Localy-Aware Inter-and Intra-Video Reconstruction for Self-Supervised Correspondence Learning

Liulei Li¹,⁶*, Tianfei Zhou², Wenguan Wang³†, Lu Yang⁴, Jianwu Li¹, Yi Yang⁵
¹ Beijing Institute of Technology ² ETH Zurich ³ ReLER, AAII, University of Technology Sydney
⁴ Beijing University of Posts and Telecommunications ⁵ CCAI, Zhejiang University ⁶ Baidu Research

https://github.com/0liliulei/LIIR

Abstract

Our target is to learn visual correspondence from unlabeled videos. We develop LIIR, a locality-aware inter-and intra-video reconstruction method that fills in three missing pieces, i.e., instance discrimination, location awareness, and spatial compactness, of self-supervised correspondence learning puzzle. First, instead of most existing efforts focusing on intra-video self-supervision only, we exploit cross-video affinities as extra negative samples within a unified, inter-and intra-video reconstruction scheme. This enables instance discriminative representation learning by contrasting desired intra-video pixel association against negative inter-video correspondence. Second, we merge position information into correspondence matching, and design a position shifting strategy to remove the side-effect of position encoding during inter-video affinity computation, making our LIIR location-sensitive. Third, to make full use of the spatial continuity nature of video data, we impose a compactness-based constraint on correspondence matching, yielding more sparse and reliable solutions. The learned representation surpasses self-supervised state-of-the-arts on label propagation tasks including objects, semantic parts, and keypoints.

1. Introduction

As a fundamental problem in computer vision, correspondence matching facilitates many applications, such as scene understanding [71], object dynamics modeling [27], and 3D reconstruction [19]. However, supervising representation for visual correspondence is not trivial, as obtaining pixel-level manual annotations is costly, and sometimes even prohibitive (due to occlusions and free-form object deformations). Although synthetic data would serve an alternative in some low-level visual correspondence tasks (e.g., optical flow estimation [2]), they limit the generalization to real scenes.

Using natural videos as a source of free supervision, i.e.,

*Work done during an internship at Baidu Research.
*Corresponding author: Wenguan Wang.

Figure 1. Performance comparison over DAVIS17 val. Our LIIR surpasses all existing self-supervised methods, and is on par with many fully-supervised ones trained with massive annotations.

self-supervised temporal correspondence learning, is considered as appealing [40]. This is because videos contain rich realistic appearance and shape variations with almost infinite supply, and deliver valuable supervisory signals from the intrinsic coherence, i.e., correlations among frames.

Along this direction, existing solutions are typically built upon a reconstruction scheme (i.e., each pixel from a ‘query’ frame is reconstructed by finding and assembling relevant pixels from adjacent frame(s)) [39,40,84], and/or adopt a cycle-consistent tracking paradigm (i.e., pixels/patches are encouraged to fall into the same location after one cycle of forward and backward tracking) [31,42,52,86,90].

Unfortunately, these successful approaches largely neglect three crucial abilities for robust temporal correspondence learning, namely instance discrimination, location awareness, and spatial compactness. First, many of them share a narrow view that only considers intra-video context for correspondence learning. As it is hard to derive a free signal from a single video for identifying different object instances, the learned features are inevitably less instance-discriminative. Second, existing methods are typically built without explicit position representation. Such design seems
counter-intuitive, given the extensive evidence that spatial position is encoded in human visual system [25] and plays a vital role when human track objects [62]. Third, as the visual world is continuous and smoothly-varying, both spatial and temporal coherence naturally exist in videos. While numerous strategies are raised to address smoothness on the time axis, far less attention has been paid to the spatial case.

To fill in these three missing pieces to the puzzle of self-supervised correspondence learning, we present a locality-aware inter-and intra-video reconstruction framework – LIRR. 

First, we augment existing intra-video analysis based correspondence learning strategy with inter-video context, which is informative for instance-level separation. This leads to an inter-and intra-video reconstruction based training objective, that inspires intra-video positive correspondence matching, but penalizes unreliable pixel associations within and cross videos. We empirically verify that our inter-and intra-video reconstruction strategy can yield more discriminative features, that encode higher-level semantics beyond low-level instance invariance modeled by previous algorithms. 

Second, to make our LIRR more location-sensitive, we learn to encode position information into the representation. Although position bias is favored for intra-video correspondence matching, it is undesired in the inter-video case. We thus devise a position shifting strategy to foster the strength and circumvent the weaknesses of position encoding. We experimentally show that, explicit position embedding benefits correspondence matching. Third, we involve a spatial compactness prior in intra-video pixel-wise affinity estimation, resulting in sparse yet compact associations. For each query pixel, the distribution of related pixels is fit by a mixture of Gaussians. This enforces each query pixel to match only a handful of spatially close pixels in an adjacent frame. Our experiments show that such compactness prior not only regularizes training, but also removes outliers during inference.

These three contributions together make LIRR a powerful framework for self-supervised correspondence learning. Without any adaptation, the learned representation is effective for various correspondence-related tasks, i.e., video object segmentation, semantic part propagation, pose tracking. On these tasks, LIRR consistently outperforms unsupervised state-of-the-arts and is comparable to, or even better than, some task-specific fully-supervised methods (e.g., Fig. 1).

2. Related Work

Self-Supervised Temporal Correspondence Learning. In the video domain, correspondence matching plays a central role in many tasks (e.g., video segmentation [27], flow estimation [15, 29] and object tracking [4]). An emerging line of work tackles this problem in a self-supervised learning paradigm, by exploiting the temporal coherence in videos. One may group these work into two major classes. The first class of methods [39, 40, 84] poses a colorization proxy task (Fig. 2(a)), i.e., reconstruct a query frame from an adjacent frame, according to their correspondence. The latter type of methods [31, 42, 52, 86, 90] performs forward and backward tracking and penalizes the inconsistency between the start and end positions of the tracked pixels or regions (Fig. 2(b)). The basic idea – cycle-consistency – is also adopted in unsupervised tracking [86, 106], optical flow [56, 111] and depth estimation [33, 103]. Though impressive, these methods miss three key elements for robust correspondence matching: instance discrimination, location awareness, and spatial compactness. In response, LIRR is equipped with three specific modules (Fig. 2(c)). First, for instance discriminative representation learning, it adopts an inter-and intra-video reconstruction scheme that formulates contrast over inter-and intra-video affinities. Second, it involves position encoding into representation learning. Third, it imposes a spatial compactness prior to both correspondence learning and inference.

We are not the first to explore inter-video context. In [52], Lu et al. raise an unsupervised learning objective, i.e., discriminate between a set of surrogate video classes, but they compute this over video-level embeddings, not sufficient for pixel-wise correspondence learning. In [87], Wang et al. also simultaneously consider inter-and intra-video representation associations, but requiring pre-aligned patch pairs. In addition, they use three loss terms to address the desired intra-inter constraints, which are, however, formulated as a single unified training objective in our case. Moreover, we take a further step by addressing location awareness and spatial compactness, instead inter-video context only. In [98], Xu et al. revisit the idea of image-level similarity learning and consider frame pairs from the same videos are positive samples and pairs from different videos are negative, however, they find negative samples (i.e., inter-video context) hurt the
performance of their model. In contrast, we formulate inter- and intra-video context through a unified, pixel-wise affinity framework that boosts instance-level discrimination without sacrificing intra-instance invariance. Our results suggest that how to make a good use of negative samples for temporal correspondence learning is still an intriguing question.

**Self-Supervised Video Representation Learning.** Correspondence learning approaches that use unlabeled video data fall in a broad field of self-supervised video representation learning. Towards learning transferable video representation, diverse pretext tasks are proposed to explore different intrinsic properties of videos as free supervisory signals, including temporal sequence ordering [17, 58, 95], predicting motion patterns [1, 18, 65, 78, 85], solving space-time cubic equations [40, 43, 52], video captioning [74, 110], video retrieval [57], etc. In contrast, we pursue a more annotation-efficient solution; similar to previous self-supervised correspondence learning methods [39, 40, 52, 84], LIR is trained using unlabeled videos only. Once trained, it can be directly applied for mask propagation, without adaptation.

**3. Our Approach**

We present LIR, a self-supervised framework that learns dense correspondence from raw videos. Before elaborating on our model design (cf. §3.2), we first review the classic reconstruction-based temporal correspondence learning strategy (cf. §3.1), which serves as the basis of our LIR.

**3.1. Preliminary: Learning Temporal Correspondence through Frame Reconstruction**

Due to the appearance continuity in video, one can consider pixels in a ‘query’ frame as being *copied* from some locations of other ‘reference’ frames. In light of this, a few studies [40, 84] raise a reconstruction-based correspondence learning scheme: each query pixel struggles to find pixels in a reference frame that can best reconstruct itself.

Formally, let \( I_q, I_r \in \mathbb{R}^{H \times W \times 3} \) respectively denote a query frame and a reference frame from the same video. They are projected into a pixel embedding space by a ConvNet encoder (e.g., ResNet [24]) \( \phi : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^{h \times w \times c} \), such that \( I_q, I_r = \phi(I_q), \phi(I_r) \). The copy operator can be approximated as an inter-frame affinity matrix \( A \in [0, 1]^{h \times w} \):

\[
A(i,j) = \frac{\exp(I_q(i) \cdot I_r(j))}{\sum_j \exp(I_q(i) \cdot I_r(j'))}, \quad i, j \in \{1, \ldots, h\}
\]

where \( A(i,j) \in [0, 1] \) refers to \((i,j)\)-th element in \( A \), signifying the similarity between pixel \( i \) in \( I_q \) and pixel \( j \) in \( I_r \), and \( \cdot \) stands for the dot product. In this way, \( A \) gives the strength of all the pixel pair-wise correspondence between \( I_q \) and \( I_r \), according to which pixel \( i \) in \( I_q \) can be reconstructed by a weighted sum of pixels in \( I_r \):

\[
\hat{I}_q(i) = \sum_j A(i,j) I_r(j).
\]

The training objective of \( \phi \) is hence defined as a reconstruction loss:

\[
\mathcal{L}_{res} = ||I_q - \hat{I}_q||_2.
\]

In practice, to avoid trivial solutions caused by information leakage, an *information bottleneck* is adopted over training samples, e.g., RGB2gray operation [84], channel-wise dropout in RGB [40] or Lab [39] colorspace. After training, the representation encoder \( \phi \) is used for correspondence matching: similar to Eq. 2, the affinity \( A \) is estimated and used to *propagate* desired pixel-level entities (e.g., instance masks, key-point maps), from a reference frame to a query frame.

**3.2. LIR: Locality-Aware Inter-and Intra-Video Reconstruction Framework**

Building upon the reconstruction-by-copy scheme, LIR is empowered with three crucial yet long overlooked abili-
ties for robust correspondence learning: instance discrimination, location awareness, and spatial compactness.

**Inter-and Intra-Video Reconstruction.** With the computation of the intra-video affinity \( A \) (Eq. 1), each query pixel is forced to distinguish its counterpart (positive) reference pixels from unrelated (negative) ones within a same video, with the indicator of the reconstruction quality \( L_{\text{res}} \) (Eqs. 2-3). As both the positive and negative samples are sourced from the same video, there is less evidence for distinction among similar object instances with intra-video appearance only (Fig. 3(a)). As one single video only contains limited content, conducting correspondence matching within videos is less challenging, and inevitably hinders the discrimination potential of the learned representation [87]. These insights motivate us to improve the intra-video affinity based reconstruction scheme by further accounting for negative across-video correspondence. Concretely, given the query \( (I_q) \) and reference \( (I_r) \) frames from the same video, an *intra-inter video affinity* \( A' \in [0, 1]^{h \times w \times h' \times w'} \) is computed:

\[
A'(i, j) = \frac{\exp(I_q(i) \cdot I_r(j))}{\sum_j \exp(I_q(i) \cdot I_r(j)) + \sum_k \sum_i \exp(I_q(i) \cdot I_r(k))},
\]

where \( \{I_n\}_n \) refer to a collection of frames, which are sampled from the whole training dataset, except the source video of \( I_q \) \( (I_r) \). By additionally considering other irrelevant videos during affinity computation, both the quantity and diversity of negative samples are greatly improved, allowing us to derive a more challenging *inter-and intra-video reconstruction* scheme (Fig. 3(b)):

\[
I_q(i) = \sum_j A'(i, j) \cdot I_r(j).
\]

With Eqs. 4-5, each pixel \( i \) in the query frame \( I_q \) is required to distinguish its counterpart pixels from massive unrelated ones, which are from not only the reference frame \( I_r \) in current video, but a huge amount of irrelevant frames \( \{I_n\}_n \) in other videos. This powerful idea, yet, is elegantly achieved by the same training objective as in Eq. 3. Note that Eq. 4 normalizes intra-video correspondence over both inter-and intra-video pixel-to-pixel relevance, while Eq. 5 only uses the pixels in the reference frame \( I_r \) for reconstruction. The rationale here is that, even if the encoder \( \phi \) wrongly matches a query pixel \( i \) with a negative but similar-looking pixel \( k \) in \( I_n \), i.e., \( \exp(I_q(i) \cdot I_r(k)) \) will be large and \( \sum_j A'(i, j) \ll 1 \), the synthesized color \( I_q(i) \) will be still very different to \( I_q(i) \) and \( \phi \) will receive a large gradient from Eq. 3. Thus \( \phi \) is driven to mine more high-level semantics and context-related clues, hence reinforcing the instance-level discrimination ability (Fig. 3(c)). Fig. 3(d) shows that the representation learned with our inter-and intra-video reconstruction strategy can distinguish similar-looking dogs nearby.

Although [87] also addresses inter-video analysis based reconstruction, it conducts embedding association on *patch level*, relying on a pre-trained tracker for patch alignment. Besides, the method consumes three loss terms for supervision, which is much more complicated than ours. Further, it separately conducts inter-and intra-video affinity based reconstruction. This is problematic; when both the reference frame in current video and an irrelevant frame from other videos contain query-like pixels, there is no explicit supervision signal to determine which one should be matched.

**Position Encoding and Position Shifting.** Plenty of literature in neuroscience has revealed that human visual system encodes both appearance and position information when we perceive and track objects [25, 48, 62]. Yet existing unsupervised correspondence methods put all focus on improving appearance based representation by ConvNets, ignoring the value of position information. Although [30, 37] suggest that ConvNets can implicitly capture position information by utilizing image boundary effects, explicit position encoding has already been a core of full attention networks (e.g., Transformer [79]), and facilitated a variety of tasks (e.g., instance segmentation [93], tracking [47], video segmentation [11]). All these indicate that position encoding deserves more attention in the field of temporal correspondence learning.

Along this direction, LIIR explicitly injects a position encoding map \( P \in \mathbb{R}^{h' \times w' \times c'} \) into the feature encoder \( \phi \):

\[
I = \phi(I, P),
\]
where $P$ is added with the output feature of the first conv layer of $\phi$ and has the same size and dimension as the conv feature. We explore three position encoding strategies:

- **2D Sinusoidal Position Embedding (2DSPE):** $P$ is given with a family of pre-defined sinusoidal functions, without introducing new trainable parameters:
  \[
  P(x, y, 2u) = \sin(x \cdot \frac{u}{w}), \quad P(x, y, 2u+1) = \cos(x \cdot \frac{u}{w}),
  \]
  \[
  P(x, y, 2v+\frac{c'}{2}) = \sin(y \cdot \frac{v}{h'}), \quad P(x, y, 2v+1+\frac{c'}{2}) = \cos(y \cdot \frac{v}{h'}),
  \]
  where $x \in [0, w')$, $y \in [0, h')$ specify the horizontal and vertical positions, $u, v \in [0, c'/4)$ specify the dimension, and $\varepsilon = 10^{-4}$. The horizontal (vertical) positions are encoded in the first (second) half of the dimensions. 2DSPE naturally handles resolutions that are unseen during training.

- **1D Absolute Position Embedding (1DAPE):** 1DAPE is the most heavy-weight strategy: the whole $P$ is a learnable parameter matrix without any constraint.

- **2D Absolute Position Embedding (2DAPE):** As in [14], two separate parameter sets: $X \in \mathbb{R}^{w \times c'/2}$ and $Y \in \mathbb{R}^{h \times c'/2}$, are learned for encoding the horizontal and vertical positions, respectively, and combined to generate $P$.

With our intra-inter video affinity (Eq. 4), exploiting position information in intra-video correspondence matching, i.e., $\{\exp(I_q(i) \cdot I_r(j))\}$, addresses local continuity resides in videos. However, for inter-video pixel relevance computation, i.e., $\{\exp(I_q(i)I_r(k))\}_{n,k}$, such position prior is undesirable, as it inspires the query pixel $i$ in $I_q$ to prefer matching these pixels with similar positions in other irrelevant videos $\{I_n\}_n$. To eliminate such position encoding induced bias from inter-video correspondence matching, we design a position shifting strategy (Fig. 4(a)). During training, for $I_n$ from other videos, we circularly shift the position encoding vectors in $P$ by a random step in horizontal and vertical axes, respectively. The reason why we adopt random shifting with circular boundary conditions, instead of random shuffling, is to preserve the spatial layout in the modulated position encoding map $P$. Then $P$ and $I_n$ are fed into $\phi$ for inter-video correspondence matching, and $P$ related gradients are abandoned if learnable 1DAPE or 2DAPE is used. Note that the standard position encoding $P$ is applied for the query ($I_q$) and reference ($I_r$) frames and updated normally. Fig. 4(b) intuitively shows that merging position information into visual representation can enable robust correspondence matching even with confusing background and fast motion. In §4.4, we will quantitatively verify that 1DAPE is more favored and indeed boosts the performance.

Spatial Compactness Prior. As the visual world is continuous and smoothly-varying, it is reasonable to assume appearances in video data change smoothly both in spatial and temporal dimensions. For correspondence learning, the temporal coherence has been extensively studied, while the spatial continuity received far less attention. To reduce search region, some existing methods [39, 40, 84] restrict correspondence matching within a local window, considering spatial regularities in a simple way. To make a better use of the spatial continuity, we augment the original reconstruction objective with an additional prior, termed as spatial compactness. Such a prior poses constraints on the spatial distribution of associated pixels, leading to sparse and coherent solutions. Specifically, given the query ($I_q$) and reference ($I_r$) frames, we expect i) each query pixel $i$ to be only matched with a small number of reference pixels, and ii) the matched reference pixels to be clustered. For a query pixel $i$ and its matching ‘heatmap’ $A_i = [A(i, j)]_{j \in [0, 1]}^{h \times w}$, w.r.t. $I_r$, we assume $A_i$ follows a mixture of $M$ 2D Gaussian distributions:

\[
|P(x, y) = \sum_{m=1}^{M} \omega_m N(x, y | \mu_m, \Sigma_m),
\]

where $(x, y)$ specifies the coordinate of a pixel location. We set $\{\mu_m = [\mu_{x,m}, \mu_{y,m}]\}_m$ as the coordinates of top-$M$ scores in $A_i$, and set $M = 2$ to address sparse robust matching. Other parameters, i.e., $\{\Sigma_m = [\sigma^2_{x,m}, 0; 0, \sigma^2_{y,m}]\}_m$, $\{\omega_m\}_m$, can be estimated efficiently from $A_i$ without incurring high computational cost. In this way, we can derive a ‘compact’ matching heatmap $A_i \in [0, 1]^{h \times w}$ for each query pixel $i$, and eventually have $\bar{A} = [\bar{A}_i]_{i \in [0, 1]}^{h \times w}$. Such spatial compactness prior $\bar{A}$ is fully aware of i) and ii), and used to regularize our representation learning:

Figure 5. Illustration of spatial compactness prior (§3.2).
For fair comparison, our feature correspondence by inspiring sparse and compact solutions. Shows that our spatial compactness prior helps build reliable matching filtering [9], while both of them can or top-k matching filtering [9], with a few of instances of our mixture-Gaussian based compactness prior, despite our different task settings. Fig. 5 shows that our spatial compactness prior helps build reliable correspondence by inspiring sparse and compact solutions. Related experiments can be found in §4.4.

### 3.3. Implementation Details

**Network Configuration.** For fair comparison, our feature encoder $\phi$ is implemented as ResNet-18 [24] as in [31, 84, 98]. Following [34, 39, 40], $\times 2$ downsampling is only made in the third residual block. Thus $\phi$ finally outputs 256 feature maps of $1/4$ size of the input, i.e., $h = \frac{H}{4}, w = \frac{W}{4}, c = 256$. The position embedding is added to feature after the first $7 \times 7$ Conv-BN-ReLU layer, i.e., $h' = \frac{h}{2}, w' = \frac{w}{2}, c' = 64$.

**Training:** LIIR is trained from scratch on two NVIDIA RTX-3090 GPUs and only uses the raw videos from Youtube-VOS [99]. Each training image is resized into $256 \times 256$ and channel-wise dropout in Lab colorspace [39] is adopted as the information bottleneck. Adam optimizer is used. At the initial 30 epochs, only intra-video reconstruction learning is adopted for warm-up, with a learning rate of $10^{-3}$ and batch size of 32. Then we conduct inter-video reconstruction learning with spatial compactness based regularization at the next 5 epochs, with a learning rate of $10^{-4}$ and batch size of 12. We online maintain a memory bank of 1,440 frames from different videos. For each query pixel, we sample 4 feature points from each stored frame, i.e., a total of $1,440 \times 4$ negative samples are used for the inter-video correspondence computation, and we employ the moving average strategy [23, 76, 97] for parameter updating.

**Testing:** Once LIIR finishes training, there is no fine-tuning when applied to downstream tasks. Note that we utilize the compactness prior enhanced inter-frame affinity $\tilde{A}$ for mask propagation. As in [39, 61], we take multiple frames as reference for the full use of temporal context: at time step $t$ when applying a to-frame body part propagation (§4.2), and pose keypoint tracking (§4.3). As in conventions [31, 84, 98], all these tasks are to propagate the first frame annotation to the whole video sequence, and we use our model to compute inter-frame dense

![Table 1. Quantitative results for video object segmentation (§4.1) on DAVIS17 [68] val. For size of datasets, we report (#raw images, length of raw videos) for self-supervised methods and (#image-level annotations, #pixel-level annotations) for supervised methods.](image)

![Figure 6. Qualitative results for video object segmentation (§4.1), on DAVIS17 [68] val (left) and Youtube-VOS [99] val (right).](image)
correspondences. In §4.4, we conduct a set of ablative studies to examine the efficacy of our essential model designs.

4.1. Results for Video Object Segmentation

Dataset. We first test our method on \( \text{val} \) sets of two popular video object segmentation datasets, \( \text{i.e., DAVIS}_{17} [68] \) and YouTube-VOS [99]. There are 30 and 474 videos in DAVIS\(_{17} \) and YouTube-VOS \( \text{val} \) sets, respectively.

Evaluation Metric. Following the official protocol [68], we use the region similarity (\( J \)) and contour accuracy (\( F \)) as the evaluation metrics. Note that the scores on YouTube-VOS are respectively reported for \( \text{seen} \) and \( \text{unseen} \) categories, obtained from the official evaluation server.

Performance on DAVIS\(_{17}\): As illustrated in Table 1, our \( \text{LIR} \) consistently outperforms all existing self-supervised methods across all the evaluation metrics. For example, it surpasses current best-performing self-supervised method, \( \text{i.e., CLTC} [34] \), in terms of mean \( J \& F \) (72.1 vs. 70.3). In addition, even without using \( \text{any} \) manual annotations for training, \( \text{LIR} \) achieves very competitive segmentation performance in comparison with some famous supervised models [5,81] trained with massive pixel-wise annotations.

Performance on YouTube-VOS \( \text{val} \). Table 2 reports performance comparison of \( \text{LIR} \) against four self-supervised competitors on YouTube-VOS \( \text{val} \). It can be observed that \( \text{LIR} \) sets new state-of-the-art. In particular, \( \text{LIR} \) yields an overall score of 69.3\%, surpassing the second-best (\( \text{i.e., CLTC} [34] \)) and third-best (\( \text{i.e., MAST} [39] \)) approaches by 2.0\% and 5.1\%, respectively. Further, \( \text{LIR} \) even outperforms some famous supervised methods (\( \text{i.e., OSVOS} [5] \) and PreMVOS [53]), especially for the \( \text{unseen} \) categories, clearly demonstrating its remarkable generalization ability.

Table 2. Quantitative results for video object segmentation (§4.1) on Youtube-VOS [99] \( \text{val} \).

| Methods          | Sup. | Overall | \( J \uparrow \) | \( F \uparrow \) | \( J \uparrow \) | \( F \uparrow \) |
|------------------|------|---------|-----------------|-----------------|-----------------|-----------------|
| Colorization[64] | ✓    | 38.9    | 43.1            | 38.6            | 36.6            | 37.4            |
| CorrFlow[40]     |      | 46.6    | 50.6            | 46.6            | 43.8            | 45.6            |
| MAST[39]         |      | 64.2    | 63.9            | 64.9            | 60.3            | 67.7            |
| CLTC[34]†        |      | 67.3    | 66.2            | 67.9            | 63.2            | 71.7            |
| \( \text{LIR} \) |      | 69.3    | 67.9            | 69.7            | 66.7            | 73.8            |

†: using task-specific model weights and architectures.

Table 3. Quantitative results for part propagation (§4.2) and pose tracking (§4.3), on VIP [107] \( \text{val} \) and JHMDB [35] \( \text{val} \).

| Methods          | Sup. | VIP mIoU \( \uparrow \) | AP \( \uparrow \) | JHMDB PCK@0.1 \( \uparrow \) | PCK@0.2 \( \uparrow \) |
|------------------|------|------------------------|-----------------|-----------------------------|-----------------------------|
| TimeCycle[30]    | ✓    | 28.9                   | 15.6            | 57.3                        | 78.1                        |
| UVC[42]          |      | 34.1                   | 17.7            | 58.6                        | 79.6                        |
| CRW[31]          |      | 38.6                   | -               | 59.3                        | 80.3                        |
| ContrastCorr[87] |      | 37.4                   | 21.6            | 61.1                        | 80.8                        |
| VFS[98]          |      | 39.9                   | -               | 60.5                        | 79.5                        |
| CLTC[34]†        |      | 37.8                   | 19.1            | 60.5                        | 82.3                        |
| JSTG[105]        |      | 40.2                   | -               | 61.4                        | 85.3                        |
| \( \text{LIR} \) |      | 41.2                   | 22.1            | 60.7                        | 81.5                        |

†: using task-specific model weights and architectures.

Qualitative Results. Fig. 6 depicts visual results on representative videos in the datasets. As seen, \( \text{LIR} \) is able to establish accurate correspondences under various challenging scenarios, \( \text{e.g., scale changes, small objects and occlusions.} \)

4.2. Results for Body Part Propagation

Dataset. We next evaluate our model performance for body part propagation. Experiments are conducted on VIP \( \text{val} \) [107], which contains 50 videos with annotations of 19 human semantic part categories (\( \text{e.g., hair, face, dress.} \))

Evaluation Metric. As suggested by VIP [107], we adopt mean intersection-over-union (mIoU) and mean Average Precision (mAP) metrics for evaluation of semantic-level and instance-level parsing, respectively.

Performance. As shown in Table 3, \( \text{LIR} \) achieves state-of-the-art performance on both semantic-level and instance-level parsing. This indicates that \( \text{LIR} \) can generate strong representation which models both cross-instance discrimination and intra-instance invariance well. Fig. 7 depicts some visualization results on two representative videos. \( \text{LIR} \) achieves temporally stable results and shows robustness to typical challenges (\( \text{e.g., pose variations, occlusions.} \))
Table 4. Detailed analysis of essential components of LIIR on DAVIS17 val and VIP [107] val. See §4.4 for details.

Table 5. A set of ablation studies on DAVIS17 [68] val and VIP [107] val. See §4.4 for details.

(a) Inter-and Intra-Video Recons. (b) Position Encoding

4.4. Diagnostic Experiment

For further detailed analysis, we conduct a series of ablative studies on DAVIS17 [68] val and VIP [107] val sets.

Key Component Analysis. We first examine the efficacy of essential components of LIIR, i.e., inter-and intra-video reconstruction, position encoding, and spatial compactness. The results are summarized in Table 4, where position encoding is implemented as 1DAPE, and compactness prior is used during both training and inference stages. When separately comparing row #2 - #4 with the baseline (MAST[39]) in row #1, we can observe that each individual module indeed boosts the performance. For example, on DAVIS17 val, inter-and intra-video reconstruction, position encoding, and spatial compactness prior respectively bring 3.4%, 1.6%, and 3.1% $\mathcal{J}\&F$ gains. This verifies our core insight that these three elements are crucial for correspondence learning. Finally, in row #5, we combine all the three components together – LIIR, and obtain the best performance. This suggests that these modules are complementary to each other, and confirms the effectiveness of our whole design.

Inter-and Intra-Video Reconstruction. We next study the impact of increasing the number of negative samples, i.e., frames from other irrelevant videos used for inter-video correspondence computation (Eq. 4). In Table 5a, row #1 gives scores of learning without considering inter-video correspondence. In this case, the results are unsatisfactory. When more negative samples are involved (i.e., 0→1,440), better performance can be achieved (i.e., 69.2→72.1 on DAVIS17 val, 38.4→41.2 on VIP val). Finally we use 1,440 negative samples for inter-video reconstruction based learning, which is the maximum number allowed by our GPU.

Position Encoding. To determine the effect of our position encoding module, we then report the performance with different encoding strategies in Table 5b. As seen, the non-learnable strategy, 2DSPE, hinders the performance, while the learnable alternatives, i.e., 1DAPE and 2DAPE, lead to better results. Compared with 2DAPE, 1DAPE is more favored, probably due to its high flexibility and capacity.

Position Shifting. We further study the influence of our position shifting strategy on performance. As shown in Table 5c, we consider two alternatives, i.e., ‘NAN’ and ‘position shuffling’. ‘NAN’ refers to using the normal position encoding map $P$ without any modulation during inter-video correspondence matching. Compared with ‘position shifting’, ‘NAN’ suffers from performance degradation (i.e., 72.1→71.3 on DAVIS17 val, 41.2→40.6 on VIP val), showing the negative effect of the position-induced bias. The other baseline, ‘position shuffling’, i.e., randomly shuffling the position encoding map for inter-video affinity computation, though better than ‘NAN’, is still worse than ‘position shifting’. This is because it destroys the spatial layouts.

Spatial Compactness Prior. The spatial compactness prior (Eq. 7) is used to regularize inter-video correspondence matching during both training and inference stages. In Table 5d, we quantitatively identify the performance contribution of our spatial compactness prior in different stages.

5. Conclusion

We presented a self-supervised temporal correspondence learning approach, LIIR, that makes contributions in three aspects. First, going beyond the popular intra-video analysis based learning scheme, we further enforce separation between intra- and inter-video pixel associations, enhancing instance-level feature discrimination. Second, with a clever position shifting strategy, we bring the advantages of position encoding into full play, while avoiding its undesirable impact at the same time. Third, a spatial compactness prior is introduced to regularize representation learning and improve correspondence inference. The effectiveness was thoroughly validated over various label propagation tasks.

Acknowledgements This work was supported in part by the Beijing Natural Science Foundation under Grant L191004, CCF-Baidu Open Fund, and ARC DECRa DE220101390.
References

[1] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In ICCV, 2015. 3
[2] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. IJCV, 92(1):1–31, 2011. 1
[3] Linchao Bao, Baoyuan Wu, and Wei Liu. Can in mrf: Video object segmentation via inference in a cnn-based higher-order spatio-temporal mrf. In CVPR, 2018. 1
[4] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In ECCV, 2016. 2
[5] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In CVPR, 2017. 3, 6, 7
[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. arXiv preprint arXiv:2104.14294, 2021. 6
[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709, 2020. 3
[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020. 3
[9] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In CVPR, 2021. 6
[10] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Re-thinking space-time networks with improved memory coverage for efficient video object segmentation. In NeurIPS, 2021. 1
[11] Hai Ci, Chunyu Wang, and Yizhou Wang. Video object segmentation by learning location-sensitive embeddings. In ECCV, 2018. 4
[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, 2009. 6
[13] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by contrast prediction. In ICCV, 2015. 3
[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. 5
[15] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philipp Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In ICCV, 2015. 2
[16] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In CVPR, 2019. 3
[17] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In CVPR, 2017. 3
[18] Chuang Gan, Boqing Gong, Kun Liu, Hao Su, and Leonidas J Guibas. Geometry guided convolutional neural networks for self-supervised video representation learning. In CVPR, 2018. 3
[19] Andreas Geiger, Julius Ziegler, and Christoph Stiller. Stereosc: Dense 3d reconstruction in real-time. In 2011 IEEE Intelligent Vehicles Symposium, 2011. 1
[20] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In ICLR, 2018. 3
[21] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In ECCV, 2020. 3
[22] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. NeurIPS, 2020. 3
[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In CVPR, 2020. 3, 6
[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016. 3, 6, 7
[25] Robert F Hess and Ian E Holliday. The coding of spatial position by the human visual system: effects of spatial scale and contrast. Vision Research, 32(6):1085–1097, 1992. 2, 4
[26] R Devin Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In ICLR, 2019. 3
[27] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. Videomatch: Matching based video object segmentation. In ECCV, 2018. 1, 2
[28] Xuhua Huang, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Fast video object segmentation with temporal aggregation network and dynamic template matching. In CVPR, 2020. 1
[29] Eddy Ilg, Nikolaus Mayer, Tommoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In CVPR, 2017. 2
[30] Md Amirul Islam, Sen Jia, and Neil DB Bruce. How much position information do convolutional neural networks encode? In ICLR, 2020. 4
[31] Allan Jabri, Andrew Owens, and Alexei A Efros. Space-time correspondence as a contrastive random walk. In NeurIPS, 2020. 1, 2, 6, 7
[32] Varun Jampani, Raghudeep Gadde, and Peter V. Gehler. Video propagation networks. In CVPR, 2017. 3
[33] Joel Janai, Fatma Guney, Anurag Ranjan, Michael Black, and Andreas Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In ECCV, 2018. 2
[104] Jae Shin Yoon, François Rameau, Jun-Sik Kim, Seokju Lee, Seunghak Shin, and In So Kweon. Pixel-level matching for video object segmentation using convolutional neural networks. In ICCV, 2017. 3

[105] Zixu Zhao, Yueming Jin, and Pheng-Ann Heng. Modelling neighbor relation in joint space-time graph for video correspondence learning. In ICCV, 2021. 6, 7

[106] Jilai Zheng, Chao Ma, Houwen Peng, and Xiaokang Yang. Learning to track objects from unlabeled videos. In ICCV, 2021. 2

[107] Qixian Zhou, Xiaodan Liang, Ke Gong, and Liang Lin. Adaptive temporal encoding network for video instance-level human parsing. In ACM MM, 2018. 7, 8

[108] Tianfei Zhou, Jianwu Li, Xueyi Li, and Ling Shao. Target-aware object discovery and association for unsupervised video multi-object segmentation. In CVPR, 2021. 3

[109] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation. In AAAI, 2020. 3

[110] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In CVPR, 2020. 3

[111] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In ECCV, 2018. 2