Mining Better Samples for Contrastive Learning of Temporal Correspondence

Sangryul Jeon[1], Dongbo Min[2,*], Seungryong Kim[3], Kwanghoon Sohn[1,*]
[1]Yonsei University, [2]Ewha Womans University, [3]Korea University
{cheonjsr,khsohn}@yonsei.ac.kr
dbmin@ewha.ac.kr, seungryong_kim@korea.ac.kr

Abstract

We present a novel framework for contrastive learning of pixel-level representation using only unlabeled video. Without the need of ground-truth annotation, our method is capable of collecting well-defined positive correspondences by measuring their confidences and well-defined negative ones by appropriately adjusting their hardness during training. This allows us to suppress the adverse impact of ambiguous matches and prevent a trivial solution from being yielded by too hard or too easy negative samples. To accomplish this, we incorporate three different criteria that ranges from a pixel-level matching confidence to a video-level one into a bottom-up pipeline, and plan a curriculum that is aware of current representation power for the adaptive hardness of negative samples during training. With the proposed method, state-of-the-art performance is attained over the latest approaches on several video label propagation tasks.

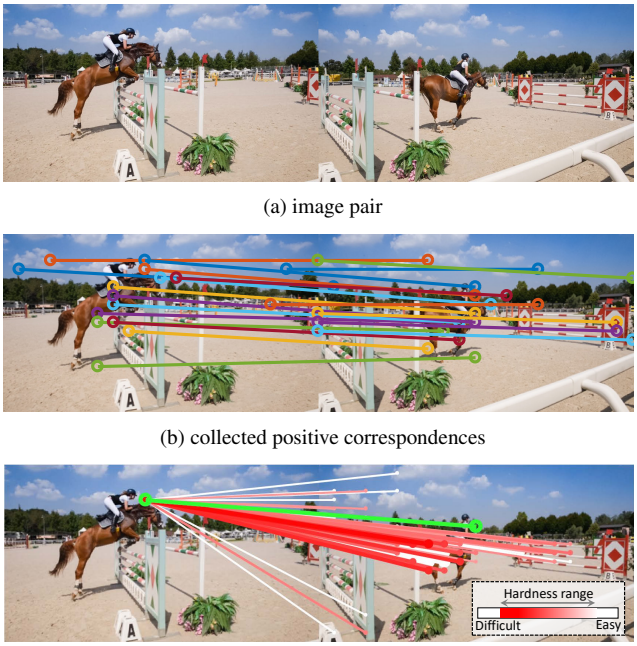
1. Introduction

Learning pixel-level representation for visual correspondence can facilitate numerous downstream applications [26, 5, 31]. In contrast to the image-level representation which demands a semantic invariance among object instances of the same category, the pixel-level representation further requires the fine-grained localization ability to discriminate a distinctive match from all possible matching candidates.

Supervising the representation for pixel-level correspondence, however, often requires costly annotations defined for all pixels. Constructing such dense annotations become even more problematic in the presence of occlusions and non-rigid object deformations. Synthetically generated data [34, 4, 43] would be an alternative of high-quality annotation maps, but it has the downside of limiting generalization to real scenes.

Several methods [51, 52, 30, 29, 28, 20] have attempted to alleviate this by leveraging abundant unlabeled videos as a source of free supervision. Unlike to the synthetic supervisions [34, 4, 43], richer appearance and shape variations captured from real world strengthen their generalization ability. Furthermore, the nature of temporal coherence in video allows the correspondences likely to exist across adjacent frames, providing useful constraints for training. Standing on these bases, they first track points over time and then learn from the inconsistency between the original points and tracked ones in a form of reconstruction.

Establishing correspondences in the unconstrained videos, however, imposes additional challenges due to the existence of temporal discontinuities. For frames sampled with a large temporal stride, the self-supervised loss often



(a) image pair

(b) collected positive correspondences

(c) collected negative correspondences for a positive one (green-colored)

Figure 1. **Visualization of collected samples used in our contrastive learning**: given (a) an image pair, we collect (b) positive correspondences that are aware of matching uncertainty, and (c) negative ones that are neither too easy nor too difficult.

*Co-corresponding author

becomes invalid in the presence of complex object deformations, illumination changes, and occlusions. This could be partially addressed by considering more matching candidates over additional adjacent frames that are likely to contain valid correspondences, *e.g.* tracking cycle with multiple lengths [52] or augmenting the model with a memory bank [28]. However, the number of ambiguous matches increases at the same time due to the larger candidates, which is problematic as the loss is evenly influenced by them.

Very recently, the concurrent work [20] casts this task into a probabilistic inference of a path through the graph constructed from an input video. In contrast to the previous works [51, 52, 30, 29, 28] that learn from a reconstruction-based loss, their consideration of negative correspondences produce better performances through the contrastive objective [37]. However, the formulation of assigning graph nodes only within an image is still challenged by occlusions where the correspondences disappear to be out of the given nodes. Furthermore, composing negative examples with all pairs of nodes that do not meet cycle-consistency constraint may let too easy negative samples to degrade the contribution of harder ones that are useful for contrastive learning.

In this work, we present a novel contrastive learning approach that is capable of collecting well-defined positive correspondences by measuring their uncertainties and well-defined negative ones by controlling their hardness during training, as exemplified in Fig. 1. Unlike previous works, our approach is able to suppress the adverse impact of ambiguous matches and simultaneously prevent a trivial solution from being yielded by too easy negative samples.

Specifically, to measure reliable matching confidence without ground-truth annotation, we formulate a bottom-up pipeline by incorporating three different criteria; Starting from checking forward-backward consistency, the initial scores are further optimized by solving the optimal transport problem, which enforces the total uncertainty for all possible matches over an image to be minimized, and then imposing a temporal coherence constraint to be less susceptible to background clutters and repetitive patterns. Furthermore, from the observation in metric learning literature [53, 46] that using too hard or too easy negative samples may produce worse representations, we collect semi-hard negative samples by specifying the upper and lower thresholds of their hardness. These thresholds are dynamically reconfigured during training with the proposed curriculum that is conditioned on the capability of the current representation. With the proposed method, state-of-the-art performance is attained over the latest approaches on several video label propagation tasks.

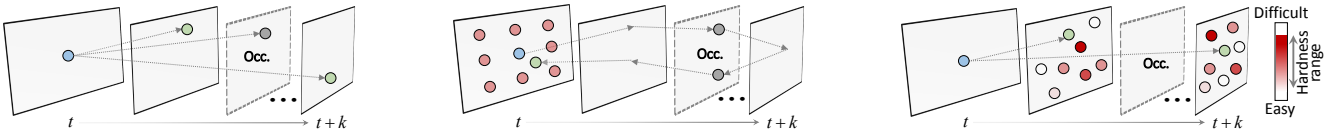
2. Related Work

Self-supervised learning of visual representation Techniques for self-supervised representation learning have recently provided remarkable results closing the gap to supervised methods [18, 8, 16]. Generally, they generate supervisory signals by holding part of the input data for defining a fixed target and then minimizing the discrepancy between the target and the predicted missing parts [39, 7, 35, 13, 14].

Yet, the rapid progress of self-supervised representation learning on an image or video has not translated into equivalent advances in learning pixel-level representation [51, 52, 30, 29, 28]. The key idea is similar to the approaches for an image or video; The proxy task is defined as tracking along video frames, and the model learns by reconstructing the attributes between given query points and the tracked ones. While these reconstruction-based approaches are often challenged due to false positive targets defined at occluded regions, we explicitly disambiguate and discard them from being utilized in the objective.

Negative mining Most recently, contrastive methods [37] have shown great performance gains by utilizing randomly sampled negative examples to normalize the objective. Though numerous variants have shown improved performances [18, 8, 16], selection strategies for negative samples has not been deeply explored. Meanwhile, it has a rich line of research in the metric learning community. Most of the literature [53, 46] observed that it is helpful to use negative samples, while showing that mining the very hardest negatives can hurt performance. Similar margin-based approach also have been popularly employed in pixel-level representation learning, including patch matching [17], depth estimation [57], and optical flow [9]. In the unsupervised setting where the annotation of correct matches is unavailable, some methods [25] attempted to heuristically set the negative matching candidates within a local window that is centered at the point assumed to be positive. However, the degree of difficulty cannot be regulated in their formulation, and choosing the negatives as the nearest candidates may be too difficult to learn, thus limiting the performance.

Optimal transport Optimal transport (OT) provides a way to estimate an optimal distance between two distributions. An advantage of OT is its robustness to noise which is useful for many computer vision applications, mainly for domain adaptation [10, 3], generative model [1, 15], and graph matching [55, 54]. OT also has been employed in recent literature for visual correspondence, such as 3D shape matching and surface registration [49], scene flow estimation [42], and semantic correspondence [32]. To ensure the reliability of initial matching costs (*i.e.* dissimilarities) during the optimization, they require ground-truth supervisions or pre-trained network parameters with ImageNet [12]. In contrast, the proposed bottom-up pipeline can yield reliable confidence scores without the need of annotations by incorporating tailored criteria for self-supervised learning including OT formulation.



(a) Reconstruction loss in [51, 52, 30, 29, 28] (b) Contrastive loss in [20] (c) Ours

Figure 2. **Illustration of the self-supervised loss by means of**: (a) reconstruction in [51, 52, 30, 29, 28], (b) contrast in [20], and (c) our method. We denote by blue, green, red, and gray circle, respectively, query, positive, negative, and occluded sample. Unlike previous works, our approach is capable of collecting well-defined positive correspondences by computing their confidence scores and well-defined negative ones by appropriately adjusting their hardness.

3. Problem Statement

Given a video sequence with an interval K as a collection of images $\{I^t, ..., I^{t+K}\}$, self-supervised learning of pixel-level representation involves first tracking a pixel $i = [i_{\mathbf{x}}, i_{\mathbf{y}}]^T$ from I^t to I^{t+K} and then minimizing the discrepancy between the original pixel and the tracked one. Analogously to the classical matching pipeline [41], tracking begins with encoding the similarity S between two pixels i and j from a pair of images I^t and I^{t+k} where $k \in [1, K]$. To this end, dense feature maps F^t and F^{t+k} are first extracted through the shared parameters $\mathbf{W}$ and its cosine distance is then computed as

$$S_{ij}^k = \langle F_i^t, F_j^{t+k} \rangle / \|F_i^t\|_2 \|F_j^{t+k}\|_2 \in \mathbb{R}^{n \times n} \quad (1)$$

where n denotes the number of pixels in an image.

Following a pioneering work of [51], several works [29, 28] proposed to learn their model by reconstructing the color of original pixel j from the tracked ones as illustrated in Fig. 2 (a):

$$\mathcal{L}_{rec}^k = \sum\nolimits_j ||\phi_j^k - \sum\nolimits_i P_{ij}^k \cdot \phi_i||_2^2, \quad (2)$$

where ϕ is the color of reconstruction target and P is the matching probability converted from similarity scores as $P_{ij}^k = \exp(S_{ij}^k) / \sum_l \exp(S_{lj}^k)$. Similar to the soft argmax operator in [24], they conducted tracking by computing the weighted sum of the attributes with corresponding matching probabilities. For an another way of tracking, some methods [52, 30] employed deterministic localizer modules with cycle-consistency constraint[1] [38, 23], *e.g.* a spatial transformer network [52] and an object-level tracker [30].

While these approaches consider various kinds of attributes for the reconstruction target, as summarized in Tab. 1, the reconsturction often becomes invalid when faced with temporal discontinuity as the target disappears due to occlusions. A possible approach to address this is to consider additional matching candidates from multiple adjacent frames that are likely to contain valid correspondences, such as augmenting the model with memory component [28] or tracking cycles of different lengths [52]. However, the number of ambiguous matches increases at the same time due to larger candidates, and they are not explicitly treated to be discarded during training.

[1]Fig. 2 (a) illustrates the methods [51, 29, 28] that track only in a forward direction of time. Meanwhile, other reconstruction-based methods [52, 30] track first forward and then backward to leverage cycle-consistency constraint.

Methods	Training objective			Curricular policy
	Att.	Confid.	Negative	
Colorization [51]	RGB	✗	✗	✗
CorrFlow [29]	Lab	✗	✗	Fixed
MAST [28]	Lab	✗	✗	✗
TimeCycle [52]	F&L	✗	✗	✗
UVC [30]	F&L	✗	✗	✗
CRW [20]	F	✗	All	✗
Ours	F	✓	Semi-hard	Dynamic

Table 1. **Comparison of recent related work.** The table indicates employed training objective and curriculum. The abbreviation of "Att." and "Confid." denotes attribute and confidence estimation, respectively. For the type of the attribute, we denote by "F" and "L", embedded feature and location, respectively.

Recent concurrent work [20] proposed to learn by finding probabilistic paths in a graph constructed from an input video. As shown in Fig. 2 (b), the pathfinding is conducted in a contrastive setting where a query pixel itself is assumed to be a positive correspondence with cycle-consistency constraint and all other ones belong to be a negative set. Denoting $P_{ij}^{a \to b}$ as a long-range matching probability from I^a to I^b, the loss function is defined as

$$\mathcal{L}_{con}^k = -\sum\nolimits_{ij} [\mathbf{I}_{n \times n} \log P^{t \to t+k} P^{t+k \to t}]_{ij} \quad (3)$$

where $\mathbf{I}_{n \times n}$ is a $n \times n$ identity matrix and $P^{t \to t+k} = \prod_{m=t}^{t+k-1} P^{m \to m+1}$.

However, their formulation assumes that total matching probabilities are conserved within a graph, *i.e.* a cycle of time, which is often violated when matches disappear due to occlusions, cutting off intermediate correspondence trajectory. Furthermore, contrasting the positive correspondence with all remaining candidates may not be the best choice for yielding a good representation. As the negative samples become mixed with hard and easy ones, the easy ones can reduce the contribution of harder ones by means of softmax normalization, causing the gradients to be vanished.

4. Method

4.1. Overview

We address the forementioned limitations effectively by collecting well-defined positive correspondences based on their uncertainties and well-defined negative ones that are neither too easy nor too difficult, as shown in Fig. 2 (c).

Fig. 3 overviews the proposed method. From a pair of query and k^{th} key feature maps, we first compute similarty scores of all possible matches with Equ. (1), and then collect positive samples by computing their confidence scores and negative samples by appropriately adjusting their hardness with the curriculum. Finally, we learn our model by contrasting the positive samples with the semi-hard negative ones collected across K time steps.

4.2. Mining Positive Correspondence

We design a bottom-up pipeline to measure the matching uncertainty consisting of three different criteria; Forward-backward consistency in a pixel-level, optimal transport optimization in an image-level, and temporal coherence constraint in a video-level.

Checking consistency To establish an initial set of confidence scores, we start from classic uncertainty measurement; checking forward-backward consistency [38, 23]. This can be done by applying the argmax operator to the similarity scores twice for forward and backward direction, respectively. Yet, assigning binary labels (0 or 1) may produce true-negative matches, *i.e.* true matches that do not meet the criterion due to the disorganized representations in early training. We alleviate this by adopting a soft consistency criterion of [44], such that

$$Q_{ij}^k = \frac{(S_{ij}^k)^2}{\max_i S_{ij}^k \cdot \max_j S_{ij}^k} \tag{4}$$

where Q_{ij} equals one if and only if the match between i and j satisfies the forward-backward consistency constraint, and becomes less than one otherwise.

Solving optimal transport problem The confidence score in Q is computed in a pairwise manner, *i.e.* individually for each pixel, and does not care about mutual relation between pixels, often leading to the many-to-one matching problem. To address this, we refine the initial scores in a non-local manner by solving the optimal transport problem such that the total uncertainty in an image is minimized as

$$\begin{aligned} &\min_{T^k} [\sum\nolimits_{ij} T_{ij}^k (1 - Q_{ij}^k)] \\ &\text{subject to} \quad T^k \mathbf{1} = \mathbf{1}, (T^k)^T \mathbf{1} = \mathbf{1}, \end{aligned} \tag{5}$$

where T is the transport plan and $\mathbf{1}$ is $n \times 1$ vector of ones. To avoid the many-to-one matching problem, both row-wise and column-wise sums of T are constrained with the regularization term that prevents too large values from being assigned to some rows or columns of T. By introducing additional regularization of the negative entropy term, we efficiently solve Equ. (5) using Sinkhorn algorithm [11] that allows us to scale to massive pixels from video sequences. Note that the solution described in Alg. 1 only consists of matrix multiplication and exponential operations.

The optimal transport has been also adopted in recent correspondence estimation approaches [32, 42, 45], but they require strong supervisions or guaranteed representation from pre-trained network to yield reliable matching costs during optimization. In contrast, we incorporate tailored criteria for self-supervised learning with optimal transport into the proposed bottom-up pipeline, enabling us to obtain reliable confidence scores in the absence of ground-truth annotation.

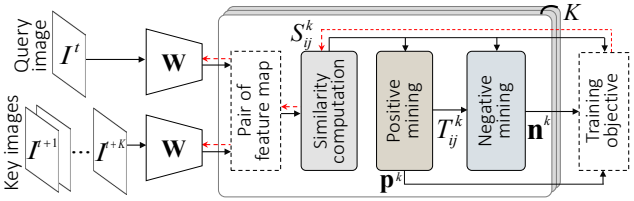


Figure 3. **Visualization of the proposed learning framework**: given a query and corresponding key images with an interval K, their feature maps are extracted through shared embedding networks with parameter $\mathbf{W}$. From a pair of query and k^{th} feature maps, we compute their similarity scores and then collect positive and negative correspondences to learn in the contrastive setting.

Imposing temporal coherence constraint Lastly, we employ the nature of temporal coherence as a video-level constraint to further make the refined confidence scores T less susceptible to ambiguous matches due to background clutter or repetitive patterns. Specifically, we retain only the matches within a local window $\mathcal{M}_i(w^k)$ centered at the query pixel i with a radius w^k:

$$C_{ij}^k = W_{ij}^k \cdot T_{ij}^k \;\text{ s.t. }\; W_{ij}^k = \begin{cases} 1, & \text{if } j \in \mathcal{M}_i(w^k) \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where the radius w^k is dilated with respect to the temporal distance k.

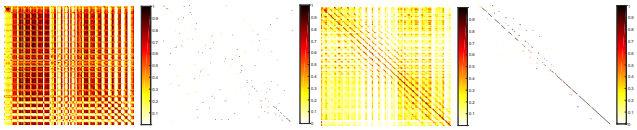
Selection Selection of positive correspondence set $\mathbf{p}$ can be done by simply picking the best match, a pair of two pixels (i, j), whose confidence score is one, such that

$$\mathbf{p}^k = \{(i, j) | C_{ij}^k = 1\}. \tag{7}$$

Without the additional network parameters or heavy computational loads, our bottom-up pipeline allows plausible matches to be retained as shown in Fig. 4 and Fig. 5.

(a) $Q = 1$ from $\mathbf{W}^0$ (b) $C = 1$ from $\mathbf{W}^0$ (c) $Q = 1$ from $\mathbf{W}$ (d) $C = 1$ from $\mathbf{W}$

Figure 4. **Visualization of the matches**: when (a) $Q = 1$, and (b) $C = 1$ computed from randomly initialized parameters (denoted by $\mathbf{W}^0$), and when (c) $Q = 1$, and (d) $C = 1$ computed from learned parameters through our method (denoted by $\mathbf{W}$).



(a) Q from $\mathbf{W}^0$ (b) T from $\mathbf{W}^0$ (c) Q from $\mathbf{W}$ (d) T from $\mathbf{W}$

Figure 5. **Visualization of the confidence scores in matrix Q and T**: (a),(b) with randomly initialized parameters (denoted by $\mathbf{W}^0$), and (c),(d) with learned parameters (denoted by $\mathbf{W}$). Due to the space limit, we visualized a partial matrix (a quarter of original).

4.3. Mining Negative Correspondence

Semi-hard negatives The employment of negative samples for contrasting positive ones has shown improved results in representation learning literature [53, 37]. However, naively using too hard or too easy negative samples may degrade the contribution of moderate ones, yielding worse representation [46]. To alleviate this, we collect semi-hard negative samples by specifying the lower and upper boundaries of their hardness. Concretely, given a query pixel u and its positive correspondence $(u, v) \in \mathbf{p}$, negative samples $\mathbf{n}$ are collected with two thresholds $m_1, m_2 \in [0, 1]$ as

$$\mathbf{n}^k(u, v) = \{(u, q) | {m_1}^k < \text{rank}(S^k_{uq}) < {m_2}^k, q \neq v\}, \tag{8}$$

where $\text{rank}(S_{uv}) \in [0, 1]$ operation returns a normalized rank of similarity score S_{uv} sorted in descending order.

Dynamic Curriculum As the degree of hardness is relatively defined with respect to the current capability of the embedding networks, it is necessary to determine appropriate hardness while the training progresses. However, in the unsupervised setting, it is nontrivial to assess the representation capability due to the lack of ground-truth annotation.

We address this by measuring the discriminability of the representation with a spatial variance of confidence score distribution, *i.e.* how distinctively a positive correspondence is established among all possible matching candidates. Accordingly, the hardness can be determined to be inversely proportional to the variance of given positive correspondence:

$$m^k \propto 1 / \sum\nolimits_{(u,v)\in\mathbf{p}^k} \text{var}_v(T^k_{uv}), \tag{9}$$

where var_v operator computes the variance over the spatial coordinates j with respect to a position v, and can be efficiently implemented in parallel using the algorithm of [47].

Algorithm 1: Training procedure of the proposed method

Input: images $\{I^t, ..., I^{t+K}\}$, **Output**: network parameter $\mathbf{W}$

Extract features $\{F^t, ..., F^{t+K}\}$
for $k = 1 : K$ **do**
  Compute pairwise similarity scores S^k
  Compute Q^k by checking consistency
  Compute T^k by solving optimal transport problem
  $*$ ***Sinkhorn algorithm*** $*$
    Initialize $\mathbf{a} = \mathbf{1}n^{-1}$, $\mathbf{U} = \exp(-(1 - Q^k)/\epsilon)$
    for $l = 1 : l_{\max}$ **do**
      $\mathbf{b} = \mathbf{1}n^{-1}/(\mathbf{U}\mathbf{a})$
      $\mathbf{a} = \mathbf{1}n^{-1}/(\mathbf{U}^T\mathbf{b})$
    end for
    T^k=diag($\mathbf{a}$)$\mathbf{U}$diag($\mathbf{b}$)
  Compute C^k with window kernel of radius w^k
  Collect positive set $\mathbf{p}^k$ from C^k
  Set boundary m^k_1 following curriculum
  Collect negative set $\mathbf{n}^k$ from S^k with m^k_1,m_2
end for
Compute gradients by minimizing Equ. (10)

As exemplified in Fig. 5 (b), the distribution of confidence scores in early training is sparsely dispersed with the high variance, thus we provide less hard negatives by setting lower thresholds. On the contray when the variance is small later in training, *e.g.* Fig. 5 (d), a higher threshold is assigned to encourage the model to overcome more difficult examples. Note that the intermediate confidence scores T are utilized here since discarding the confidence scores outside of the local window in Equ. (6) may also remove the information needed to compute the current representation power. In practice, we fixed the upper threshod m_2 during training and adaptively controlled the lower one m_1 following the curriculum.

As shown in Tab. 1, previous works paid little attention on learning with curriculum. Though [29] used the scheduled sampling strategy [2], their curriculum is fixed during training. Some of them [52, 20] attempted to learn from multiple cycles as shorter ones may ease learning, but the number of cycle is not guided by the curriculum.

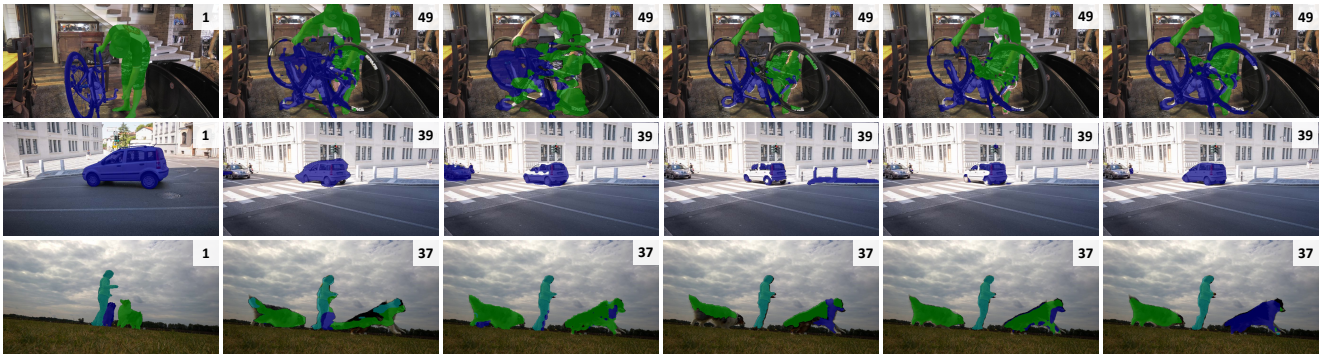
4.4. Contrastive Learning

Finally, we minimize the following objective:

$$\mathcal{L} = \sum_k \sum_{(i,j)\in\mathbf{p}^k} -\log \frac{\exp(S^k_{ij}/\tau)}{\sum_l \exp(S^k_{\mathbf{n}(i,j)}/\tau) + \exp(S^k_{ij}/\tau)}, \tag{10}$$

Methods	Backbone	Supervised	Dataset (Size)	$\mathcal{J}\&\mathcal{F}_{\text{mean}}$	$\mathcal{J}_{\text{mean}}$	$\mathcal{J}_{\text{recall}}$	$\mathcal{F}_{\text{mean}}$	$\mathcal{F}_{\text{recall}}$
Colorization† [51]	ResNet-18	✗	Kinetics (800 hours)	34.0	34.6	34.1	32.7	26.8
CorrFlow† [29]	ResNet-18	✗	OxUvA (14 hours)	50.3	48.4	53.2	52.2	56.0
MAST† [28]	ResNet-18	✗	YT-VOS (5.58 hours)	65.5	63.3	73.2	67.6	77.7
TimeCycle [52]	ResNet-50	✗	VLOG (344 hours)	48.7	46.4	50.0	50.0	48.0
UVC [30]	ResNet-18	✗	Kinetics (800 hours)	60.9	59.3	68.8	62.7	70.9
CRW [20]	ResNet-18	✗	Kinetics (800 hours)	67.6	64.8	76.1	70.2	82.1
Ours	ResNet-18	✗	YT-VOS (5.58 hours)	**70.3**	**67.9**	**78.2**	**72.6**	**83.7**
ResNet [19]	ResNet-18	✓	I (1.28M, 0)	62.9	60.6	69.9	65.2	73.8
OSVOS [6]	VGG-16	✓	I/D (1.28M, 10k)	60.3	56.6	63.8	63.9	73.8
FEELVOS [50]	Xception-65	✓	I/C/D/YT-VOS (1.28M, 663k)	71.5	69.1	79.1	74.0	83.8
STM [36]	ResNet-50	✓	I/D/YT-VOS (1.28M, 164k)	81.8	79.2	-	84.3	-

Table 2. **Quantitative results for video object segmentation on DAVIS-2017 validation set.** For the datasets, we denote as I=ImageNet, C=COCO, D=DAVIS, P=PASCAL-VOC, and a tuple by the numbers of image-level and pixel-level annotations. The methods denoted by † use its own label propagation algorithm. Result of [19, 30, 52] is borrowed from [20].



(a) first frame (b) TimeCycle [52] (c) UVC [30] (d) MAST [28] (e) CRW [20] (f) Ours

Figure 6. **Qualitative results for video object segmentation on DAVIS-2017 validation set**: (a) first frame with given annotation, and propagated label using correspondences obtained from (b) TimeCycle [52], (c) UVC [30], (d) MAST [28], (e) CRW [20], and (f) our method. The number at the right upper corner of each image indexes corresponding frame order.

where l denotes a number of collected negative samples and τ is a temperature. Unlike [51, 52, 30, 29, 28, 20], our loss computes the gradients only at the collected positive correspondences $\mathbf{p}$ thereby preventing the distraction of ambiguous matches. Furthermore, we avoid the gradient vanishing problem due to too easy negative ones by adjusting the hardness of negative samples $\mathbf{n}$. Alg. 1 summarizes the overall training procedure of our method.

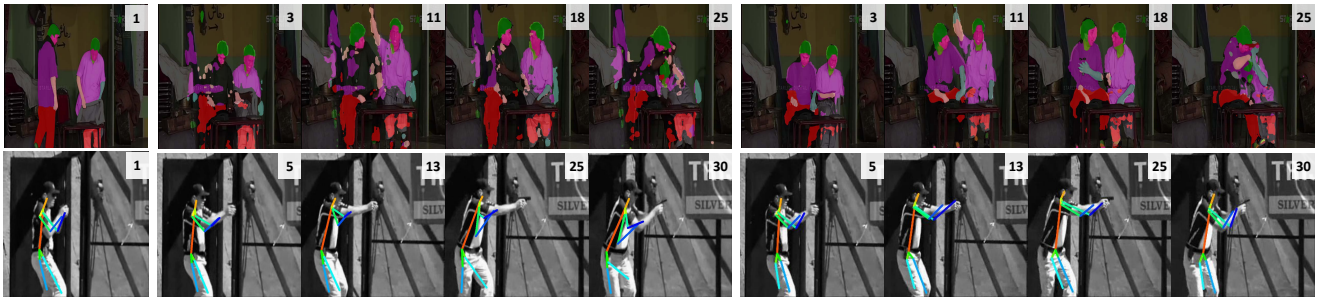
4.5. Implementation Details

As our backbone, we adopt the ResNet-18 network architecture [19] modified to increase the spatial resolution of the convolutional feature map by a factor of four, *i.e.* downsampling factor of 1/8. The parameters of sinkhorn algorithm [11] are set following [32], such that a weighting term of additional entropy regularization ϵ to 0.05 and max iteration $l_{\max}$ to 30. The temperature τ is set to 0.03. The number of interval K is determined as 5 according to the ablation study in Sec. 5.3. A set of window radii w^k are dilated with respect to the temporal length k such that $\{2, 2, 3, 5, 5\}$. For training data, raw video sequences from YouTube-VOS [56] training set are utilized, which contains $3,471$ videos for 94 different object categories with the total length of 5.58 hours. The input images are all resized into 256×256, and the resulting feature maps have a size of 32×32. We train our model using Adam optimizer [27] for 1M iterations with 12 sequences per batch and a learning rate of 10^{-4}. The boundaries for negative mining are initially set to $\{m_1, m_2\} = \{0, 0.9\}$ when the training starts, and the maximum threshold for m_1 is set to 0.8.

5. Experiments

5.1. Experimental Settings

The evaluation of the learned representation is conducted on video label propagation tasks; Given the ground-truth annotation at the first frame, labels are propagated to the rest of the frames using the representation of our model to compute dense correspondences. Due to the rapid progress in this research line, the label propagation algorithm of the state-of-the art methods is not standardized. For a fair comparison, we simply follow the same label propagation algorithm of the best approach for each evaluation task. Generally, the algorithms average the inferences from additional spatial and temporal context in video to obtain the final propagated label. The details of each algorithm are described in the supplemental material, including more qualitative results and performance analyses of our method.



(a) first frame (b) UVC [30] (c) Ours

Figure 7. **Qualitative results for part segmentation and pose tracking on VIP (top) and JHMDB (bottom) validation set, respectively**: (a) first frame with given annotation, and propagated label using correspondences obtained from (b) UVC [30], and (c) our method.

Methods	Sup.	Overall	Seen		Unseen	
			$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Colorization [51]	✗	38.9	43.1	38.6	36.6	37.4
CorrFlow [29]	✗	46.6	50.6	46.6	43.8	45.6
MAST [28]	✗	64.2	63.9	64.9	60.3	67.7
Ours	✗	**67.3**	**66.2**	**67.9**	**63.2**	**71.7**
OSVOS [6]	✓	58.8	59.8	60.5	54.2	60.7
PreMVOS [33]	✓	66.9	71.4	75.9	56.5	63.7
STM [36]	✓	79.4	79.7	84.2	72.8	80.9

Table 3. **Quantitative results for video object segmentation on Youtube-VOS validation set**. Following the protocol of [56], we categorize the performances into "Seen" and "Unseen" classes.

Methods	Sup.	VIP [58]		JHMDB [22]	
		mIoU	AP	$\alpha = 0.1$	$\alpha = 0.2$
TimeCycle [52]	✗	28.9	15.6	57.3	78.1
UVC [30]	✗	34.1	17.7	58.6	79.8
CRW [20]	✗	36.0	-	59.0	**83.2**
Ours	✗	**37.8**	**19.1**	**60.5**	82.3
ResNet-18 [19]	✓	31.8	12.6	53.8	74.6
ATEN [58]	✓	37.9	24.1	-	-
TSN [48]	✓	-	-	68.7	92.1

Table 4. **Quantitative results for part segmentation and pose tracking on VIP and JHMDB validation set, respectively.** We denote α by the used threshold for PCK values.

5.2. Results

Video object segmentation We first evaluate our model on two widely-used datasets for video object segmentation task, DAVIS-2017 [40] and Youtube-VOS [56]. The performances are reported with two standard metrics, namely region overlapping ($\mathcal{J}$) and contour accuracy ($\mathcal{F}$).

For the evaluation on DAVIS-2017 [40] validation set, we use the label propagation algorithm of [20] and resize the input images into 480×480 resolution. As summarized in Tab. 2, despite of using smaller training dataset, the proposed model clearly outperforms all other self-supervised methods, exhibiting even competitive performances to the fully-supervised techniques. This indicates that collecting informative positive and negative samples for contrastive learning greatly effects the quality of the representations. The qualitative results in Fig. 6 also demonstrate that the representation from our model can effectively deal with the ambiguities between temporally distant frames such as large illumination change, complex object deformation and motion-blurred region.

We also examine our method on Youtube-VOS [56] validation set by hiring the label propagation algorithm of [28]. Compared to the methods [51, 29, 28] that only consider reconstruction targets to learn, a large gain reported in Tab. 3 confirms that employing hard negative samples for contrastive learning enables us to learn stronger representation.

Part segmentation and pose tracking We also evaluated our model on the validation set of video instance parsing (VIP) dataset [58] for semantic human part segmentation, and JHMDB benchmark [22] for human keypoints tracking. Compared to the other datasets described above, these benchmarks [58, 22] enable us to validate more precise correspondence. We follow the evaluation protocol of [52, 30], resizing images into 560×560 for part segmentation, and 320×320 for pose tracking. For the evaluation metrics, we use the mean intersection-over-union (IoU) and mean average precision (AP) to measure instance-level human parsing, and probability of correct keypoint (PCK) metric to measure the accuracy between tracked keypoints and the ground-truth one with a threshold α.

Fig. 7 shows that our method can localize fine-grained details of the object in both semantic part segmentation and keypoint tracking tasks. In particular, as reported in Tab. 4, our results show better performances in terms IoU, AP and PCK at $\alpha = 0.1$ metrics compared to the method [20] that consider all possible negative samples during training. This reveals that adjusting hardness of negative samples with respect to the current capability of representation can help guide a representation to be more discriminative.

5.3. Ablation study

To examine the effects of our components, we conduct a series of ablation studies on the validation set of DAVIS-2017 [40] by giving variety to one component at a time.

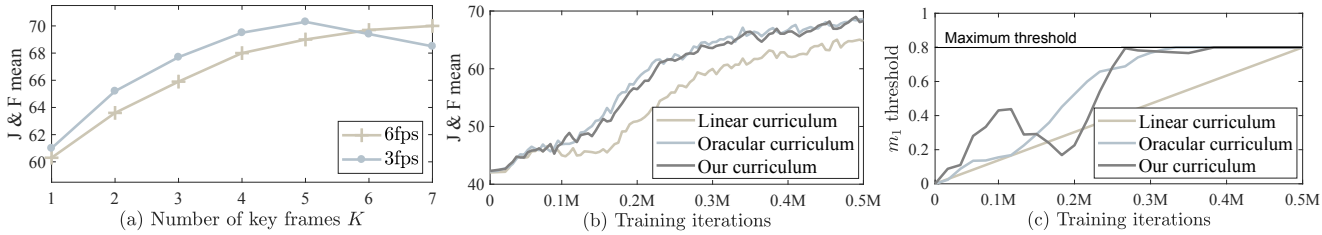


Figure 8. **Convergence analysis**: (a) the performance with respect to various numbers of key frames K, and (b) with respect to various curricula used in training. The convergence of threshold m_1 is also visualized in (c).

Consistency $(S \rightarrow Q)$	Optimal transport $(Q \rightarrow T)$	Coherence $(T \rightarrow C)$	$\mathcal{J}\&\mathcal{F}_{\text{mean}}$
✓	✓	✗	68.7
✓	✗	✓	65.1
✗	✓	✓	63.7
✓	✓	✓	**70.3**

(a) uncertainty criteria

p	**n**	m_1,m_2	Curriculum	$\mathcal{J}\&\mathcal{F}_{\text{mean}}$
✓	✗	✗	-	64.6
✓	✓	✗	-	66.4
✓	✓	✓	✗	67.6
✓	✓	✓	linear	68.9
✓	✓	✓	dynamic	**70.3**

(b) curricular negative mining

Table 5. **Ablation study on DAVIS-2017 validation set**: for different components in (a) uncertainty criteria and (b) curricular negative mining. Note that, in (a), the input of the removed criterion is directly provided as an input for the following criterion. In (b), when only positive samples **p** are used during training, we utilize a reconstruction-based loss function (Equ. (2)) with the attribute of "feature space".

Interval length To study the effect of using different interval of time, we evaluate our model with varying the number of key frames K from 1 to 7 and the frame-rate from 6 fps to 3 fps. As shown in Fig. 8 (a), the measurement of matching uncertainty allows us to learn from the frames that are temporally distant. As the performance converges within $4-6$ lengths, we choose to use 5 frames with 3 fps rate due to its efficiency and optimal performance.

Uncertainty criteria We also report the quantitative assessment when one of criteria is removed from our bottom-up pipeline in Tab. 5 (a). As shown in 3^{rd} row, directly applying optimal transport problem to the similarity scores similar to [32, 42, 45] degrades the performance due to the unguranteed matching costs in the self-supervised setting. However, our bottom-up formulation enables us to yield reliable confidence scores in the absence of ground-truth annotation, highlighting the importance of incorporating those constraints in a unified fashion.

Curricular negative mining To validate the effectiveness of our negative mining strategy, four different baselines are considered here; 1) learning only with positive correspondences **p** by minimizing Equ. (2), learning in the contrastive setting 2) with all possible negative samples similar to [20], with semi-hard negatives **n** collected 3) by fixing m_1 and m_2, and 4) by linearly increasing the lower threshold m_1 to the upper one m_2. As shown in Tab. 5 (b), the proposed curricular negative mining method leads to substantial performance gain over these baseline approaches.

We also validate the assumption of our dynamic curriculum that the current representation power can be measured by the distinctiveness of positive correspondence. For this end, we additionally suppose there exists an oracular curriculum with the access to the annotation of validation set, thereby the actual capacity of embedding networks can be monitored. Accordingly, the oracular curriculum adaptively increases lower threshold m_1 with respect to the performance on a validation set. Fig. 8 (b) compares the accuracies over $0.5M$ iterations when training with the proposed curriculum, including two baseline curricula. We find that the oracular and proposed curriculum converge to roughly the same accuracy, while the linear curriculum appears to converge more slowly to the lower accuracy. From Fig. 8 (c), we observe that the proposed method naturally pushes the threshold up to 0.8 (the highest allowed threshold) around $0.4M$ iterations. These phenomena confirm our two conjectures that 1) assigning a higher threshold later in training is desirable as the model is encouraged to overcome more difficult examples and 2) the evaluation of the representation can be replaced by measuring its discriminability.

6. Conclusion

We presented a novel self-supervised framework for learning pixel-level representation from only unlabeled videos. Our approach is able to collect well-defined positive correspondences by measuring their uncertainties and well-defined negative ones by controlling their hardness during training. The outstanding performance was validated through extensive experiments on various label propagation tasks.

Acknowledgements : This work was supported by IITP grant funded by the Korea government (MSIT) (No.2020-0-00056, To create AI systems that act appropriately and effectively in novel situations that occur in open worlds) and the Yonsei University Research Fund of 2021 (2021-22-0001).

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *In: ICML*, 2017. 2

[2] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179, 2015. 5

[3] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 447–463, 2018. 2

[4] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. *In: ECCV*, 2012. 1

[5] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on robotics*, 32(6):1309–1332, 2016. 1

[6] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 221–230, 2017. 6, 7

[7] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018. 2

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *ICML*, 2020. 2

[9] C. B. Choy, Y. Gwak, and S. Savarese. Universal correspondence network. *In:NIPS*, 2016. 2

[10] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016. 2

[11] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013. 4, 6

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[13] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In *Advances in Neural Information Processing Systems*, pages 10542–10552, 2019. 2

[14] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in neural information processing systems*, pages 766–774, 2014. 2

[15] Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR. 2

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[17] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg. Matchnet: Unifying feature and metric learning for patch-based matching. *In: CVPR*, 2015. 2

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7

[20] Allan Jabri, Andrew Owens, and Alexei A Efros. Space-time correspondence as a contrastive random walk. *Advances in Neural Information Processing Systems*, 2020. 1, 2, 3, 5, 6, 7, 8

[21] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[22] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199, 2013. 7

[23] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Forward-backward error: Automatic detection of tracking failures. In *2010 20th International Conference on Pattern Recognition*, pages 2756–2759. IEEE, 2010. 3, 4

[24] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017. 3

[25] Seungryong Kim, Stephen Lin, Sangryul Jeon, Dongbo Min, and Kwanghoon Sohn. Recurrent transformer networks for semantic correspondence. In *Advances in Neural Information Processing Systems*, 2018. 2

[26] S. Kim, D. Min, B. Ham, S. Ryu, M. N. Do, and K. Sohn. Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. *In: CVPR*, 2015. 1

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[28] Zihang Lai, Erika Lu, and Weidi Xie. Mast: A memory-augmented self-supervised tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2020. 1, 2, 3, 6, 7

[29] Z. Lai and W. Xie. Self-supervised learning for video correspondence flow. In *BMVC*, 2019. 1, 2, 3, 5, 6, 7

[30] Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. In *Advances in Neural Information Processing Systems*, pages 318–328, 2019. 1, 2, 3, 6, 7

[31] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):978–994, 2010. 1

[32] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4463–4472, 2020. 2, 4, 6, 8

[33] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *Asian Conference on Computer Vision*, pages 565–580. Springer, 2018. 7

[34] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 1

[35] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9359–9367, 2018. 2

[36] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9226–9235, 2019. 6, 7

[37] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 5

[38] Pan Pan, Fatih Porikli, and Dan Schonfeld. Recurrent tracking using multifold consistency. 2009. 3, 4

[39] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[40] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016. 7

[41] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 3

[42] Gilles Puy, Alexandre Boulch, and Renaud Marlet. Flot: Scene flow on point clouds guided by optimal transport. In *European Conference on Computer Vision*. 2, 4, 8

[43] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Convolutional neural network architecture for geometric matching. *In: CVPR*, 2017. 1

[44] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *Advances in Neural Information Processing Systems*, pages 1658–1669, 2018. 4

[45] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4938–4947, 2020. 4, 8

[46] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2, 5

[47] Erich Schubert and Michael Gertz. Numerically stable parallel computation of (co-) variance. In *Proceedings of the 30th International Conference on Scientific and Statistical Database Management*, pages 1–12, 2018. 5

[48] Jie Song, Limin Wang, Luc Van Gool, and Otmar Hilliges. Thin-slicing network: A deep structured model for pose estimation in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4220–4229, 2017. 7

[49] Zhengyu Su, Yalin Wang, Rui Shi, Wei Zeng, Jian Sun, Feng Luo, and Xianfeng Gu. Optimal mass transport for shape matching and comparison. *IEEE transactions on pattern analysis and machine intelligence*, 37(11):2246–2259, 2015. 2

[50] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9481–9490, 2019. 6

[51] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 391–408, 2018. 1, 2, 3, 6, 7

[52] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019. 1, 2, 3, 5, 6, 7

[53] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2840–2848, 2017. 2, 5

[54] Hongteng Xu, Dixin Luo, and Lawrence Carin. Scalable gromov-wasserstein learning for graph partitioning and matching. In *Advances in neural information processing systems*, pages 3052–3062, 2019. 2

[55] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin. Gromov-wasserstein learning for graph matching and node embedding. *In:ICML*, 2019. 2

[56] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 6, 7

[57] J. Zbontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. *JMLR*, 17(1):2287–2318, 2016. 2

[58] Qixian Zhou, Xiaodan Liang, Ke Gong, and Liang Lin. Adaptive temporal encoding network for video instance-level human parsing. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1527–1535, 2018. 7