16] learns the space-time correspondence through a frame-wise contrastive loss. However, both methods are trained on video datasets and try to realize the spatial and temporal feature learning in a unified framework, which is sub-optimal for each of them. Recently, as mentioned in [13], the contrastive model [5][14] pre-trained on image data shows impressive performance for video correspondence. This motivates us to design a framework that learns the spatial and temporal feature independently with image and video data.

In this paper, we decouple video correspondence learning into two separate processes, including spatial and temporal feature learning. To achieve this, we first train the model in a contrastive learning paradigm on ImageNet [3], which gives the model the ability to capture object appearance. Then, instead of training with a large video dataset, *i.e.*, Kinetics [2] with 300k videos, we perform the temporal feature learning on YouTube-VOS [17], which consists of 3.5k videos. However, apart from the severe information lost and temporal discontinuity due to large spatial down-sampling on frames, directly fine-tuning the old model with only new data will leads to a well-known phenomenon of catastrophic forgetting [9]. To address the first problem, we propose a novel pyramid learning framework. The frame reconstruction is applied at different levels of network to better exploit the free temporal supervision. As you can see in Figure 1 (a), the pixels of target and reference frame with higher resolution have a lower chance of occurring temporal discontinuity, which provides a more accurate local correlation map. Thus we design a new loss named **local correlation distillation loss** that supports explicitly learning of the correlation map in the region with high uncertainty, which is achieved by taking the finest local correlation map as pseudo labels. At the same time, as shown in Figure 1 (b), we freeze the model pre-trained on ImageNet as teacher. Then a **global correlation distillation loss** is proposed to keep the student the ability of capturing object-level correspondence which is closely related to object appearance modeling.

To sum up, our main contributions include: (a) We propose a decoupled learning paradigm for self-supervised video correspondence, including spatial and temporal feature learning. (b) We propose a pyramid learning framework with local correlation distillation to enable the model to estimate fine-grained correspondence. (c) We introduce a global correlation distillation loss to keep the ability of capturing object appearance when training on video dataset. (d) Last but not least, we verify the proposed approach in a series of correspondence-related tasks including video object segmentation, human parts propagation and pose tracking. Our approach consistently outperforms previous state-of-the-art self-supervised methods and is even comparable with some task-specific fully-supervised algorithms.

76 2 Related Work

77 **Self-supervised learning for video correspondence.** Recent approaches focus on learning corre-
 78 spondence from unlabeled videos in a self-supervised manner. The task requires the model to have
 79 the ability to capture object appearance and estimate the fine-grained correspondence between frames
 80 at the same time, which has proceeded along two different dimensions: reconstruction-based methods
 81 and cycle-consistency-based methods. In the first type, a query point is reconstructed from adjacent
 82 frames while the latter performs forward-backward tracking with the objective of minimize the cycle
 83 inconsistency. Through getting promising results, most methods address the problem considering only
 84 one perspective. VFS learns the spatial temporal representation through a frame-wise contrastive loss
 85 while DUL and Contra try to realize the spatial and temporal feature learning in a unified framework
 86 by exploiting the inter-video constraint, which may result in sub-optimal performance. In this paper,
 87 we learn a better representation by decoupling the video correspondence learning to two separate
 88 processes including spatial and temporal feature learning.

89 **Self-supervised spatial feature learning.** Self-supervised spatial feature learning aims to learn
 90 discriminative feature of object appearance with unlabeled data, which recently has get promising
 91 result with contrastive learning. In an early work [44], the constrastive learning is formulated
 92 as instance discrimination task, which requires the model to return low values for similar pairs
 93 and high values for dissimilar pairs. Recently, the performance is further improved by creating a
 94 dynamic memory-bank, introducing online clustering and avoiding the use of negative pairs. Further
 95 more, DensCL, InsLoc and DetCo propose various pre-text tasks to adapt the contrastive learning
 96 to tasks which require dense prediction. Even through showing superior performance for video
 97 correspondence learning [11], the contrastive model pre-trained on image data still struggles to
 98 modeling the fine-grained correspondence between video frames.

99 **Self-supervised temporal feature learning.** Compared to spatial feature learning, temporal feature
 100 learning focus on learning the motion information of video, which is closely related to optical flow
 101 and motion estimation. Most methods [12][11] directly regress the groundtruth optical flow produced
 102 by synthetic datasets thus suffering from severe domain shift. To deal with the problem, [1][3] try to
 103 learn the dense correspondence on real video without any label by minimizing the photometric error
 104 between the raw frame and its reconstruction. However, the video frames usually contain temporal
 105 discontinuity including dramatic appearance change and occlusions, which seriously degrades the
 106 capability of the method. [12][14] alleviate the problem by utilizing the optical flow predictions from
 107 teacher model to guide the learning of student model in the region with occlusions. In this paper,
 108 we address the issue from two perspectives: (1) Learning a better representation that is invariant for
 109 appearance change. (2) Supervising the correlation by taking the finest corelation as pseudo labels.

110 3 Approach

111 The basic idea of our method is to decouple video correspondence learning into two separate
 112 processes including spatial feature learning and temporal feature learning. We first train our model
 113 using contrastive loss with image data to learn the object appearance that is invariant to viewpoint.
 114 Then, we perform the temporal feature learning on a small video dataset to learn the fine-grained
 115 correspondence between frames. To deal with the problems including temporal discontinuity and
 116 catastrophic forgetting when training on video data, a pyramid learning framework is proposed with
 117 local and global correlation distillation loss.

118 3.1 Spatial Feature Learning

119 The spatial feature mainly describes appearance of objects involved in a image. Spatial feature
 120 learning is analogous to that of instance discrimination, and thus easily benefits from the recent
 121 advancements brought by contrastive learning. We firstly briefly reviewing instance discrimination
 122 objective in contrastive learning. Given an encoded query $\mathbf{q} \in \mathbb{R}^d$ and a set of encoded key vectors
 123 $\mathcal{K} = \{\mathbf{k}^+, \mathbf{k}_1^-, \mathbf{k}_2^-, \dots, \mathbf{k}_K^-\}$ which consists of one positive key $\mathbf{k}^+ \in \mathbb{R}^d$ and K negative keys
 124 $\mathcal{K}^- = \{\mathbf{k}_j^-\}$, where d denotes the embedding dimension. The query and its positive key are
 125 generated from same instance with two different augmentations, while the negative keys ref to other
 126 instances. The objective of instance discrimination is to maximize the similarity between the query \mathbf{q}
 127 and the positive key \mathbf{k}^+ while remains query distinct to all negative keys \mathcal{K}^- . Thus, a contrastive

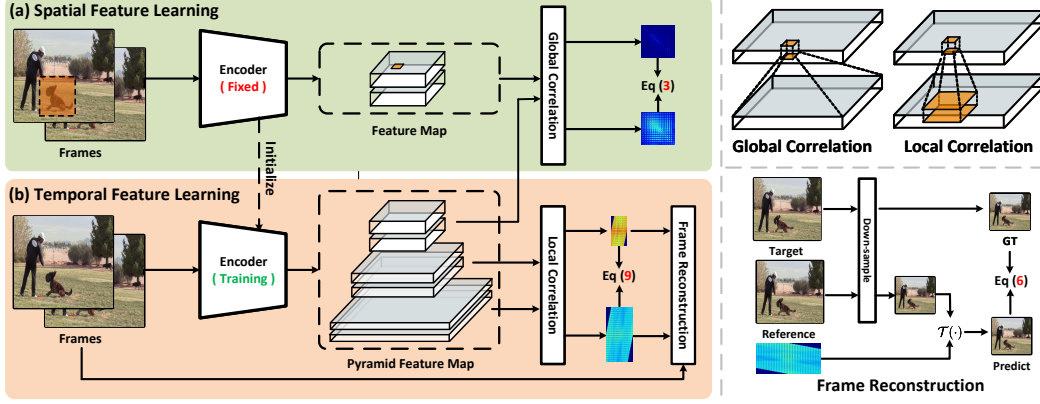


Figure 2: An overview of our spatial temporal feature learning framework. Our method decouples video correspondence learning into two separate processes including spatial feature learning and temporal feature learning. Specifically, the **spatial feature learning** first exploits the contrastive loss which is analogous to that of instance discrimination to learn the object appearance with image data. Then we perform the self-supervised training with video data in the next step. To maintain the ability to capture object appearance, we fix the pre-trained network as teacher and a global correlation distillation is devised. For **temporal feature learning**, we propose a pyramid learning framework where the frame reconstruction is devised at each levels of network. As the same time, we introduce a novel loss named local correlation distillation loss that supports explicitly learning of the correlation map at the region with high uncertainty, which is achieved by taking the finest local correlation map as pseudo labels.

loss is presented in InfoNCE with a softmax formulation:

$$\mathcal{L}_{\text{ncc}} = -\log \frac{\exp(\mathbf{q}^T \mathbf{k}^+ / \tau_c)}{\exp(\mathbf{q}^T \mathbf{k}^+ / \tau_c) + \sum_{i=1}^K \exp(\mathbf{q}^T \mathbf{k}_i^- / \tau_c)}, \quad (1)$$

where the similarity is measured via dot product, and τ_c is the temperature hyper-parameter. MoCo [14] builds a dynamic memory bank to maintain a large number of negative samples with a moving-averaged encoder. DetCo [15] further improves the contrastive loss \mathcal{L}_{ncc} by introducing a global and local contrastive learning to enhance local representation for dense prediction. In this paper, we adopt the same framework as MoCo [14] and DetCo [15] to learn a appearance model for most of our experiments.

Global correlation distillation. After main training with contrastive loss, we get a encoder ϕ . Then we continuously train it on video data to learn the fine-grained correspondence (See section 3.2). However, directly fine-tuning the old model with only new data will lead to a well-known phenomenon of catastrophic forgetting, which degrades the performance. Thus we introduce a global correlation distillation loss in order to maintain the ability to capture object appearance. More specifically, we first fix the feature encoder ϕ as teacher denoted as ϕ_t . Given a pair of video frames consisting of target and reference frame $I_t, I_r \in \mathbb{R}^{H \times W \times 3}$, the ϕ maps them to a pair of feature embeddings $F_t^l, F_r^l \in \mathbb{R}^{h^l \times w^l \times d^l}$, where $l \in \{0, 1, \dots, N\}$ is the index of each pyramid level and the smaller number represents the coarser level. Here l is set to N . For each query point $F_t^l(i)$ and key point $F_r^l(j)$, we have global correlation $a_{i,j}$ (See Figure 2) using a softmax over similarities *w.r.t.* all keys in reference frame, *i.e.*:

$$a_{i,j} = \frac{\exp(F_t^l(i) \cdot F_r^l(j) / \tau)}{\sum_n \exp(F_t^l(i) \cdot F_r^l(n) / \tau)}, i, j, n \in \{1, \dots, h^l w^l\}, \quad (2)$$

Where ‘ \cdot ’ stands for the dot product. Each point in F_t^l and F_r^l covers a relatively large region **since the output stride is set to 32 in our feature encoder**. Thus we can form the correlation as object-level correspondence which is closely related to object appearance. We generate the pseudo labels of global correlation distillation by computing the global correlation $a_{i,j}^t$ for each query with teacher ϕ_t . The global correlation distillation loss is defined to minimize the mean squared error between a and a^t :

$$\mathcal{L}_{\text{gc}} = \|a - a^t\|_2^2, \quad (3)$$

3.2 Temporal Feature Learning

We then perform temporal feature learning right after spatial feature learning. Temporal feature learning aims to learn the fine-grained correspondence between video frames. Recently, a few studies [40, 84] introduce a reconstruction-based correspondence learning scheme, where each query pixel in the target frame can be reconstructed by leveraging the information of adjacent reference frames within a local region. More specifically, the target and reference frame I_t, I_r are projected into a fine-grained pixel embedding space. We denoted these embedding as $F_t, F_r \in \mathbb{R}^{hw \times d}$. For each query pixel i in I_t , we can calculate the local correlation $c_{i,j}$ with reference frame within a local region with a softmax formulation:

$$c_{i,j} = \frac{\exp(F_t(i) \cdot F_r(j)/\tau)}{\sum_n \exp(F_t(i) \cdot F_r(n)/\tau)}, i \in \{1, \dots, hw\}, j, n \in \mathcal{N}(i), \quad (4)$$

Where $\mathcal{N}(i)$ is the index set with a limited range of r pixels for pixel i (see Figure 2). Then each query pixel i in target frame can be reconstructed by a weighted-sum of pixels in $\mathcal{N}(i)$, according to the local correlation map $c \in \mathbb{R}^{hw \times (r)^2}$:

$$\hat{I}_t(i) = \sum_{j \in \mathcal{N}(i)} c_{i,j} I_r(j), \quad (5)$$

We regard the above process as a transformation function for all query pixels and denotes it as: $\hat{I}_t = \mathcal{T}(c, I_t)$. Then the reconstruction loss is defined as L1 distance between \hat{I}_t and I_t :

$$\mathcal{L}_{\text{rec}} = \|I_t - \hat{I}_t\|_1, \quad (6)$$

However, the Eq 5 should only be applied when the feature embedding has same size with video frame. Thus the stride of ϕ must be set to 1, which introduces large memory and computation cost. One possible solution is to apply down-sampling on target frame. MAST propose a image feature alignment module which sampling the pixel at center of strided convolution kernels. However, down-sampling with large rate would not only cause severe information lost of supervision but also result in more pixel occlusions between video frames, which obviously degrades the representation of temporal feature learning. To alleviate the problem, we design a pyramid learning framework consisting of pyramid frame reconstruction and local correlation distillation with entropy-based selection.

Pyramid frame reconstruction. As you can see in Figure 2, we obtain a pair of feature pyramids $\{F_t^l\}_{l=1}^{N-1}, \{F_r^l\}_{l=1}^{N-1}$. Then we get the pyramid local correlation map $\{c^l\}_{l=1}^{N-1}$ at each level by utilizing Eq 4 with different local range r^l . As the same time, we adopt a same down-sampling method as [1] to obtain a pair of frame pyramids $\{I_t^l\}_{l=1}^{N-1}, \{I_r^l\}_{l=1}^{N-1}$, which has same shape with the feature pyramids at each level. Given the c^l, I_t^l and I_r^l , we apply the pyramid reconstruction loss at each level:

$$\mathcal{L}_{\text{rec}}^p = \sum_l \|I_t^l - \mathcal{T}(c^l, I_r^l)\|_1, \quad (7)$$

By doing this, we are able to exploit more free temporal supervision and get better temporal representation at intermediate level.

Local correlation distillation. The bottom level of the frame pyramid contains more rich information and suffer less occlusions for temporal feature learning due to relatively small down-sampling rate, which may result in more accurate local correlation map. Inspired by it, we design a novel local correlation distillation loss which explicitly make constraint on the final local correlation map $c^{N-1} \in \mathbb{R}^{h^{N-1} w^{N-1} \times (r^{N-1})^2}$. We first compute the local correlation map c^{N-2} at level $N-2$ and then apply correlation down-sampling [16] to get pseudo labels c^t with the same size as c^{N-1} . Then the local correlation distillation loss \mathcal{L}_{lc} is adopt to minimize the mean squared error between c^{N-1} and c^t .

Entropy-based selection. The correlation of each query *w.r.t* reference frame indicates more uncertainty when having smooth distribution, which should be paid more attention when applying distillation. Thus we calculate the entropy for each query i :

$$\mathcal{H}(i) = \sum_j -\log c_{i,j}^{N-1}, \quad (8)$$

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

$$\mathcal{L}_{lc}^e = \sum_i m_i \left\| c_{i,:}^{N-1} - c_{i,:}^t \right\|_2^2, \quad (9)$$

Eventually, our training loss of temporal feature learning is defined as: $\mathcal{L}_t = \mathcal{L}_{rec}^p + \alpha \mathcal{L}_{lc}^e$. The final loss of training on video data is a weighted sum of \mathcal{L}_t and a regularization term \mathcal{L}_{gc} introduced in Section 3.1:

$$\mathcal{L} = \mathcal{L}_t + \beta \mathcal{L}_{gc}, \quad (10)$$

4 Experiments

We verify the merit of our method in a series of correspondence-related tasks, including semi-supervised video object segmentation, pose keypoints tracking, human parts segmentation propagation. In this section we will first introduce our experiments settings including details of implementation and evaluation. Then detailed ablation studies are performed to explain how each component of our method works. Last but not least, we finally report the performance comparison with state-of-the-art methods to further verify the effectiveness of our method.

4.1 Implementation Details

Architectures. We exploit the encoder ϕ with both ResNet-18 and ResNet-50 [27, 37] for self-supervised training. Following prior works [12], we reduce the stride of convolutional layers in ϕ to increase the spatial resolution of feature maps on layer *res4* by a factor of 4 or 8 (*i.e.*, downsampling rate 8 or 4). Further more we set the stride of *res5* layer to 4 in order to increase the receptive field for spatial feature learning.

Training. We first train our model using contrastive loss for 200 epochs on ImageNet following most hyper-parameters settings of [76]. Then we perform temporal feature learning on YouTube-VOS train set [16] which consists of 3.5k videos. In this stage, the video frame is resized into 256×256 and channel-wise dropout in Lab color space [39] is adopted as the information bottleneck. We train the encoder for 100k iterations with mini-batch of 128 using Adam as our optimizer. The initial learning rate is set to $1e-4$ with a cosine (half-period) learning rate schedule. The frame reconstruction is applied on both *res3* and *res4* layer while we realize global correlation distillation on *res5* layer.

Evaluation. We follow the same evaluation protocol and downstream tasks in [33, 71]. We directly utilize 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 *res4* layer with a spatial resolution of $\frac{H}{8} \times \frac{W}{8}$ or $\frac{H}{4} \times \frac{W}{4}$. To propagate the semantic labels from the initial ground-truth annotation, the recurrent inference strategy is applied following recent works [12]. More specifically, the semantical label of first frame as well as previous predictions are propagated to the current frame with the help of affinity between video frames. We evaluate our method over three downstream tasks including semi-supervised video object segmentation in DAVIS-2017 [8], human part propagation in VIP [10] and human pose tracking in JHMDB [9].

4.2 Ablation Study

The ablation study is performed with semi-supervised video object segmentation on DAVIS-2017 validation set. Following the official protocol [68], 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 conduct a series of experiments to prove the effectiveness of each component. The stride of encoder is all set to 8 for training and evaluation.

Temporal feature learning. We first examine how each design in temporal feature learning impacts the overall performance, which are shown in Table 5 (a) 1 \sim 4. To have a clear look, we train the model from scratch on YouTube-VOS when examining the efficacy of our components for temporal feature learning. The baseline is to apply frame reconstruction \mathcal{L}_{rec} . The p , \mathcal{L}_{lc} and e represents pyramid frame reconstruction, local correlation distillation without and with entropy-based selection. From the table, we can see leveraging more supervision of reconstruction at each level leads to an improvement in the range of 0.8%. With the guidance of a more fine-grained local correlation

#	\mathcal{L}_{nce}	\mathcal{L}_{gc}	\mathcal{L}_{rec}	p	\mathcal{L}_{lc}	e	Dataset	Arch	$\mathcal{J}\&\mathcal{F}_m \uparrow$
1			✓				YTV	Res18	64.6
2			✓	✓			YTV	Res18	65.4
3			✓	✓	✓		YTV	Res18	68.1
4			✓	✓	✓	✓	YTV	Res18	69.0
5	✓		✓	✓	✓	✓	I + YTV	Res18	69.3
6	✓	✓	✓	✓	✓	✓	I + YTV	Res18	70.5

(a) Ablation study of each component.

Method	Dataset	Arch	$\mathcal{J}\&\mathcal{F}_m \uparrow$
\mathcal{L}_{nce}	I	Res50	66.5
w \mathcal{L}_t	I + YTV	Res50	69.6
w $\mathcal{L}_t + \text{LwF}$	I + YTV	Res50	69.9
w $\mathcal{L}_t + \mathcal{L}_{gc}$	I + YTV	Res50	71.3

(b) Ablation study of \mathcal{L}_{gc} .

Table 5: Ablation study for each component in spatial and temporal feature learning. The "p" and "e" in (a) correspond to pyramid frame reconstruction and entropy-based selection. Models in (b) are all pre-trained on ImageNet with contrastive loss and models with "w" are subsequently trained on YouTube-VOS using different methods. I: ImageNet [11]. YTV: YouTube-VOS [12].

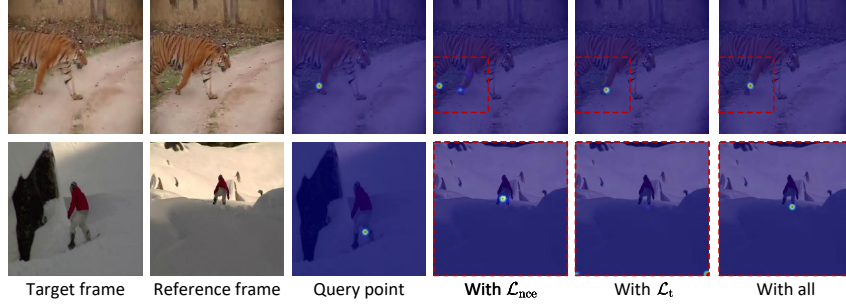


Figure 3: Visualization of the ablation study. Given a query point randomly sampled in target frame, we visualize the result of computing the local correlation and global correlation map *w.r.t.* reference frame. The dashed line in red represents the range of computing correlation map *w.r.t.* query point. The reference frame is randomly sampled in the memory bank of inference strategy [11].

map, \mathcal{L}_{lc} boosts up the accuracy from 65.4% to 68.1%. Moreover, enforcing the local correlation distillation to focus on the region with higher entropy leads to a performance gain in the range of 1%. By fusing the above components, the performance finally reaches 69.0%.

Spatial feature learning. We investigate the effect of training with each component in spatial feature learning. The results are shown in Table 3 (a) 5 ~ 6, with the help of the pre-training on ImageNet using contrastive loss, the performance of our method reaches 69.3%. Moreover, the global correlation distillation loss \mathcal{L}_{gc} boosts up the performance from 69.3% to 70.5% by keeping the ability of the model to capture object-level correspondence which is closely related to object appearance modeling.

Further exploitation of \mathcal{L}_{gc} . Directly fine-tuning the model pre-trained with contrastive loss \mathcal{L}_{nce} will lead to a well-known phenomenon of catastrophic forgetting, which is closely related to continual learning. To further verify the effectiveness of \mathcal{L}_{gc} , we exploit a general continual model LwF [23] based on knowledge distillation apart from directly fine-tuning on video dataset. We modify the framework of LwF (see Appendix) to adapt to the paradigm of self-supervised learning and adopt ResNet-50 as our encoder which has a stronger ability to capture object appearance. The results are shown on Table 5 (b). All methods achieve better results attributed to the proposed temporal feature learning while our method using \mathcal{L}_{gc} gets the best performance.

Further analysis. We give a further analysis here based on the above experiments. On the one hand, temporal feature learning helps to learn the fine-grained correspondence related to motion estimation between frames, which is unable to accomplish by training an appearance model. As you can see in the first row of Figure 3, the appearance model trained with \mathcal{L}_{nce} is misled by two patches at different locations (*i.e.* two feats of the tiger) which has similar appearance, while the model trained with \mathcal{L}_t tends to learn a better temporal representation for fine-grained correspondence. However, in the second row of Figure 3, the model trained with \mathcal{L}_t fails to capture temporal correspondence with a local correlation when facing severe temporal discontinuity while the model trained with \mathcal{L}_{nce} is able to correct the mistakes by tracking the points based on the object appearance (see with \mathcal{L}_{nce} and with all).

Method	Sup.	Arch	Stride	Dataset		$\mathcal{J} \& \mathcal{F}_m \uparrow$	$\mathcal{J}_m \uparrow$	$\mathcal{F}_m \uparrow$
				Image	Video			
Supervised	✓	ResNet-18	8	ImageNet	-	62.9	60.6	65.2
MoCo		ResNet-18	8	ImageNet	-	60.8	58.6	63.1
SimSiam		ResNet-18	8	ImageNet	-	62.0	60.0	64.0
Colorization		ResNet-18	8	-	Kinetics	34.0	34.6	32.7
CorrFlow		ResNet-18	8	-	OxUvA	50.3	48.4	52.2
MuG		ResNet-18	8	-	OxUvA	54.3	52.6	56.1
ContrastCorr		ResNet-18	8	COCO	TrackingNet	63.0	60.5	65.5
UVC		ResNet-18	8	COCO	Kinetics	57.8	56.3	59.2
VFS		ResNet-18	8	-	Kinetics	66.7	64.0	69.4
CRW		ResNet-18	8	-	Kinetics	67.6	64.8	70.2
JSTG		ResNet-18	8	-	Kinetics	68.7	65.8	71.6
DUL		ResNet-18	8	-	YTV	69.3	67.1	71.6
MAST		ResNet-18	4	-	YTV	65.5	63.3	67.6
MAMP		ResNet-18	4	-	YTV	69.7	68.3	71.2
Ours		ResNet-18	8	-	YTV	69.0		
Ours		ResNet-18	8	ImageNet	YTV	70.5		
Ours		ResNet-18	4	-	YTV	71.3		
Ours		ResNet-18	4	ImageNet	YTV	73.2		
<hr/>								
Supervised	✓	ResNet-50	8	ImageNet	-	66.0	63.7	68.4
MoCo		ResNet-50	8	ImageNet	-	65.4	63.2	67.6
SimSiam		ResNet-50	8	ImageNet	-	66.3	64.5	68.2
TimeCycle		ResNet-50	8	-	VLOG	48.7	46.4	50.0
UVC		ResNet-50	8	COCO	Kinetics	57.8	56.3	59.2
VINCE		ResNet-50	8	-	Kinetics	65.6	63.4	67.8
SeCo		ResNet-50	8	-	Kinetics	60.6	60.4	62.8
VFS		ResNet-50	8	-	Kinetics	68.9		
Ours		ResNet-50	8	ImageNet	YTV	71.3		
SiamMask	✓	ResNet-50	-	I + C	YTV	56.4	54.3	58.5
OnAVOS	✓	ResNet-38	-	I + C + P	D	65.4	61.6	69.1
OSVOS-S	✓	VGG-16	-	I + P	D	68.0	64.7	71.3

Table 2: **Quantitative results for video object segmentation on validation set of DAVIS-2017.** We show results of state-of-the-art self-supervised methods and some supervised methods for comparison. "Dataset" represents the dataset(s) for pre-training, including: I:ImageNet (1.28m). C:COCO (30k). O:OxUvA (14h). T:TrackingNet (300h). K:Kinetics (800h). V:VLOG (344h). YTV:YouTube-VOS (5h). D:DAVIS 2017 (-). P:PASCAL-VOC (-). We report the data size for self-supervised methods (total number/duration of image/video dataset).

274 4.3 Comparison with State-of-the-art

275 **Results for video object segmentation.** We compare our method against previous self-supervised
276 methods in Table 2. For fair comparison, we report both results with setting the stride of encoder
277 to 4 and 8. The results are all reported with layer *res4* across all methods. Our method achieves
278 state-of-the-art performance using both ResNet-18 and ResNet-50. For ResNet-18, our method with
279 a stride of 8 achieves 70.5%, making a absolute performance improvement by 1.2% over all baselines
280 using same architecture. Benefiting from less information lost for temporal feature learning by setting
281 the stride of encoder to 4, the performance of method reaches 73.2%, leading a performance gain of
282 3.5% over MAMP, which consistently verify the idea of our methods. Besides, we found our method
283 trained with only temporal feature learning reaches 69.0%/71.2% with stride of 8/4, surpassing all
284 methods pre-trained on Kinetics and TrackingNet which have much bigger size than YouTube-VOS.
285 For ResNet-50, our method still outperforms all baseline methods by 2.4%. More remarkably, our
286 method even outperforms some task-specific fully-supervised algorithms [14][15].

287 Results for human part propagation.

288 Next, we evaluate our method on human
289 part tracking. Experiments are con-
290 ducted on the validation set of VIP [11],
291 which consists of 50 videos with 19
292 human semantic part classes, requiring
293 more precise matching than DAVIS. Fol-
294 lowing [45], we adopt mean intersection-
295 over-union (mIoU) as our evaluation met-

Methods	Sup.	VIP		JHMDDB	
		mIoU \uparrow	PCK@0.1 \uparrow	PCK@0.2 \uparrow	
ResNet-18	✓	31.9	53.8	74.6	
TimeCycle		28.9	57.3	78.1	
UVC		34.1	58.6	79.6	
CRW		38.6	59.3	80.3	
ContrastCorr		37.4	61.1	80.8	
VFS		39.9	60.5	79.5	
CLTC		37.8	60.5	82.3	
JSTG		40.2	61.4	85.3	
Ours		41.0	63.1	82.9	
<hr/>					
8 ATEN	✓	37.9	-	-	
Thin-Slicing Net	✓	-	68.7	92.1	

Table 3: **Quantitative results for human part propagation and human pose tracking.** We show results of state-of-the-

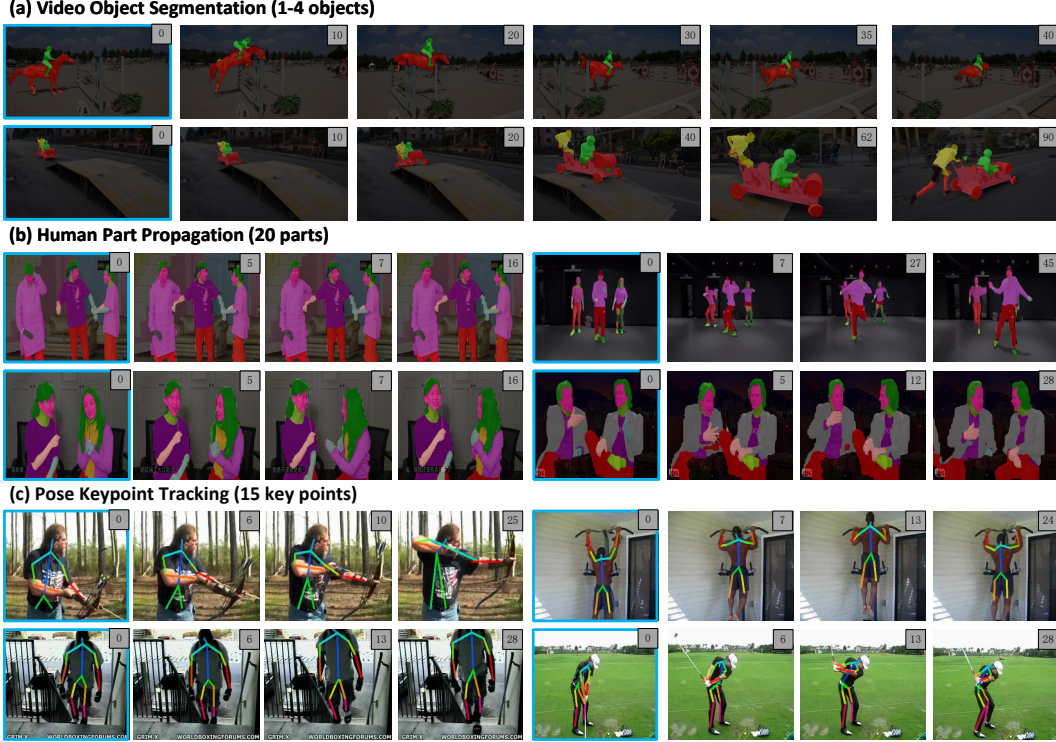


Figure 4: Qualitative results for label propagation. Given a first frame with different annotation highlighted with a blue outline, we propagate it to current frame without fine-tuning. (a) Video object segmentation on DAVIS-2017. (b) Human part propagation on VIP. (c) Pose keypoint tracking on JHMDB.

ric and resize the video frames to 560×560 . All models except Time-Cycle [12] are set to ResNet-18 with a stride of 8 for fair comparison. The results are shown in Table 3. Our method achieves state-of-the-art performance, surpassing all previous state-of-the-art by 0.8%. Notably, our model outperforms ATEN [18] which is specifically designed for the task using human annotations. Figure 4 (b) depicts some visualization results on several representative videos.

Results for human pose tracking. We then make performance comparison on the downstream task of human pose tracking. We conduct the experiments on the validation of JHMDB [35] which has 268 videos. The annotations consist of 15 body joints for each person. Probability of correct keypoint(PCK) [101] is utilized here to examine the accuracy between result and ground-truth with different thresholds. Follow the evaluation protocol of [24, 15], we resize the video frames to 320×320 . The results in Table 3 show a consistently performance gain over previous methods, which successfully demonstrates the transferability of our method to different downstream tasks. The visualization results in Figure 4 (c) show the robustness of our approach towards various challenges.

5 Conclusions

We then make performance comparison on the downstream task of human pose tracking. We conduct the experiments on the validation of JHMDB [35] which has 268 videos. The annotations consist of 15 body joints for each person. Probability of correct keypoint(PCK) [101] is utilized here to examine the accuracy between result and ground-truth with a threshold σ . The results in Table 3 show a consistently performance gain over previous methods, which successfully demonstrates the transferability of our method to different downstream tasks. We then make performance comparison

on the downstream task of human pose tracking. We conduct the experiments on the validation of JHMDB [35] which has 268 videos. The annotations consist of 15 body joints for each person. Probability of correct keypoint(PCK) [101] is utilized here to examine the accuracy between result and ground-truth with different thresholds. The results in Table 3 show a consistently performance gain over previous methods, which successfully demonstrates the transferability of our method to different downstream tasks.

References

- [1] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *CVPR*, 2017.
- [2] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.
- [3] J. Deng. A large-scale hierarchical image database. In *CVPR*, 2009.
- [4] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox. FlowNet: Learning optical flow with convolutional networks. In *ICCV*, 2015.
- [5] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.
- [6] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial intelligence*, 1981.
- [7] A. Jabri, A. Owens, and A. Efros. Space-time correspondence as a contrastive random walk. In *NeurIPS*, 2020.
- [8] Z. Lai, E. Lu, and W. Xie. Mast: A memory-augmented self-supervised tracker. In *CVPR*, 2020.
- [9] Z. Li and D. Hoiem. Learning without forgetting. *IEEE transactions on PAMI*, 2017.
- [10] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019.
- [11] N. Wang, W. Zhou, and H. Li. Contrastive transformation for self-supervised correspondence learning. In *AAAI*, 2020.
- [12] X. Wang, A. Jabri, and A. A. Efros. Learning correspondence from the cycle-consistency of time. In *CVPR*, 2019.
- [13] Z. Wang, H. Zhao, Y.-L. Li, S. Wang, P. Torr, and L. Bertinetto. Do different tracking tasks require different appearance models? In *NeurIPS*, 2021.
- [14] E. Xie, J. Ding, W. Wang, X. Zhan, H. Xu, P. Sun, Z. Li, and P. Luo. Detco: Unsupervised contrastive learning for object detection. In *ICCV*, 2021.
- [15] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu. Pose flow: Efficient online pose tracking. In *BMVC*, 2018.
- [16] J. Xu and X. Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. In *ICCV*, 2021.
- [17] N. Xu, L. Yang, Y. Fan, D. Yue, Y. Liang, J. Yang, and T. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to **[Yes]**, **[No]**, or **[N/A]**. You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? **[Yes]**
- Did you include the license to the code and datasets? **[No]** The code and the data are proprietary.

- 366 • Did you include the license to the code and datasets? [N/A]
- 367 Please do not modify the questions and only use the provided macros for your answers. Note that the
 368 Checklist section does not count towards the page limit. In your paper, please delete this instructions
 369 block and only keep the Checklist section heading above along with the questions/answers below.
- 370 1. For all authors...
- 371 (a) Do the main claims made in the abstract and introduction accurately reflect the paper's
 372 contributions and scope? **[TODO]**
- 373 (b) Did you describe the limitations of your work? **[TODO]**
- 374 (c) Did you discuss any potential negative societal impacts of your work? **[TODO]**
- 375 (d) Have you read the ethics review guidelines and ensured that your paper conforms to
 376 them? **[TODO]**
- 377 2. If you are including theoretical results...
- 378 (a) Did you state the full set of assumptions of all theoretical results? **[TODO]**
- 379 (b) Did you include complete proofs of all theoretical results? **[TODO]**
- 380 3. If you ran experiments...
- 381 (a) Did you include the code, data, and instructions needed to reproduce the main experi-
 382 mental results (either in the supplemental material or as a URL)? **[TODO]**
- 383 (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they
 384 were chosen)? **[TODO]**
- 385 (c) Did you report error bars (e.g., with respect to the random seed after running experi-
 386 ments multiple times)? **[TODO]**
- 387 (d) Did you include the total amount of compute and the type of resources used (e.g., type
 388 of GPUs, internal cluster, or cloud provider)? **[TODO]**
- 389 4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
- 390 (a) If your work uses existing assets, did you cite the creators? **[TODO]**
- 391 (b) Did you mention the license of the assets? **[TODO]**
- 392 (c) Did you include any new assets either in the supplemental material or as a URL?
 393 **[TODO]**
- 394 (d) Did you discuss whether and how consent was obtained from people whose data you're
 395 using/curating? **[TODO]**
- 396 (e) Did you discuss whether the data you are using/curating contains personally identifiable
 397 information or offensive content? **[TODO]**
- 398 5. If you used crowdsourcing or conducted research with human subjects...
- 399 (a) Did you include the full text of instructions given to participants and screenshots, if
 400 applicable? **[TODO]**
- 401 (b) Did you describe any potential participant risks, with links to Institutional Review
 402 Board (IRB) approvals, if applicable? **[TODO]**
- 403 (c) Did you include the estimated hourly wage paid to participants and the total amount
 404 spent on participant compensation? **[TODO]**

405 A Appendix

406 Optionally include extra information (complete proofs, additional experiments and plots) in the
 407 appendix. This section will often be part of the supplemental material.