Joint Self-Supervised Video Alignment and Action Segmentation

Ali Shah Ali† Syed Ahmed Mahmood† Mubin Saeed Andrey Konin
M. Zeeshan Zia Quoc-Huy Tran

Retrocausal, Inc., Redmond, WA
www.retrocausal.ai

Abstract

We introduce a novel approach for simultaneous self-supervised video alignment and action segmentation based on a unified optimal transport framework. In particular, we first tackle self-supervised video alignment by developing a fused Gromov-Wasserstein optimal transport formulation with a structural prior, which trains efficiently on GPUs and needs only a few iterations for solving the optimal transport problem. Our single-task method achieves the state-of-the-art performance on multiple video alignment benchmarks and outperforms VAVA, which relies on a traditional Kantorovich optimal transport formulation with an optimality prior. Furthermore, we extend our approach by proposing a unified optimal transport framework for joint self-supervised video alignment and action segmentation, which requires training and storing a single model and saves both time and memory consumption as compared to two different single-task models. Extensive evaluations on several video alignment and action segmentation datasets demonstrate that our multi-task method achieves comparable video alignment yet superior action segmentation results over previous methods in video alignment and action segmentation respectively. Finally, to the best of our knowledge, this is the first work to unify video alignment and action segmentation into a single model.

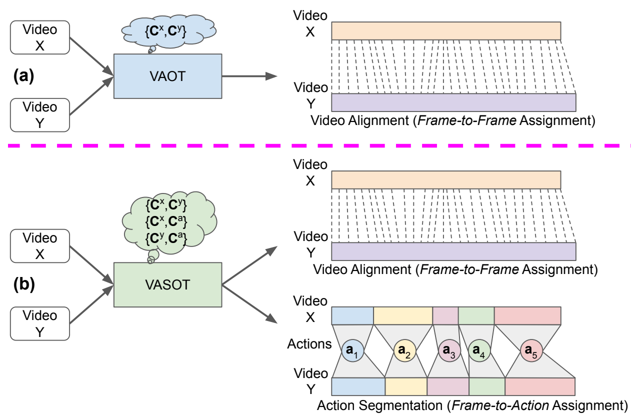


Figure 1. (a) Our self-supervised video alignment method (VAOT) based on a fused Gromov-Wasserstein optimal transport with structural priors $\{\mathbf{C}^x, \mathbf{C}^y\}$. (b) Our joint self-supervised video alignment and action segmentation method (VASOT) based on a unified optimal transport with structural priors $\{\mathbf{C}^x, \mathbf{C}^y\}$ for video alignment and $\{\mathbf{C}^x, \mathbf{C}^a\}$ and $\{\mathbf{C}^y, \mathbf{C}^a\}$ for action segmentation.

1. Introduction

Though the past decade has witnessed remarkable progress in human activity understanding in videos, the majority of the research efforts have been invested in action recognition [8, 20, 67, 74], which categorizes simple actions in short videos. In this paper, we study the two less-explored problems, i.e., temporal video alignment (*frame-to-frame* assignment), which establishes framewise correspondences between long videos recording a complex activity, and temporal action segmentation (*frame-to-action* assignment), which assigns frames of long videos capturing a multi-phase activity to phase/action labels. Since acquiring per-frame annotations for supervised training is generally difficult and costly, we are interested in self-supervised approaches for video alignment and action segmentation.

One popular group of self-supervised video alignment methods rely on global alignment techniques widely used in time series literature. For example, LAV [25] utilizes dynamic time warping [12] by assuming monotonic orderings and no background/redundant frames. VAVA [44] relaxes the above assumptions by incorporating an optimality prior into a standard Kantorovich optimal transport framework [11], along with an inter-video contrastive term and an intra-video contrastive term. However, it is challenging to balance multiple losses as well as handle repeated actions. Similarly, self-supervised action segmentation meth-

† indicates joint first author.
{alishah,ahmed,mubin,andrey,zeeshan,huy}@retrocausal.ai.

ods based on optimal transport have been introduced, including TOT [34] and UFSA [69] which may suffer in cases of order variations, unbalanced segmentation, and repeated actions. ASOT [78] addresses these drawbacks via a fused Gromov-Wasserstein optimal transport framework with a structural prior, which outperforms previous works in self-supervised action segmentation. Lastly, though both self-supervised video align and action segmentation require fine-grained temporal understanding of videos, their interaction in a multi-task learning setup has not been explored.

Motivated by the above observations, we first propose *VAOT* (see Fig. 1(a)), a novel self-supervised video alignment approach based on a fused Gromov-Wasserstein optimal transport formulation with a structural prior, which tackles order variations, background/redundant frames, and repeated actions in a single global alignment framework. Our single-task model trains efficiently on GPUs and requires few iterations to derive the optimal transport solution, while outperforming previous methods, including VAVA [44], on video alignment datasets. Moreover, we develop *VASOT* (see Fig. 1(b)), a joint self-supervised video alignment and action segmentation approach by exploring the relationship between self-supervised video alignment and action segmentation via a unified optimal transport framework. Our multi-task model performs on par with prior works on video alignment benchmarks yet establishes the new state of the art on action segmentation benchmarks. In addition, our joint model requires training and storing a single model and saves both time and memory consumption as compared to two separate single-task models. Lastly, we observe in Sec. 4.3 that, in a multi-task learning setting, action segmentation provides little boost to video alignment results, whereas video alignment increases action segmentation performance significantly.

In summary, our contributions include:

- We propose a fused Gromov-Wasserstein optimal transport formulation with a structural prior for self-supervised video alignment, outperforming previous methods. Our single-task method learns efficiently on GPUs, needing few iterations to obtain the optimal transport solution.
- We develop a unified optimal transport-based approach for simultaneous self-supervised video alignment and action segmentation, yielding comparable video alignment but superior action segmentation results over previous methods. Our joint approach requires training and storing a single model, saving both time and memory usage.
- We conduct extensive experiments on several video alignment and action segmentation datasets, i.e., Pouring, Penn Action, IKEA ASM, 50 Salads, YouTube Instructions, Breakfast, and Desktop Assembly, to validate the advantages of our single-task and multi-task methods. To our best knowledge, our work is the first to combine video alignment with action segmentation.

2. Related Work

Self-Supervised Learning. Early self-supervised learning methods focus on designing image-based pretext tasks with pseudo-labels as supervision signals for learning representations such as image colorization [36, 37], object counting [45, 49], predicting rotations [23], solving puzzles [5, 30], image inpainting [27], and image clustering [6, 7]. The above image-based approaches mostly extract spatial cues from the image content. Recently, significant efforts have been invested in video-based self-supervised learning methods, which exploit both spatial and temporal information. Examples of video-based pretext tasks include forecasting future frames [1, 15, 59, 72], enforcing temporal coherence [24, 48, 82], predicting temporal order [22, 38, 47, 76], arrow of time [52, 75], and pace [4, 73, 79], and utilizing contrastive learning [13, 21, 26, 53]. More recently, skeleton-based self-supervised learning methods with skeleton-based pretext tasks, e.g., skeleton inpainting [81], motion prediction [61], skeleton sequence alignment [35, 68], and utilizing neighborhood consistency [57], motion continuity [62], and multiple pretext tasks [43], have been introduced. These skeleton-based approaches may suffer from human pose estimation errors and missing context details. Here, we leverage video alignment and/or action segmentation as our video-based pretext tasks.

Video Alignment. Self-supervised video alignment has attracted a great amount of research interest in recent years. TCC [18] enforces cycle consistencies between corresponding frames across videos for learning representations. Recently, GTCC [17] extends TCC [18] by proposing multi-cycle consistencies for tackling repeated actions. Both TCC [18] and GTCC [17] perform local alignment by aligning each frame separately. Motivated by global alignment techniques for time series, methods which align the video as a whole have been introduced. LAV [25] which assumes monotonic orderings and no background/redundant frames relies on dynamic time warping [12]. To handle non-monotonic orderings and background/redundant frames, VAVA [44] employs a traditional Kantorovich optimal transport formulation [11] with an optimality prior. In this work, we propose a fused Gromov-Wasserstein optimal transport formulation with a structural prior which handles order variations, background/redundant frames, and repeated actions in a single global alignment framework. In addition, we develop a unified optimal transport-based approach for joint video alignment and action segmentation.

Action Segmentation. Initial works in self-supervised action segmentation perform representation learning and offline clustering as disjoint steps. Please see Ding et al. [16] for a recent survey. CTE [33] trains a temporal embedding first and then employs K-Means to cluster the embedded representations. To enhance CTE [33], VTE [71] and ASAL [42] introduce a visual embedding and an action embedding respectively. The above methods separate representation learning and offline clustering, prohibiting effective communications between the two modules. Recently, methods which jointly conduct representation learning and online clustering have been developed, e.g., UDE [63], TOT [34], and UFSA [69]. However, their performance may deteriorate in cases of order variations, unbalanced segmentation, and repeated actions. To overcome these limitations, ASOT [78] introduces a fused Gromov-Wasserstein optimal transport formulation with a structural prior. Here, we propose a similar optimal transport formulation for video alignment, outperforming previous methods on video alignment datasets, and extend it to a joint video alignment and action segmentation model, establishing the new state of the art on action segmentation benchmarks.

Optimal Transport with Structured Data. Optimal transport underpins several computer vision and machine learning applications. A comprehensive review of optimal transport and its applications in machine learning is presented in Khamis et al. [29]. Optimal transport applications in computer vision include keypoint matching [46, 54], point set registration [56], object detection [14], object tracking [39], video alignment [44], and procedure learning [10]. For problems with structured data, a Gromov-Wasserstein optimal transport with a structural prior is frequently employed, e.g., graph matching [77], brain image registration [65], and action segmentation [78]. In this paper, we develop a Gromov-Wasserstein optimal transport with a structural prior for video alignment and extend it to joint video alignment and action segmentation. To our best knowledge, our work is the first to incorporate video alignment and action segmentation into a unified optimal transport framework.

3. Our Approach

We describe in this section our main contributions, namely a self-supervised video alignment approach (VAOT) in Sec. 3.1 and a joint self-supervised video alignment and action segmentation approach (VASOT) in Sec. 3.2.

Notations. First of all, $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i,j} A_{ij} B_{ij}$ denotes the dot product of $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times m}$, $\mathbf{1}_n \in \mathbb{R}^n$ models a vector of ones, and $[n] = \{1, \ldots, n\}$ represents a discrete set of n elements. Next, $\Delta_m \subset \mathbb{R}^m$ denotes the $(m-1)$ dimensional probability simplex, while $\Delta_m^n \subset \mathbb{R}^{m \times n}$ models the Cartesian product space consisting of n such simplexes. Furthermore, $X = \{x_1, \ldots, x_N\}$ and $Y = \{y_1, \ldots, y_M\}$ denote two input videos of N and M frames respectively. Let $f_{\boldsymbol{\theta}}$ with learnable parameters $\boldsymbol{\theta}$ be the embedding function, frame-level embeddings of X and Y are expressed as $\mathbf{X} = f_{\boldsymbol{\theta}}(X) \in \mathbb{R}^{N \times D}$ and $\mathbf{Y} = f_{\boldsymbol{\theta}}(Y) \in \mathbb{R}^{M \times D}$, where D is the embedding vector length. Lastly, K learnable action centroids are expressed by $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_K] \in \mathbb{R}^{D \times K}$.

3.1. Self-Supervised Video Alignment

3.1.1. Optimal Transport with Structured Data

Kantorovich Optimal Transport. We briefly describe the conventional optimal transport formulation, also known as Kantorovich optiomal transport (KOT) [64], in the discrete setting. The KOT problem aims to find the minimum-cost coupling $\mathbf{T}^\star$ between histograms $\mathbf{p} \in \Delta_n$ and $\mathbf{q} \in \Delta_m$ with a ground cost $\mathbf{C} \in \mathbb{R}_+^{n \times m}$ and is written as:

$$\operatorname*{argmin}_{\mathbf{T} \in \mathcal{T}_{\mathbf{p},\mathbf{q}}} \quad \mathcal{F}_{\text{KOT}}(\mathbf{C}, \mathbf{T}) = \langle \mathbf{C}, \mathbf{T} \rangle, \tag{1}$$

with $\mathcal{T}_{\mathbf{p},\mathbf{q}} = \{\mathbf{T} \in \mathbb{R}_+^{n \times m} \mid \mathbf{T}\mathbf{1}_m = \mathbf{p}, \mathbf{T}^\top \mathbf{1}_n = \mathbf{q}\}$. The coupling $\mathbf{T}$ is regarded as the *soft assignment* between elements in the supports of $\mathbf{p}$ and $\mathbf{q}$, i.e., discrete sets $[n]$ and $[m]$. For video alignment, $\mathbf{T} \in \mathbb{R}_+^{N \times M}$ represents the assignment between frames of X and Y.

Gromov-Wasserstein Optimal Transport. For histograms defined over incomparable spaces, Gromov-Wasserstein (GW) optimal transport [51] is typically employed as:

$$\operatorname*{argmin}_{\mathbf{T} \in \mathcal{T}_{\mathbf{p},\mathbf{q}}} \quad \mathcal{F}_{\text{GW}}(\mathbf{C}^x, \mathbf{C}^y, \mathbf{T}) = \sum_{\substack{i,k \in [n] \\ j,l \in [m]}} L(\mathbf{C}^x_{ik}, \mathbf{C}^y_{jl}) \mathbf{T}_{ij} \mathbf{T}_{kl}. \tag{2}$$

Here, $(\mathbf{C}^x, \mathbf{p}) \in \mathbb{R}^{n \times n} \times \Delta_n$ and $(\mathbf{C}^y, \mathbf{q}) \in \mathbb{R}^{m \times m} \times \Delta_m$ represent two (metric, measure) pairs respectively, while distance matrices $\mathbf{C}^x$ and $\mathbf{C}^y$ describe metrics defined over supports $[n]$ and $[m]$ respectively. Note that there is no metric defined *between* supports $[n]$ and $[m]$ in the GW setting. $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ denotes a cost function minimizing discrepancies between distance matrix elements. We utilize the GW formulation to impose *structural priors* $\mathbf{C}^x$ and $\mathbf{C}^y$ on the transport map for video alignment (i.e., temporal consistency), which we will describe in the next section.

Fused Gromov-Wasserstein Optimal Transport. Fused Gromov-Wasserstein (FGW) optimal transport [66, 70] which merges KOT and GW formulations is often used for problems with known ground cost and structural prior. Let $\alpha \in [0, 1]$, the FGW problem is expressed as:

$$\operatorname*{argmin}_{\mathbf{T} \in \mathcal{T}_{\mathbf{p},\mathbf{q}}} \quad \mathcal{F}_{\text{FGW}}(\mathbf{C}, \mathbf{C}^x, \mathbf{C}^y, \mathbf{T}) = (1-\alpha)\mathcal{F}_{\text{KOT}}(\mathbf{C}, \mathbf{T}) + \alpha \mathcal{F}_{\text{GW}}(\mathbf{C}^x, \mathbf{C}^y, \mathbf{T}). \tag{3}$$

For video alignment, the KOT objective encourages visual similarity between corresponding frames of X and Y, while the GW objective enforces structural properties on the resulting alignment (i.e., temporal consistency).

Balanced Optimal Transport. The above optimal transport problems impose *balanced assignment* constraints $\mathbf{T} \in \mathcal{T}_{\mathbf{p},\mathbf{q}} = \{\mathbf{T} \in \mathbb{R}_+^{n \times m} \mid \mathbf{T}\mathbf{1}_m = \mathbf{p}, \mathbf{T}^\top \mathbf{1}_n = \mathbf{q}\}$. Recent works have relaxed these constraints by replacing (one [78] or both [65]) marginal constraints on $\mathbf{T}$ with penalty terms in the objective, yielding (*partial* [78] or *full* [65]) *unbalanced assignment* constraints. For video alignment, we adopt the full unbalanced formulation [65] but the results are worse than those of the balanced formulation, as we will show later in Sec. 4.1. This is likely because it is difficult to balance multiple losses (the full unbalanced formulation yields two extra penalty terms) and video alignment is generally more balanced than action segmentation (the number of frames is much larger than the number of actions).

3.1.2. Video Alignment Optimal Transport

Here, we adapt the above balanced FGW optimal transport in Eq. 3 for video alignment, yielding our proposed video alignment optimal transport (VAOT). Let us denote $\mathbf{p} = \frac{1}{N}\mathbf{1}_N$ and $\mathbf{q} = \frac{1}{M}\mathbf{1}_M$ as histograms defined over the sets of N frames in X and M frames in Y, represented by $[N]$ and $[M]$ respectively. The solution $\mathbf{T}^\star \in \mathbb{R}_+^{N\times M}$ between $[N]$ and $[M]$ represents the soft assignment between frames of X and Y. For a frame x_i in X, the corresponding frame $y_{j^\star}$ in Y is specified by $j^\star = \mathrm{argmax}_j\, \mathbf{T}^\star_{ij}$. Below we will discuss our cost matrices $\{\mathbf{C}, \mathbf{C}^x, \mathbf{C}^y\}$ for the FGW problem in Eq. 3, deriving the solution $\mathbf{T}^\star$ efficiently, and handling background/redundant frames.

Visual Cue. The KOT subproblem in Eq. 3 includes the cost matrix $\mathbf{C}$ which measures the difference in visual content of X and Y and is defined as $\mathbf{C}_{ij} = 1 - \frac{\mathbf{x}_i^\top \mathbf{y}_j}{\|\mathbf{x}_i\|_2 \|\mathbf{y}_j\|_2}$, with frame embeddings $\mathbf{x}_i = f_{\boldsymbol{\theta}}(x_i)$ and $\mathbf{y}_j = f_{\boldsymbol{\theta}}(y_j)$.

Structural Prior. For the GW subproblem in Eq. 3, we define $L(a,b) = ab$ and cost matrices $\mathbf{C}^x \in \mathbb{R}_+^{N\times N}$ over frames of X and $\mathbf{C}^y \in \mathbb{R}_+^{M\times M}$ over frames of Y as:

$$\mathbf{C}^x_{ik} = \begin{cases} \frac{1}{r} & 1 \le \delta_{ik} \le Nr \\ 0 & \text{otherwise} \end{cases}, \mathbf{C}^y_{jl} = \begin{cases} 0 & 1 \le \delta_{jl} \le Mr \\ 1 & \text{otherwise} \end{cases}. \tag{4}$$

Here, $\delta_{ik} = |i-k|$, $\delta_{jl} = |j-l|$, and a radius parameter $r \in (0,1]$. The GW component encourages temporal consistency over $\mathbf{T}$. In particular, assigning temporally nearby frames in X ($\delta_{ik} \le Nr$) to temporally distant frames in Y ($\delta_{jl} > Mr$) incurs a cost ($L(\mathbf{C}^x_{ik}, \mathbf{C}^y_{jl}) = \frac{1}{r}$), whereas mapping temporally nearby frames in X ($\delta_{ik} \le Nr$) to temporally adjacent frames in Y ($\delta_{jl} \le Mr$) or mapping temporally distant frames in X ($\delta_{ik} > Nr$) to temporally remote frames in Y ($\delta_{jl} > Mr$) incurs no cost ($L(\mathbf{C}^x_{ik}, \mathbf{C}^y_{jl}) = 0$). The GW component is capable of handling order variations and repeated actions, as shown in ASOT [78].

Fast Numerical Solver for VAOT. The GW component in Eq. 3 can be computed efficiently as $\mathcal{F}_{\text{GW}}(\mathbf{C}^x, \mathbf{C}^y, \mathbf{T}) = \langle \mathbf{C}^x\mathbf{T}\mathbf{C}^y, \mathbf{T}\rangle$ since the cost function $L(a,b) = ab$ can be factorized [51]. In addition, by adding an entropy regularization term $-\epsilon H(\mathbf{T})$, with $H(\mathbf{T}) = -\sum_{i,j} T_{ij} \log T_{ij}$ and $\epsilon > 0$, to the FGW formulation in Eq. 3, we can obtain the solution $\mathbf{T}^\star$ efficiently via projected mirror descent [51], which can be run on GPUs. Our solver often converges in less than 25 iterations. By exploiting the sparse structures of $\mathbf{C}^x$ and $\mathbf{C}^y$, each iteration has $O(NM)$ time complexity.

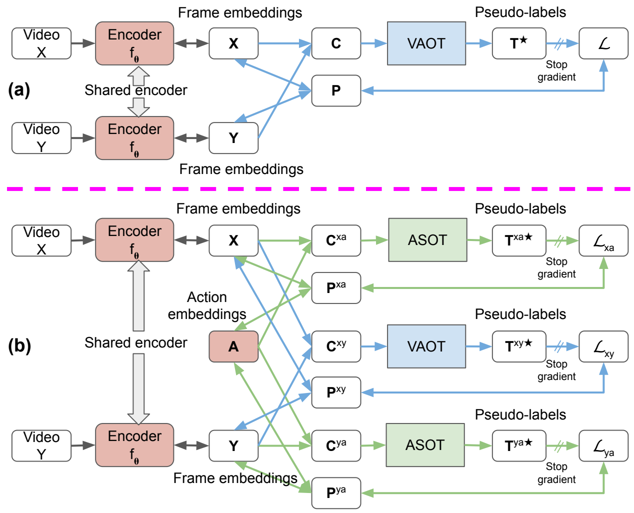


Figure 2. (a) Our self-supervised video alignment method (VAOT). (b) Our joint self-supervised video alignment and action segmentation method (VASOT). Learnable parameters are shown in red. Arrows denote computation/gradient flows (blue and green represent video alignment and action segmentation respectively).

Background/Redundant Frame Handling. To tackle background/redundant frames, we follow VAVA [44] to add a *virtual frame* to X and Y so that background/redundant frames are explicitly assigned to it. Specifically, we append an extra row and column to $\mathbf{T}$ and expand other variables accordingly. If the assignment probability of x_i ($i \le N$) with every y_j ($j \le M$) is smaller than a threshold parameter ζ, we match x_i with the virtual frame y_{M+1}. Similarly, if the assignment probability of y_j ($j \le M$) with every x_i ($i \le N$) is smaller than ζ, we match y_j with x_{N+1}. Note that virtual frames and their associated frames are excluded from computing the losses. As shown in Sec. 4.1, handling background/redundant frames leads to performance gain.

3.1.3. Self-Supervised Learning

We now present our self-supervised learning framework for video alignment in Fig. 2(a). We utilize the above VAOT module to compute pseudo-labels as supervision signals for training the frame encoder. We learn the parameters $\boldsymbol{\theta}$ of the frame encoder by minimizing the cross-entropy loss between normalized similarities $\mathbf{P}$ (computed based on frame embeddings $\mathbf{X}$ and $\mathbf{Y}$) and pseudo-labels $\mathbf{T}^\star$ (obtained by VAOT). We first define normalized similarities $\mathbf{P} \in \Delta_M^N$ as:

$$\mathbf{P}_{ij} = \frac{\exp(\mathbf{X}\mathbf{Y}^\top/\tau)_{ij}}{\sum_l \exp(\mathbf{X}\mathbf{Y}^\top/\tau)_{il}}, \tag{5}$$

with $\tau > 0$ denoting a temperature scaling parameter. In addition, pseudo-labels $\mathbf{T}^\star$ are obtained by solving the balanced FGW problem in Eq. 3 but with the augmented KOT cost matrix $\tilde{\mathbf{C}} = \mathbf{C} + \rho\mathbf{R}$, with $\rho \geq 0$. Here, we follow VAVA [44] to impose the *temporal prior* $\mathbf{R}$, defined as $\mathbf{R}_{ij} = |i/N - j/M|$, which encourages the coupling $\mathbf{T}$ to have a banded diagonal shape and temporally nearby frames in X to be matched with temporally adjacent frames in Y, yielding improved performance as seen in Sec. 4.1. Finally, our self-supervised learning loss is written as:

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{j=1}^{M} \mathbf{T}^{\star}_{ij} \log \mathbf{P}_{ij}. \tag{6}$$

Note we do not back-propagate gradients through $\mathbf{T}^{\star}$.

3.2. Joint Self-Supervised Video Alignment and Action Segmentation

3.2.1. Self-Supervised Video Alignment vs. Action Segmentation

Sec. 3.1 presents our self-supervised video alignment approach (VAOT) developed based on an FGW optimal transport formulation with a structural prior, which has previously been adopted for self-supervised action segmentation by ASOT [78]. Below we discuss differences between self-supervised video alignment and action segmentation, which lead to distinct design choices for VAOT and ASOT [78]. Firstly, as illustrated in Fig. 1, video alignment performs finer-grained *frame-to-frame* assignment, as compared to coarser-grained *frame-to-action* assignment in action segmentation, which causes our removal of $\mathbf{A}$ and our cost matrices $\{\mathbf{C}, \mathbf{C}^x, \mathbf{C}^y\}$ in Sec. 3.1 to be different from ASOT [78]. Secondly, video alignment generally has a more balanced assignment than action segmentation (the number of frames is much larger than the number of actions) and it is hard to balance multiple losses (the full unbalanced formulation adds two extra penalty terms). Thus, a balanced FGW formulation performs the best for VAOT, whereas a partial unbalanced FGW formulation is preferred in ASOT [78]. Thirdly, a background action class, to which background/redundant frames are assigned, is typically included in the K actions for action segmentation, whereas it is not already defined for video alignment. Thus, VAOT adds a virtual frame to tackle background/redundant frames.

3.2.2. Self-Supervised Multi-Task Learning

Since both self-supervised video alignment and action segmentation exploit fine-grained temporal information in videos, we propose a self-supervised multi-task learning framework for joint video alignment and action segmentation. In particular, we combine our VAOT module for video alignment with ASOT [78] for action segmentation into a unified optimal transport-based approach (VASOT), which is illustrated in Fig. 2(b). Here, we update variable names in VAOT from $\{\mathbf{C}, \mathbf{P}, \mathbf{T}^{\star}, \mathcal{L}\}$ to $\{\mathbf{C}^{xy}, \mathbf{P}^{xy}, \mathbf{T}^{xy\star}, \mathcal{L}_{xy}\}$ respectively for video alignment between X and Y, while introducing new variables $\{\mathbf{C}^{xa}, \mathbf{P}^{xa}, \mathbf{T}^{xa\star}, \mathcal{L}_{xa}\}$ and $\{\mathbf{C}^{ya}, \mathbf{P}^{ya}, \mathbf{T}^{ya\star}, \mathcal{L}_{ya}\}$ for action segmentation on X and Y respectively. The parameters $\boldsymbol{\theta}$ of the frame encoder and the action embeddings $\mathbf{A}$ are trained by using the below combination of self-supervised learning losses:

$$\mathcal{L}_{joint} = w_{align}\mathcal{L}_{xy} + w_{seg}(\mathcal{L}_{xa} + \mathcal{L}_{ya}), \tag{7}$$

where $w_{align} \geq 0$ and $w_{seg} \geq 0$ denote the weights for the video alignment loss $\mathcal{L}_{xy}$ and the action segmentation losses $\mathcal{L}_{xa}$ and $\mathcal{L}_{ya}$ respectively, $\mathcal{L}_{xy}$ is the cross-entropy loss between normalized similarities $\mathbf{P}^{xy}$ (computed between $\mathbf{X}$ and $\mathbf{Y}$) and pseudo-labels $\mathbf{T}^{xy\star}$ (obtained by VAOT), while $\mathcal{L}_{xa}$ is the cross-entropy loss between normalized similarities $\mathbf{P}^{xa}$ (computed based on $\mathbf{X}$ and $\mathbf{A}$) and pseudo-labels $\mathbf{T}^{xa\star}$ (derived by ASOT [78]) and $\mathcal{L}_{ya}$ is the cross-entropy loss between normalized similarities $\mathbf{P}^{ya}$ (computed between $\mathbf{Y}$ and $\mathbf{A}$) and pseudo-labels $\mathbf{T}^{ya\star}$ (obtained by ASOT [78]). This is in contrast with VAOT or ASOT [78], where $\boldsymbol{\theta}$ is solely learned by using either $\mathcal{L}_{xy}$ or $\mathcal{L}_{xa}$ and $\mathcal{L}_{ya}$ respectively. We find balancing $w_{align} = w_{seg} = 1$ yields good results for both video alignment and action segmentation, as seen in Sec. 4.2. Our joint model requires training and storing a single model, saving both time and memory usage as compared to two single-task models. As we observe in Sec. 4.3, in a multi-task learning setting, action segmentation offers little benefit to video alignment performance, while video alignment boosts action segmentation results substantially.

4. Experiments

Datasets. We benchmark our VAOT and VASOT approaches for video alignment using *three* datasets, including monotonic datasets, i.e., Pouring [55] and Penn Action [80], and in-the-wild dataset, i.e., IKEA ASM [3]. Pouring includes videos of humans pouring liquids and Penn Action comprises of videos of humans playing sports, while IKEA ASM videos capture humans assembling furniture. All methods have the same training and validation splits. Moreover, for action segmentation evaluation, we use *four* datasets, including in-the-wild datasets, i.e., Breakfast [32], 50 Salads [60], and YouTube Instructions [2], and monotonic dataset, i.e., Desktop Assembly [34]. Breakfast and 50 Salads videos show cooking activities, while YouTube Instructions consists of instructional videos and Desktop Assembly includes videos of an assembly activity. All methods are trained and tested on the same set of videos. For 50 Salads, we evaluate at two action granularity levels, i.e., *Mid* with 19 actions and *Eval* with 12 actions. Finally, for datasets with many activities, i.e., Penn Action, Breakfast, and YouTube Instructions, we train and test the methods per activity and report the average results.

	Method	Acc@0.1	Acc@0.5	Acc@1.0	Progress	τ	AP@5	AP@10	AP@15
IKEA ASM	w/o Structural Prior	<u>30.29</u>	<u>35.52</u>	<u>37.81</u>	-	-	27.54	27.33	27.15
	w/o Temporal Prior	17.84	17.84	17.84	-	-	15.63	15.64	15.56
	w/o Balanced Assignment	17.84	20.71	25.24	-	-	15.49	15.69	15.78
	w/o Virtual Frame	30.16	34.49	36.10	-	-	<u>29.57</u>	<u>29.24</u>	<u>28.87</u>
	All	**33.73**	**36.42**	**38.64**	-	-	**31.49**	**31.92**	**32.01**

Table 1. Ablation analysis results. **Bold** and <u>underline</u> denote the best and second best respectively.

Implementation Details. For fair comparison purposes, our VAOT and VASOT approaches for video alignment utilize the same ResNet-50 encoder as recent self-supervised video alignment methods [17, 18, 25, 44]. Similarly, our VASOT approach for action segmentation employs the same MLP encoder as state-of-the-art self-supervised action segmentation methods [33, 34, 78]. Action embeddings $\mathbf{A}$ are initialized via K-Means, while the number of clusters K is set to the ground truth value. We implement our methods in PyTorch [50] and use ADAM optimization [31]. Please refer to our supplementary material for more details.

Competing Methods. We compare our VAOT and VASOT approaches against prior self-supervised video alignment methods, namely SAL [47], TCN [55], TCC [18], LAV [25], VAVA [44], and GTCC [17]. VAVA [44], which integrates an optimality prior into a classical Kantorovich optimal transport, is the closest to our VAOT approach. Also, we test our VASOT approach against previous self-supervised action segmentation methods, namely CTE [33], VTE [71], UDE [63], ASAL [42], TOT [34], UFSA [69], ASOT [78], and HVQ [58]. ASOT [78] is the single-task baseline for action segmentation, which we adopt in Sec. 3.2 for our multi-task VASOT approach.

Evaluation Metrics. To evaluate our VAOT and VASOT approaches for video alignment, we compute *four* metrics on the validation set, i.e., phase classification (*Acc@*$\{$*0.1,0.5,1.0*$\}$), phase progression (*Progress*), video alignment (τ), and fine-grained frame retrieval (*AP@*$\{$*5,10,15*$\}$). Prior to that, we train the model on the training set and freeze it, and then train an SVM classifier or linear regressor on top of frozen features. Note that as mentioned in [17, 18, 25, 44], Progress and τ are only defined for monotonic datasets and hence are not computed for IKEA ASM. Also, to test our VASOT approach for action segmentation, we calculate *three* metrics, i.e., mean over frames (*MoF*), F1 score (*F1*), and mean intersection over union (*mIoU*). Before that, we train the model, obtain predicted action segments, and perform Hungarian matching between predicted and ground truth action clusters.

4.1. Ablation Analysis Results

We first study the impacts of design choices in VAOT in Sec. 3.1. We show the IKEA ASM results in Tab. 1. Please see our supplementary material for the Pouring results.

Effect of Structural Prior. The structural priors $\{\mathbf{C}^x, \mathbf{C}^y\}$ defined in Eq. 4 are used to encourage temporal consistency on the transport map $\mathbf{T}$. We analyze the impact of the structural priors by removing the GW subproblem in Eq. 3 (via setting $\alpha = 0$), yields worse results, as reported in Tab. 1. This demonstrates the importance of the structural priors and temporal consistency in our VAOT approach.

Effect of Temporal Prior. Removing the temporal prior $\mathbf{R}$ described in Sec. 3.1.3 (by setting $\rho = 0$), leading to notable performance drops in Tab. 1. This validates the contribution of the temporal prior in our VAOT approach.

Effect of Balanced Assignment. When the balanced assignment formulation is replaced by the full unbalanced assignment formulation, the performance degrades significantly, as shown in Tab. 1. This indicates that the balanced assignment formulation is preferred for the video alignment problem, as we discussed previously in Sec. 3.1.1.

Effect of Virtual Frame. Virtual frames (described in Sec. 3.1.2) are used to tackle background/redundant frames. From Tab. 1, removing virtual frames negatively affects the robustness and hence performance of our VAOT approach.

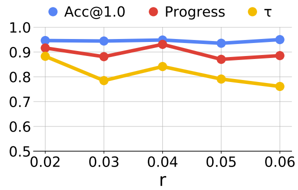
4.2. Sensitivity Analysis Results

Here, we conduct sensitivity analyses on hyperparameters of VAOT and VASOT. Fig. 3 shows the results. We use Pouring in Figs. 3(a-e) and Desktop Assembly in Fig. 3(f). The results of ϵ are provided in our supplementary material.

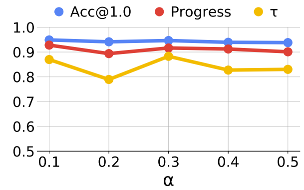
Effect of r and α. From Fig. 3(a), Acc@1.0 remains stable and Progress shows small variation across all studied values of r, whereas τ is the most sensitive metric, peaking at $r = 0.02$ and decreasing as r increases. Similar observations can be made for α in Fig. 3(b), where Acc@1.0 and Progress are mostly stable, whereas τ fluctuates the most, performing the best with $\alpha = 0.3$. Furthermore, we find that $r = 0.02$ and $\alpha = 0.3$ also work the best for the remaining datasets.

Effect of ρ and ζ. For ρ in Fig. 3(c), it can be seen that Acc@1.0 and Progress remain mostly stable, whereas τ varies the most, performing the best at $\rho = 0.35$. Similarly, for ζ in Fig. 3(d), we observe that τ is the most sensitive metric, peaking at $\zeta = 0.5$, whereas Acc@1.0 and Progress remain steady across the analyzed value range of ζ. Moreover, we notice that $\rho = 0.35$ and $\zeta = 0.5$ also yield the best results for the remaining datasets.

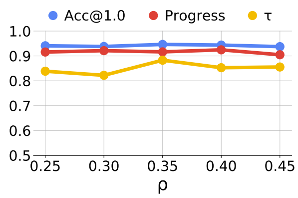
Effect of w_{seg} and w_{align}. We study the relationship



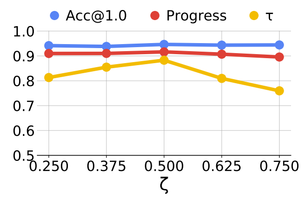
(a) Structural prior radius r.



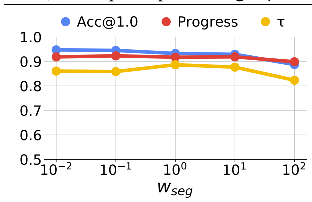
(b) Gromov-Wasserstein weight α.



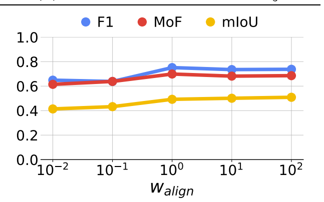
(c) Temporal prior weight ρ.



(d) Virtual frame threshold ζ.



(e) Segmentation loss weight w_{seg} (here, $w_{align} = 1$).



(f) Alignment loss weight w_{align} (here, $w_{seg} = 1$).

Figure 3. Sensitivity analysis results. Note that (a-d) are for VAOT, while (e-f) are for VASOT.

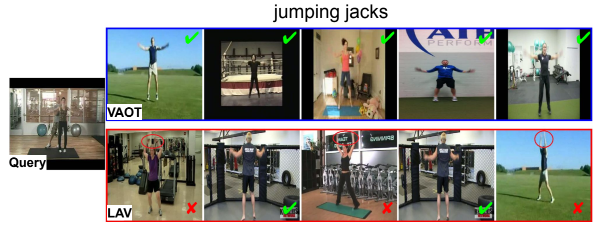


Figure 4. Fine-grained frame retrieval results on Penn Action. The query image is on the left, while on the right are the top 5 matching images retrieved by VAOT (blue box) and LAV (red box).

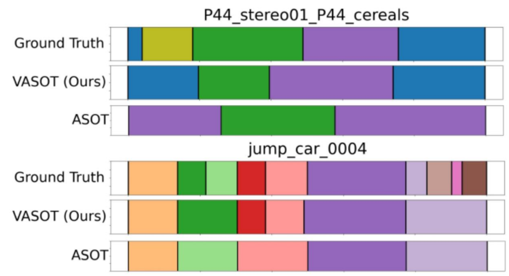


Figure 5. Action segmentation results on Breakfast (top) and YouTube Instructions (bottom).

between action segmentation and video alignment in our multi-task VASOT approach by varying their weights w_{seg} and w_{align} in Eq. 7. In particular, we report alignment results with various w_{seg} values (while keeping $w_{align} = 1$) in Fig. 3(e), and segmentation results with various w_{align} values (while keeping $w_{seg} = 1$) in Fig. 3(f). It can be seen from Fig. 3(e) that alignment results drop as w_{seg} increases and the best overall alignment performance is obtained with $w_{seg} = 1$. In contrast, it is clear from Fig. 3(f) that segmentation results improve as w_{align} increases and become steady after w_{align} reaches 1. Therefore, in a multi-task learning setup, action segmentation provides little boost to video alignment results, whereas video alignment increases action segmentation performance notably. Moreover, with balancing $w_{seg} = w_{align} = 1$, VASOT achieves good results for both video alignment and action segmentation. For the next section, we set $w_{seg} = w_{align} = 1$ for VASOT.

4.3. State-of-the-Art Comparison Results

Video Alignment Comparison Results. We now benchmark our VAOT and VASOT approaches against previous self-supervised video alignment methods and present quantitative results in Tab. 2. Firstly, it is evident from Tab. 2 that our VAOT approach achieves the best overall performance across all datasets, outperforming all competing methods, i.e., SAL [47], TCN [55], TCC [18], LAV [25], VAVA [44], and GTCC [17]. Especially, VAOT shows major improvements over VAVA [44] on the in-the-wild IKEA ASM dataset. The results confirm the advantage of our FGW formulation with a structural prior over the classical Kantorovich formulation with an optimality prior in VAVA [44]. Next, Fig. 4 shows some qualitative results, where VAOT retrieves all 5 correct frames with the same action (*Hands at shoulder*) as the query image, while LAV [25] obtains 3 incorrect frames (*Hands above head*), highlighted by red ovals. Lastly, similar to Fig. 3(e), we find that action segmentation offers little benefit to video alignment in a multi-task learning setup, and our multi-task VASOT approach performs mostly similarly to our single-task VAOT approach in Tab. 2. This is likely because video alignment is a more complex problem involving finer-grained *frame-to-frame* assignment, as compared to coarser-grained *frame-to-action* assignment in action segmentation. Nevertheless, VASOT obtains mostly favorable results over prior works.

Action Segmentation Comparison Results. We test our VASOT approach against state-of-the-art self-supervised action segmentation methods and include quantitative results in Tab. 3. Firstly, from Tab. 3, VASOT consistently achieves the best results across all metrics and datasets, outperforming all competing methods, i.e., CTE [33], VTE [71], UDE [63], ASAL [42], TOT [34], UFSA [69], ASOT [78], and HVQ [58]. While our multi-task VASOT approach shows small gains over the single-task ASOT baseline [78] on Breakfast and YouTube Instructions, our improvements on 50 Salads and Desktop Assembly are sub-

	Method	Acc@0.1	Acc@0.5	Acc@1.0	Progress	τ	AP@5	AP@10	AP@15
Pouring	SAL [47]	85.68	87.84	88.02	74.51	73.31	84.05	83.77	83.79
	TCN [55]	89.19	90.39	90.35	80.57	86.69	83.56	83.31	83.01
	TCC [18]	89.23	91.43	91.82	80.30	85.16	87.16	86.68	86.54
	LAV [25]	91.61	92.82	92.84	80.54	85.61	89.13	89.13	89.22
	VAVA [44]	91.65	91.79	92.45	83.61	87.55	90.05	89.92	90.17
	GTCC [17]	71.20	89.20	93.50	85.80	88.10	-	-	-
	VAOT (Ours)	91.80	**92.88**	**94.63**	91.63	88.28	**91.34**	**90.56**	**90.29**
	VASOT (Ours)	**91.93**	92.12	93.04	**91.71**	**88.64**	91.03	90.44	90.21
Penn Action	SAL [47]	74.87	78.26	79.96	59.43	63.36	76.04	75.77	75.61
	TCN [55]	81.99	83.67	84.04	67.62	73.28	77.84	77.51	77.28
	TCC [18]	81.26	83.35	84.45	67.26	73.53	76.74	76.27	75.88
	LAV [25]	83.56	83.95	84.25	66.13	80.47	79.13	78.98	78.90
	VAVA [44]	83.89	84.23	84.48	70.91	80.53	81.52	80.47	80.67
	GTCC [17]	78.30	81.20	81.30	70.80	88.30	-	-	-
	VAOT (Ours)	83.96	85.35	86.92	**84.31**	**88.99**	**81.62**	**81.03**	**80.68**
	VASOT (Ours)	**84.17**	**85.54**	**87.66**	83.39	88.70	78.85	78.38	78.03
IKEA ASM	SAL [47]	22.94	23.43	25.46	-	-	14.28	14.04	14.10
	TCN [55]	22.51	25.47	25.88	-	-	17.37	17.03	16.96
	TCC [18]	22.70	25.04	25.63	-	-	18.03	17.53	17.20
	LAV [25]	23.19	25.47	25.54	-	-	20.14	19.35	19.21
	VAVA [44]	29.12	29.95	29.10	-	-	26.42	25.73	25.80
	VAOT (Ours)	**33.73**	**36.42**	**38.64**	-	-	**31.49**	**31.92**	**32.01**
	VASOT (Ours)	29.96	30.78	31.02	-	-	30.29	30.37	30.42

Table 2. Video alignment comparison results. **Bold** and underline denote the best and second best respectively.

	Method	Breakfast	YouTube Instructions	50 Salads (Mid)	50 Salads (Eval)	Desktop Assembly
		MoF / F1 / mIoU	MoF / F1 / mIoU	MoF / F1 / mIoU	MoF / F1 / mIoU	MoF / F1 / mIoU
Full-Dataset Evaluation	CTE* [33]	41.8 / 26.4 / -	39.0 / 28.3 / -	30.2 / - / -	35.5 / - / -	47.6 / 44.9 / -
	CTE† [33]	47.2 / 27.0 / 14.9	35.9 / 28.0 / 9.9	30.1 / 25.5 / 17.9	35.0 / 35.5 / 21.6	- / - / -
	VTE [71]	48.1 / - / -	- / 29.9 / -	24.2 / - / -	30.6 / - / -	- / - / -
	UDE [63]	47.4 / 31.9 / -	43.8 / 29.6 / -	- / - / -	42.2 / 34.4 / -	- / - / -
	ASAL [42]	52.5 / 37.9 / -	44.9 / 32.1 / -	34.4 / - / -	39.2 / - / -	- / - / -
	TOT [34]	47.5 / 31.0 / -	40.6 / 30.0 / -	31.8 / - / -	47.4 / 42.8 / -	56.3 / 51.7 / -
	TOT+ [34]	39.0 / 30.3 / -	45.3 / 32.9 / -	34.3 / - / -	44.5 / 48.2 / -	58.1 / 53.4 / -
	UFSA (M) [69]	- / - / -	43.2 / 30.5 / -	- / - / -	47.8 / 34.8 / -	- / - / -
	UFSA (T) [69]	52.1 / 38.0 / -	49.6 / 32.4 / -	36.7 / 30.4 / -	55.8 / 50.3 / -	65.4 / 63.0 / -
	ASOT [78]	56.1 / 38.3 / 18.6	52.9 / 35.1 / 24.7	46.2 / 37.4 / 24.9	59.3 / 53.6 / 30.1	70.4 / 68.0 / 45.9
	HVQ [58]	54.4 / **39.7** / -	50.3 / 35.1 / -	- / - / -	- / - / -	- / - / -
	VASOT (Ours)	**57.5** / 39.0 / **18.8**	**53.2 / 35.7 / 25.2**	**47.2 / 41.3 / 26.1**	**60.6 / 57.4 / 34.5**	**70.9 / 75.1 / 49.3**

Table 3. Action segmentation comparison results. **Bold** and underline denote the best and second best respectively.

stantial. The results validate the benefit of fusing video alignment with action segmentation and demonstrate that video alignment boosts action segmentation results notably in a multi-task learning setup. Moreover, Fig. 5 shows some qualitative results, where VASOT predicts segmentations which capture action boundaries more accurately and are more closely aligned with ground truth than ASOT [78].

5. Conclusion

This paper presents a novel approach for joint self-supervised video alignment and action segmentation. We first develop a fused Gromov-Wasserstein optimal transport with a structural prior for self-supervised video alignment, outperforming prior works. Our single-task method trains efficiently on GPUs and needs few iterations to derive the optimal transport solution. Next, we extend our approach to a unified optimal transport framework for joint self-supervised video alignment and action segmentation, yielding similar video alignment yet better action segmentation results than prior works. Our multi-task method requires training and storing a single model and saves both time and memory usage. To our best knowledge, our work is the first to explore the relationship between video alignment and action segmentation. Our future works will explore deep supervision [40, 41], complex weighting in multi-task learning [9, 28], and other potential applications (e.g., joint keypoint matching and clustering [19, 54]).

References

[1] Unaiza Ahsan, Chen Sun, and Irfan Essa. Discrimnet: Semi-supervised action recognition from videos using generative adversarial networks. *arXiv preprint*, 2018. 2

[2] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4575–4583, 2016. 5

[3] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 847–859, 2021. 5

[4] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9922–9931, 2020. 2

[5] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019. 2

[6] 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

[7] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2959–2968, 2019. 2

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[9] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018. 8

[10] Sayeed Shafayet Chowdhury, Soumyadeep Chandra, and Kaushik Roy. Opel: Optimal transport guided procedure learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[11] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 1, 2

[12] Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In *International conference on machine learning*, pages 894–903. PMLR, 2017. 1, 2

[13] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, 219:103406, 2022. 2

[14] Henri De Plaen, Pierre-François De Plaen, Johan AK Suykens, Marc Proesmans, Tinne Tuytelaars, and Luc Van Gool. Unbalanced optimal transport: A unified framework for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3198–3207, 2023. 3

[15] Ali Diba, Vivek Sharma, Luc Van Gool, and Rainer Stiefelhagen. Dynamonet: Dynamic action and motion network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6192–6201, 2019. 2

[16] Guodong Ding, Fadime Sener, and Angela Yao. Temporal action segmentation: An analysis of modern techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[17] Gerard Donahue and Ehsan Elhamifar. Learning to predict activity progress by self-supervised video alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18667–18677, 2024. 2, 6, 7, 8

[18] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 6, 7, 8

[19] Francisco J Estrada, Pascal Fua, Vincent Lepetit, and Sabine Susstrunk. Appearance-based keypoint clustering. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1279–1286. IEEE, 2009. 8

[20] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1

[21] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3299–3309, 2021. 2

[22] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3636–3645, 2017. 2

[23] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 2

[24] Ross Goroshin, Joan Bruna, Jonathan Tompson, David Eigen, and Yann LeCun. Unsupervised learning of spatiotemporally coherent metrics. In *Proceedings of the IEEE international conference on computer vision*, pages 4086–4093, 2015. 2

[25] Sanjay Haresh, Sateesh Kumar, Huseyin Coskun, Shahram N Syed, Andrey Konin, Zeeshan Zia, and Quoc-Huy Tran. Learning by aligning videos in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5548–5558, 2021. 1, 2, 6, 7, 8

[26] Kai Hu, Jie Shao, Yuan Liu, Bhiksha Raj, Marios Savvides, and Zhiqiang Shen. Contrast and order representations for video self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7939–7949, 2021. 2

[27] Simon Jenni, Hailin Jin, and Paolo Favaro. Steering self-supervised feature learning beyond local pixel statistics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6408–6417, 2020. 2

[28] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018. 8

[29] Abdelwahed Khamis, Russell Tsuchida, Mohamed Tarek, Vivien Rolland, and Lars Petersson. Scalable optimal transport methods in machine learning: A contemporary survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[30] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8545–8552, 2019. 2

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

[32] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014. 5

[33] Anna Kukleva, Hilde Kuehne, Fadime Sener, and Jurgen Gall. Unsupervised learning of action classes with continuous temporal embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12066–12074, 2019. 2, 6, 7, 8

[34] Sateesh Kumar, Sanjay Haresh, Awais Ahmed, Andrey Konin, M Zeeshan Zia, and Quoc-Huy Tran. Unsupervised action segmentation by joint representation learning and online clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20174–20185, 2022. 2, 3, 5, 6, 7, 8

[35] Taein Kwon, Bugra Tekin, Siyu Tang, and Marc Pollefeys. Context-aware sequence alignment using 4d skeletal augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8172–8182, 2022. 2

[36] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European conference on computer vision*, pages 577–593. Springer, 2016. 2

[37] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6874–6883, 2017. 2

[38] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017. 2

[39] Seonghak Lee, Jisoo Park, Radu Timofte, and Junseok Kwon. Sota: Sequential optimal transport approximation for visual tracking in wild scenario. *IEEE Access*, 2024. 3

[40] Chi Li, M Zeeshan Zia, Quoc-Huy Tran, Xiang Yu, Gregory D Hager, and Manmohan Chandraker. Deep supervision with shape concepts for occlusion-aware 3d object parsing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5465–5474, 2017. 8

[41] Chi Li, M Zeeshan Zia, Quoc-Huy Tran, Xiang Yu, Gregory D Hager, and Manmohan Chandraker. Deep supervision with intermediate concepts. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1828–1843, 2018. 8

[42] Jun Li and Sinisa Todorovic. Action shuffle alternating learning for unsupervised action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12628–12636, 2021. 2, 6, 7, 8

[43] Lilang Lin, Sijie Song, Wenhan Yang, and Jiaying Liu. Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2490–2498, 2020. 2

[44] Weizhe Liu, Bugra Tekin, Huseyin Coskun, Vibhav Vineet, Pascal Fua, and Marc Pollefeys. Learning to align sequential actions in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2181–2191, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[45] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7661–7669, 2018. 2

[46] 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. 3

[47] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016. 2, 6, 7, 8

[48] Hossein Mobahi, Ronan Collobert, and Jason Weston. Deep learning from temporal coherence in video. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 737–744, 2009. 2

[49] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5898–5906, 2017. 2

[50] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[51] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pages 2664–2672. PMLR, 2016. 3, 4

[52] Lyndsey C Pickup, Zheng Pan, Donglai Wei, YiChang Shih, Changshui Zhang, Andrew Zisserman, Bernhard Scholkopf,

and William T Freeman. Seeing the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2035–2042, 2014. 2

[53] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021. 2

[54] 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. 3, 8

[55] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1134–1141. IEEE, 2018. 5, 6, 7, 8

[56] Zhengyang Shen, Jean Feydy, Peirong Liu, Ariel H Curiale, Ruben San Jose Estepar, Raul San Jose Estepar, and Marc Niethammer. Accurate point cloud registration with robust optimal transport. *Advances in Neural Information Processing Systems*, 34:5373–5389, 2021. 3

[57] Chenyang Si, Xuecheng Nie, Wei Wang, Liang Wang, Tieniu Tan, and Jiashi Feng. Adversarial self-supervised learning for semi-supervised 3d action recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 35–51. Springer, 2020. 2

[58] Federico Spurio, Emad Bahrami, Gianpiero Francesca, and Juergen Gall. Hierarchical vector quantization for unsupervised action segmentation. *arXiv preprint arXiv:2412.17640*, 2024. 6, 7, 8

[59] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015. 2

[60] Sebastian Stein and Stephen J McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 729–738, 2013. 5

[61] Kun Su, Xiulong Liu, and Eli Shlizerman. Predict & cluster: Unsupervised skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9631–9640, 2020. 2

[62] Yukun Su, Guosheng Lin, and Qingyao Wu. Self-supervised 3d skeleton action representation learning with motion consistency and continuity. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13328–13338, 2021. 2

[63] Sirnam Swetha, Hilde Kuehne, Yogesh S Rawat, and Mubarak Shah. Unsupervised discriminative embedding for sub-action learning in complex activities. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2588–2592. IEEE, 2021. 3, 6, 7, 8

[64] Matthew Thorpe. Introduction to optimal transport. *Notes of Course at University of Cambridge*, 2018. 3

[65] Alexis Thual, Quang Huy Tran, Tatiana Zemskova, Nicolas Courty, Rémi Flamary, Stanislas Dehaene, and Bertrand Thirion. Aligning individual brains with fused unbalanced gromov wasserstein. *Advances in neural information processing systems*, 35:21792–21804, 2022. 3, 4

[66] Vayer Titouan, Nicolas Courty, Romain Tavenard, and Rémi Flamary. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pages 6275–6284. PMLR, 2019. 3

[67] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 1

[68] Quoc-Huy Tran, Muhammad Ahmed, Murad Popattia, M Hassan Ahmed, Andrey Konin, and M Zeeshan Zia. Learning by aligning 2d skeleton sequences and multimodality fusion. In *European Conference on Computer Vision*, pages 141–161. Springer, 2024. 2

[69] Quoc-Huy Tran, Ahmed Mehmood, Muhammad Ahmed, Muhammad Naufil, Anas Zafar, Andrey Konin, and Zeeshan Zia. Permutation-aware activity segmentation via unsupervised frame-to-segment alignment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6426–6436, 2024. 2, 3, 6, 7, 8

[70] Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused gromov-wasserstein distance for structured objects. *Algorithms*, 13(9):212, 2020. 3

[71] Rosaura G VidalMata, Walter J Scheirer, Anna Kukleva, David Cox, and Hilde Kuehne. Joint visual-temporal embedding for unsupervised learning of actions in untrimmed sequences. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1238–1247, 2021. 2, 6, 7, 8

[72] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances in neural information processing systems*, pages 613–621, 2016. 2

[73] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 504–521. Springer, 2020. 2

[74] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 1

[75] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8052–8060, 2018. 2

[76] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019. 2

[77] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke. Gromov-wasserstein learning for graph matching and node embedding. In *International conference on machine learning*, pages 6932–6941. PMLR, 2019. 3

[78] Ming Xu and Stephen Gould. Temporally consistent unbalanced optimal transport for unsupervised action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14618–14627, 2024. 2, 3, 4, 5, 6, 7, 8

[79] Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervised spatio-temporal representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6548–6557, 2020. 2

[80] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE international conference on computer vision*, pages 2248–2255, 2013. 5

[81] Nenggan Zheng, Jun Wen, Risheng Liu, Liangqu Long, Jianhua Dai, and Zhefeng Gong. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2

[82] Will Y Zou, Andrew Y Ng, and Kai Yu. Unsupervised learning of visual invariance with temporal coherence. In *NIPS 2011 workshop on deep learning and unsupervised feature learning*, 2011. 2