Learning to Align Sequential Actions in the Wild

Weizhe Liu[1*] Bugra Tekin[2] Huseyin Coskun[3] Vibhav Vineet[2] Pascal Fua[4] Marc Pollefeys[2,5]
[1] Tencent AI Lab [2] Microsoft [3] Technische Universität München [4] EPFL [5] ETH Zurich

Abstract

State-of-the-art methods for self-supervised sequential action alignment rely on deep networks that find correspondences across videos in time. They either learn frame-to-frame mapping across sequences, which does not leverage temporal information, or assume monotonic alignment between each video pair, which ignores variations in the order of actions. As such, these methods are not able to deal with common real-world scenarios that involve background frames or videos that contain non-monotonic sequence of actions.

In this paper, we propose an approach to align sequential actions in the wild that involve diverse temporal variations. To this end, we propose an approach to enforce temporal priors on the optimal transport matrix, which leverages temporal consistency, while allowing for variations in the order of actions. Our model accounts for both monotonic and non-monotonic sequences and handles background frames that should not be aligned. We demonstrate that our approach consistently outperforms the state-of-the-art in self-supervised sequential action representation learning on four different benchmark datasets. Code is publicly available at https://github.com/weizheliu/VAVA.

1. Introduction

Understanding human activities in video sequences is important for applications such as human-computer interaction, video analysis, robot learning, and surveillance. In recent years, a significant amount of research has focused on supervised, coarse-scale action understanding. Most of the work focuses on predicting explicit classes for clips corresponding to a certain limited set of action categories in a supervised fashion [8, 11, 33, 52, 53, 55]. While giving a categorical understanding of human behavior, such techniques do not provide a fine-grained analysis of human action. Furthermore, the dependence on per-frame labels requires a large amount of human effort that does not scale up to many different types of subjects, environments, and scenarios. For such supervised methods, it is also not always clear what exhaustive set of labels is required for a fine-grained understanding of videos. Thus, recent papers [16, 28] advocate self-supervised learning of video representation without frame-wise action labels. They rely on the fact that human activities often involve many sequential steps in a predictable order. To drink water, one might grab a mug, drink, and then put the mug down. To change a tire, one would first lift the vehicle off the ground, remove the wheel, and replace it by a spare one. Assuming the order is set, visual representations can be learned from multiple videos of the same activity by temporal alignment of the frames.

This is often done by monotonically aligning the frames [25], which assumes that actions always occur in the same order. However, in most real-world sequences, this is not the case and temporal deviations such as those depicted by Fig. 1 do occur. They can be summarized as follows:

- *Background frames*: Frames that are not related to the major activity and should therefore not be aligned. For example, you might get a phone call while changing the tire. In this case, the "phone call" frames are background frames that are not related to the major activity and should be ignored.

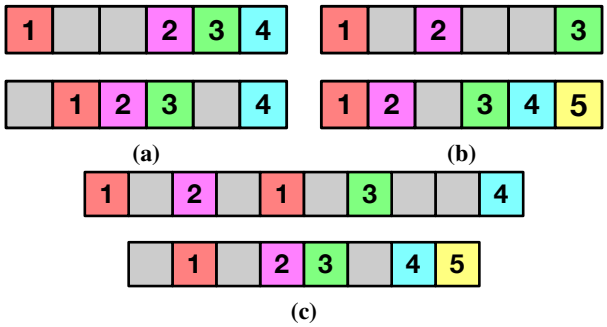


Figure 1. **Temporal Variations [15]. (a)** *Background frames*, depicted as gray blocks, are not related to the major activity. **(b)** Frames with number 4 and 5 are *redundant frames* which only exist in one sequence but not in the other. **(c)** Frames with action 1 in the first sequence occur before and after action 2 and forms a sequence of *non-monotonic frames*. Our approach explicitly tackles all of these temporal variations and is suitable for aligning sequential actions in a broad context.

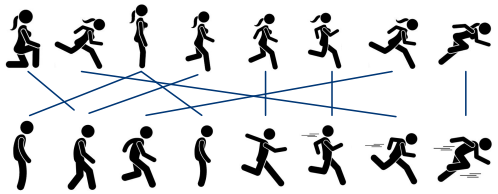
*This work was completed during an internship at Microsoft Mixed Reality & AI Lab.

- *Redundant frames*: Frames that only exist in one sequence but not in the other. For example, one person might put on gloves before changing the tire while another does not. In this case, the "glove wearing" frames are redundant and should be ignored as well.
- *Non-monotonic frames*: Frames that occur in non-monotonic order. For example, while changing the tire, you lift your vehicle off the ground and try to remove the tire, only to realize that you have not lifted high enough. You then go back to the previous action, that is, lifting, before proceeding with the remaining actions.

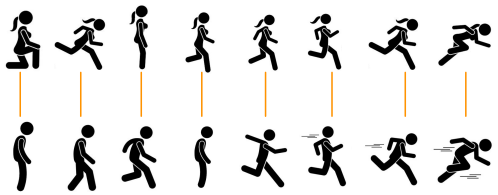
Our method aims at tackling all these cases and reduces the stringent assumptions of earlier work on the temporal sequence of actions. For this purpose, we propose an approach for learning temporal correspondences across videos through a novel alignment framework. Our model accounts for temporal variations exhibited across real-world sequences with a differentiable deep network formulation that relies on an optimal transport loss. While optimal transport is able to align non-monotonic sequences based on frame-wise matching of the features computed from individual frames, it ignores temporal smoothness and ordering relationships of the videos. To remedy this, we introduce temporal priors on the transportation matrix that the optimal transport algorithm takes as input. This accounts for the temporal structure of the sequence and enforces time consistency during alignment in a flexible way. This is unlike previous work that either ignores temporal priors within sequences [16] or enforces monotonic alignment between pairs of videos [28], as depicted by Fig. 2.

In particular, we enforce a temporal prior by modeling the diagonal of the optimal transport matrix with an adaptive Gaussian Mixture Model (GMM). Our temporal prior effectively favors transportation of one sequence to the elements in the nearby temporal positions of the other sequence, and, hence respects the overall temporal structure and order of the sequences during alignment. At the same time, our optimal transport based formulation aims to find ideal frame-wise matches and handles non-monotonic frames. To explicitly handle background and redundant frames, we further propose an approach, which introduces an additional virtual frame in the optimal transport matrix so that unmatched frames are explicitly assigned to it. Furthermore, since enforcing temporal priors on video alignment generally suffers from converging to trivial solutions [25], we introduce a novel inter-video contrastive loss to regularize the learning process. In particular, our contrastive loss optimizes for disentangled video representations, i.e., videos that are close in terms of their similarity given by optimal transport are mapped to spatially nearby points in the embedding space and vice versa.

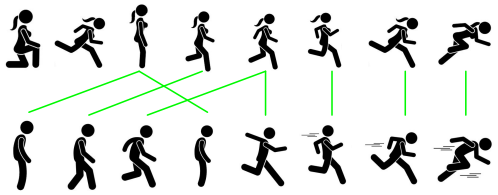
Our contributions can be summarized as follows: First, we propose a self-supervised learning approach that aligns



(a) Alignment without Temporal Prior



(b) Monotonic Alignment



(c) VAVA

Figure 2. **Types of Alignment Priors.** Two example sequences of people running. The first sequence shows a person starting from a professional crouching position while the second sequence shows another person starting from a normal standing position. **(a)** Alignment without temporal prior is based on pure appearance similarity, therefore the starting "crouching" action of the first sequence is incorrectly aligned with the "speeding up" action in the second one. **(b)** Monotonic alignment enforces pure monotonic order, therefore even though two actions look quite different, they can be incorrectly aligned. **(c)** By contrast, our approach, VAVA, enforces temporal priors to address non-monotonic frames and gracefully handles unmatched frames (*e.g.* crouching position in the first sequence), resulting in accurate alignment between sequences.

sequential actions in-the-wild, which feature a diverse set of temporal variations. Second, we enforce adaptive temporal priors on optimal transport, which could efficiently handle non-monotonic frames while respecting the local temporal structure of sequences. Third, we extend the optimal transport formulation with an additional virtual frame that actively handles redundant and background frames that should not be matched. Finally, to prevent our model from converging to trivial solutions, we propose a novel contrastive loss term that regularize the learning of optimal transport matrix. In Sec. 4, we show quantitatively that these contributions allow us to reliably learn robust temporal correspondences and align sequential actions in real-world settings. Our self-supervised approach, which we call *Variation-Aware Video Alignment (VAVA)*, uses temporal alignment as a pretext to learn visual representations that are effective in downstream tasks, such as action phase classification and tracking the progress of an action, and significantly outperforms state-

of-the-art methods on four different benchmark datasets.

2. Related Work

Self-Supervised Video Representation Learning. Temporal information in videos provide rich supervision signal to learn strong spatio-temporal representations [14, 18, 45]. This contrasts to single-image based approaches [9, 17, 19, 23, 29, 31, 32, 34, 35, 39, 42, 59, 61] that only rely on spatial signal. Misra *et al.* [40] introduce the idea of learning such visual representations by estimating the order of shuffled video frames. Inspired by the success of this approach, several recent papers focused on designing a novel pretext task using temporal information, such as predicting future frames [13, 49, 54] or their embeddings [21, 27]; estimating the order of frames [10, 20, 36, 40, 57] or the direction of video [56]. Another line of research focuses on using temporal coherence [6, 24, 26, 41, 62, 63] as supervision signal.

However, these methods usually optimize over a single video at a time, therefore they exploit less information compared with approaches that jointly optimize over a pair of videos [16, 28]. Furthermore, such visual representations are learned by maximizing the similarity of two randomly cropped and augmented clips from the same video [4, 18, 38, 43, 45]. This requires training videos that contain the exact same single action. However, in real world scenarios, a complex human activity typically involves multiple actions and even background frames. Another limitation of these approaches is that they aim to learn coarse clip-wise visual representations, therefore they are not suitable for frame-wise downstream tasks like fine-grained action recognition. In contrast to these methods, we propose a self-supervised learning strategy that can learn frame-wise representations from unconstrained videos that involve sequential actions.

Video Alignment is rather straightforward to address if the videos are synchronized. This can be done by using existing methods such as CCA [2, 3] and DTW [7]. Recent trend in computer vision [47] leverages deep networks and proposes to align videos by learning self-supervised visual representations from videos with the same human activity. In this regard, Sermanet *et al.* [47] propose to learn cross-sequence visual representation by aligning synchronized multi-view videos that record exactly the same human actions from different viewpoints. As synchronized multi-view videos are not always available, this approach cannot be generalized to unconstrained settings. Dwibedi *et al.* [16] address this issue by finding frame correspondences across unsynchronized videos with cycle consistency loss, however, this approach only looks for local matches across sequences and does not explicitly account for the global temporal structure of the videos.

Maybe the most similar works to our approach are [25, 28], which align video pairs with the assumption of strictly monotonic temporal order. As we explained in the introduction, this assumption is too strong and seldom happens naturally in real-world scenarios. In contrast to these methods, our approach does not require synchronized videos and learns to align video sequences from in-the-wild settings, which includes temporal variations, such as background frames, redundant frames and non-monotonic frames. As shown in Sec. 4, our approach consistently outperforms above methods and the margin is even larger if there were temporal variations.

Optimal Transport. Optimal transport measures the dissimilarity between two probability distributions over a metric space. Given feature vectors associated to each entity and matrix of distances between them, it provides a way to establish correspondences between features that minimize the sum of distances. Besides, it also provides guarantees of optimality, separability, and completeness. These desirable properties have been leveraged for many different tasks, such as scene flow estimation [37, 44], object detection [22], domain adaptation [58], classification [48] and point matching [46] that matches features in spatial domain. However, none of them focuses on sequence alignment as we do. One potential reason is that vanilla optimal transport formulation does not account for temporal priors, therefore the alignment is less reliable in the time domain, as depicted by Fig. 2(a). One exception is [50], which uses optimal transport only to measure the distance between skeleton sequences and does not learn a visual representation as we do. Besides, it only enforces monotonic temporal priors without accounting for the cases of temporal variations, therefore is less flexible than our approach which specifically addresses such situations.

3. Approach

In this section, we first formalize the problem of self-supervised representation learning by aligning frames from pairs of video sequences (Sec. 3.1). After that, we present our approach for incorporating temporal priors in optimal transport to leverage temporal information and handle non-monotonic frames (Sec. 3.2). We then propose an effective way to deal with background and redundant frames (Sec. 3.3). Finally, we provide a summary of our loss function and model details (Sec. 3.4).

3.1. Alignment by Optimal Transport

Given two sequences of video frames $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_N]$ and $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_M]$, we take their respective embeddings to be $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$ and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_M]$. $\mathbf{X}$ and $\mathbf{Y}$ are computed with an encoder network ϕ, as depicted in Fig. 3.

If frames $\mathbf{s}_i$ and $\mathbf{v}_j$ represent the same fine-grained action, the distance between their respective embeddings, $\mathbf{x}_i$ and $\mathbf{y}_j$, should be small, otherwise, the distance should be

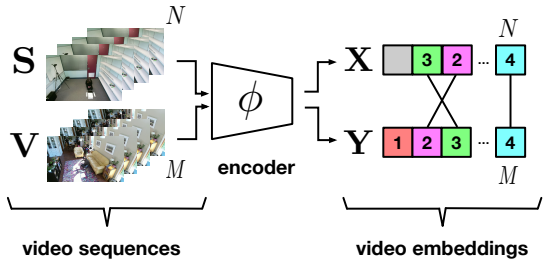


Figure 3. **Encoder Network and Video Embeddings.**

large. Given such embeddings, Optimal Transport (OT) can be used to align two such sequences by first computing an $N \times M$ distance matrix, $\boldsymbol{D}$, whose components are Euclidean distances between embedding vectors, that is, $d(\mathbf{x}_i, \mathbf{y}_j) = \|\mathbf{x}_i - \mathbf{y}_j\|$. The optimal assignment, $d_O(\mathbf{X}, \mathbf{Y})$, between the embeddings can be found by solving the following optimization problem:

$$d_O(\mathbf{X}, \mathbf{Y}) := \min_{\boldsymbol{T} \in U(\boldsymbol{\alpha}, \boldsymbol{\beta})} < \boldsymbol{T}, \boldsymbol{D} > \quad (1)$$

Here, $< \cdot, \cdot >$ is the Frobenius dot product, and, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_N)$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_M)$ are non-negative weights that sum to one and denote the relative importance of individual frames. As we have no reason to weigh one frame more than the others, we take $\alpha_i = 1/N$ and $\beta_j = 1/M$, for all i and j. The set of all feasible transport matrices is represented with U. A valid transportation matrix in U satisfies that the row and column-wise sum are equal to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ [12], in particular:

$$U(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \{\boldsymbol{T} \in \mathbb{R}^{N \times M} | \boldsymbol{T}\mathbf{1}_M = \boldsymbol{\alpha}, \boldsymbol{T}^\top \mathbf{1}_N = \boldsymbol{\beta}\} \quad (2)$$

Eq. 1 can be solved with linear programming, however, this is a computationally expensive procedure and is not suitable for training purposes. To address this issue, Cuturi [12] proposes to regularize OT problem with an additional entropy term and solves it using Sinkhorn algorithm.

$$d_O(\mathbf{X}, \mathbf{Y}) := \min_{\boldsymbol{T} \in U(\boldsymbol{\alpha}, \boldsymbol{\beta})} < \boldsymbol{T}, \boldsymbol{D} > -\upsilon h(\boldsymbol{D}) \quad (3)$$

where h is an entropy term that regularizes the problem and υ is a small scalar coefficient. Here, the entries of the transport matrix, *i.e.* the $t_{i,j}$ coefficients of $\boldsymbol{T}$, can be understood to be proportional to the probability that frame i in $\mathbf{S}$ is aligned with frame j in $\mathbf{V}$. A large value of distance $d_{i,j}$ would correspond to a small value of $t_{i,j}$, which implies that these two frames are dissimilar and thus have a low chance of alignment. The benefit of such formulation is that we can enforce temporal priors by modeling $\boldsymbol{T}$ to follow a predefined temporal distribution.

3.2. Enforcing Temporal Priors

While optimal transport measures the minimum cost of aligning two sequences, it completely ignores temporal ordering relationships and therefore does not exploit the temporal consistency that we know to be present in video sequences. In most cases, given multiple videos of the same activity, the temporal position of one sequence should only be aligned to elements in the nearby temporal positions of the other sequence. In an extreme case where the two sequences are perfectly aligned, the transport matrix $\boldsymbol{T}$ should be diagonal. In practice, this is far too strong a constraint. As depicted by Fig. 4, the activity may start a bit earlier in one sequence than the other; it may be faster; the actions in one of the video sequences may be monotonic while the other ones are not.

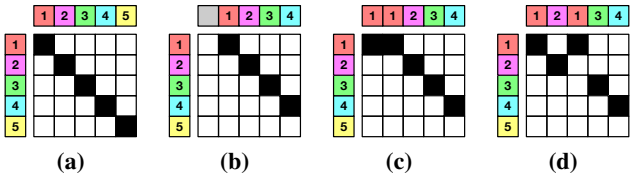


Figure 4. **Assignment Variations.** (a) Two videos strictly follow the same temporal order, the assignment matrix has peak values along the diagonal. (b) The activity from one video starts a bit earlier than the other one, hence the assignment matrix has peak values parallel to the diagonal. (c) The action of one video is slower than the other, thus the assignment matrix has peak values that are near the diagonal without being strictly parallel to it. (d) Actions follow monotonic order in one sequence but not in the other.

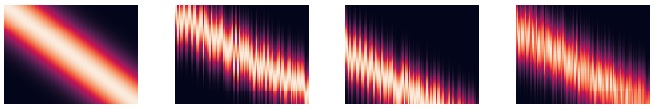
To capture temporal variations across sequences, while being able to optimally align two videos, we propose to enforce temporal priors on the optimal transport problem. To this end, we propose a novel prior distribution of transport matrix with an adaptive Gaussian Mixture Model (GMM) that comprises of two temporal priors.

The first prior, which we call as *Consistency Prior*, favors transportation of one sequence to the elements in the nearby temporal positions of the other sequence, and hence respects the overall temporal structure and the consistency in the order of the actions across sequences. With this prior, assignment matrix is very likely to have peak values along the diagonal and the values should gradually decrease along the direction perpendicular to the diagonal, as depicted by Fig. 5(a). We can model this situation with a two-dimensional distribution, in which the distribution along any line perpendicular to the diagonal is a Gaussian distribution centered on the diagonal. We model the *Consistency Prior* on the assignment matrix with a Gaussian as follows

$$\boldsymbol{P}_c(i,j) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{l_c^2(i,j)}{2\sigma^2}}, \quad (4)$$

where $l_c(i,j)$ is the distance from the position (i,j) to the diagonal

$$l_c(i,j) = \frac{|i/N - j/M|}{\sqrt{1/N^2 + 1/M^2}}. \quad (5)$$



(a) (b) (c) (d)

Figure 5. **Toy Example of Temporal Priors.** Light color denotes high alignment probability. (a) Consistency prior with the peak values appearing along the diagonal. (b) Ground-truth probability for which there exist many non-monotonic frames. (c) Optimality prior from transportation matrix, with the peak values appearing on the locations of most similar pairs in the embedding space. (d) Gaussian mixture of (b) and (c), which more accurately represents the ground-truth, shown in (b), as compared to (a) or (c).

This prior, while modeling consistency across sequences, does not allow for handling unconstrained non-monotonic sequences. For example, in the extreme case of actions being performed in the exact reverse order across two sequences, the consistency prior would not be able to capture temporal variations. Similarly, for two sequences, in which there exists many non-monotonic frames, as depicted by Fig. 5(b), this probability distribution would not ideally model the alignment.

To be able to explicitly deal with non-monotonic sequences, we propose another prior, which we call, *Optimality Prior*. Recall that the transport matrix $\boldsymbol{T}$ we compute in Eq. 3 during the training process indicates the rough alignment between two video sequences and changes dynamically according to the temporal variations across sequences. We exploit this transport matrix to model another temporal prior. In particular, as depicted by Fig. 5(c), we model our prior, such that the distribution along any line perpendicular to the diagonal is a Gaussian, centered at the intersection of the most likely alignment based on the transport matrix. We model the *Optimality Prior* on the assignment matrix with

$$\boldsymbol{P}_o(i,j) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{l_o^2(i,j)}{2\sigma^2}} , \tag{6}$$

where $l_o(i,j)$ is the average distance from the position (i,j), to the frame locations that give the optimal alignment, (i, j_o) and (i_o, j), given by the transport matrix

$$l_o(i,j) = \frac{|i/N - i_o/N| + |j/M - j_o/M|}{2\sqrt{1/N^2 + 1/M^2}} . \tag{7}$$

In short, Consistency Prior $\boldsymbol{P}_c$ represents the general case, in which the sequence pairs follow the same coarse ordering, while Optimality Prior $\boldsymbol{P}_o$ models the potential temporal variations across sequences. As shown by Fig. 5(d), the ground truth distribution is more accurately represented by the combination of these two priors, which we formulate using a Gaussian Mixture Model, as follows:

$$\boldsymbol{P}(i,j) = \psi \boldsymbol{P_c}(i,j) + (1-\psi)\boldsymbol{P_o}(i,j) , \tag{8}$$

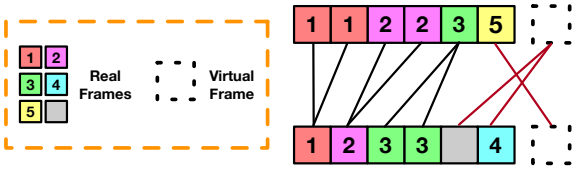


Figure 6. **Virtual Frame.** Virtual frame enables the model to handle unmatched frames that should not be aligned. Redundant frames (shown with 4 and 5) and background frames (shown in gray) are explicitly assigned to it.

where $\psi \in [0,1]$ is a weighting parameter that we set to 1.0 initially and decrease gradually over time to account for the fact that the learned transport matrix is less reliable in the very beginning of the training and becomes more robust in later stages. By enforcing temporal priors on optimal transport, our model is able to adaptively handle non-monotonic frames and temporal variations.

3.3. Handling Background and Redundant Frames

Consistency and *Optimality* temporal priors enable our model to handle non-monotonic frames between video sequences. However they do not explicitly handle background and redundant frames, introduced in Sec. 1. To be able to account for such frames in our model, we introduce an additional *virtual frame* in the transport matrix so that unmatched frames are explicitly assigned to it, as shown in Fig. 6.

To this end, we augment the transport matrix, $\boldsymbol{T} \in \mathbb{R}^{N\times M}$, with an additional entry for each sequence to obtain $\hat{\boldsymbol{T}} \in \mathbb{R}^{(N+1)\times(M+1)}$. The set of all feasible transportation matrices introduced in Eq. 2 then becomes

$$U(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) := \{\hat{\boldsymbol{T}} \in \mathbb{R}^{(N+1)\times(M+1)} | \hat{\boldsymbol{T}}\mathbf{1}_{M+1} = \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{T}}^\top \mathbf{1}_{N+1} = \hat{\boldsymbol{\beta}}\}$$

where $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are the weight vectors expanded with one extra element to account for the virtual frame. If the chance of alignment with all the real frames is less than a certain threshold value, ζ, we align this frame to the virtual frame instead. Note that many frames can be aligned with the virtual frame and the virtual frame does not follow the temporal priors we defined in Sec. 3.2.

3.4. Training Loss

VAVA Loss. Our model accounts for temporal variations exhibited across real-world sequences with a differentiable formulation that relies on an optimal transport loss. We regularize our loss function by exploiting temporal priors, as explained in Sec. 3.2. For the Consistency Prior described in Eq. 4, the large values of the transport matrix $\hat{\boldsymbol{T}}$ should be along the diagonal and the rest of the values should be small for other regions. Such a structure of the transport

matrix can be measured with

$$I_c(\hat{\boldsymbol{T}}) = \sum_{i=1}^{N+1}\sum_{j=1}^{M+1} \frac{t_{ij}}{(\frac{i}{N+1} - \frac{j}{M+1})^2 + 1} \ . \quad (9)$$

where we add one more row and column for virtual frames as explained in Sec. 3.3, $I_c(\hat{\boldsymbol{T}})$ in Eq. 9 is referred to as *inverse difference moment* in literature [1, 50] and will have large values for the region along the diagonal.

For the Optimality Prior described in Eq. 6, in which, large values appear in the most likely alignment locations given by the transport matrix, a similar structure can be captured with

$$I_o(\hat{\boldsymbol{T}}) = \sum_{i=1}^{N+1}\sum_{j=1}^{M+1} \frac{t_{ij}}{\frac{1}{2}d_o + 1} \quad (10)$$

$$d_o = (\frac{i - i_o}{N+1})^2 + (\frac{j - j_o}{M+1})^2$$

Our overall temporal prior that combines the Consistency Prior and the Optimality Prior can then be represented with the following loss function defined on the transport matrix

$$I(\hat{\boldsymbol{T}}) = \psi I_c(\hat{\boldsymbol{T}}) + (1-\psi) I_o(\hat{\boldsymbol{T}}) \ , \quad (11)$$

with the same ψ as we defined in Eq. 8. For a smooth alignment, we further enforce the expected distribution to be similar to the temporal priors by minimizing the Kullback-Leibler(KL) divergence between the two matrices

$$KL(\hat{\boldsymbol{T}}||\hat{\boldsymbol{P}}) = \sum_{i=1}^{N+1}\sum_{j=1}^{M+1} t_{ij} \log \frac{t_{ij}}{p_{ij}} \ , \quad (12)$$

where $\hat{\boldsymbol{P}}$ is as defined in Eq. 8, except that it is augmented with the virtual frame. Our variation-aware video alignment (VAVA) loss would therefore be defined by combining the temporal priors and the KL divergence within the optimal transport formulation (Eq. 3):

$$L_{vava} = d_O(\mathbf{X}, \mathbf{Y}) - \lambda_1 I(\hat{\boldsymbol{T}}) + \lambda_2 KL(\hat{\boldsymbol{T}}||\hat{\boldsymbol{P}}) \ , \quad (13)$$

where $d_O(\mathbf{X}, \mathbf{Y})$ is the Sinkhorn distance [12], as defined in Eq. 3, with extra row and column for virtual frames; λ_1 and λ_2 are hyper-parameters to weigh the two loss terms.

Contrastive Regularization. Enforcing temporal priors on video alignment generally suffers from converging to trivial solutions [28, 50]. The previous work [28] employs an *intra-video* contrastive loss term to regularize the training process. The *intra-video* contrastive loss for a given video embedding, $\mathbf{X}$, is defined as

$$C(\mathbf{X}) = \sum_{i=1}^{N+1}\sum_{j=1}^{M+1} \mathbb{1}_{|i-j|>\delta} \boldsymbol{W}(i,j) max(0, \lambda_3 - \hat{\boldsymbol{D}}_{\mathbf{X}}(i,j)) + \mathbb{1}_{|i-j|\leq\delta} \boldsymbol{W}(i,j) \hat{\boldsymbol{D}}_{\mathbf{X}}(i,j), \quad (14)$$

where $\mathbb{1}$ is an indicator function, which is 1, if the condition is met, and, 0 otherwise. $\boldsymbol{W}(i,j) = (i-j)^2 + 1$, is the distance in frame index and $\boldsymbol{D}_{\mathbf{X}}(i,j) = ||\mathbf{x}_i - \mathbf{x}_j||$, is the distance in the embedding space. δ is a window size for separating temporally far away and close frames and λ_3 is a margin parameter. This loss encourages close frames to be nearby in the embedding space, while penalizing temporally far away frames.

In our approach, in addition to using an intra-video contrastive loss term, we introduce an *optimal transport guided inter-video contrastive loss* to regularize the training process. In particular, we propose to contrast video pairs based on their similarity given by optimal transport. As discussed in Sec. 3.2, our transport matrix provides an estimate of the alignment between two sequences in the training stage. We leverage this information to enforce an *inter-video contrastive loss*:

$$C(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{N+1}\sum_{j=1}^{M+1} -\mathbb{1}_{\bar{\boldsymbol{A}}(i,j)} \hat{\boldsymbol{D}}_{\mathbf{X},\mathbf{Y}}(i,j) + \mathbb{1}_{\boldsymbol{A}(i,j)} \hat{\boldsymbol{D}}_{\mathbf{X},\mathbf{Y}}(i,j) \quad (15)$$

where $\boldsymbol{A}(i,j)$ denotes frames, i and j, that yield the largest $t_{i,j}$ value on our transport matrix for each row or column, while $\bar{\boldsymbol{A}}(i,j)$ denotes frames, i and j, that are least likely for alignment, which have the smallest $t_{i,j}$ values. This loss encourages frames to have similar latent embeddings if they are expected to be aligned by our optimal transport formulation, and, if not, they are enforced to have dissimilar latent embeddings. Our total regularization term is defined by

$$L_{cr} = C(\mathbf{X}) + C(\mathbf{Y}) + C(\mathbf{X}, \mathbf{Y}) \ . \quad (16)$$

Final Loss. Our final loss is obtained by combining the VAVA loss that enforces temporal priors on optimal transport (Eq. 13), along with contrastive regularization terms (Eq. 16) that optimize for disentangled representations of frames within and across sequences.

$$L_{all} = L_{vava} + \gamma L_{cr} \ . \quad (17)$$

Here, γ is a hyper-parameter to weigh the influence of the regularization term.

4. Evaluation

Datasets. We evaluate our approach on four different challenging datasets, namely COIN [51], IKEA ASM [5], Pouring [47], and Penn Action [60]. COIN and IKEA ASM datasets exhibit large temporal variations and comprise of *background frames*, *redundant frames* and *non-monotonic frames*, as described in Sec. 1. We therefore use them to demonstrate the effectiveness of our approach in aligning sequential actions in unconstrained environments. Pouring and Penn Action datasets do not contain any such temporal variation, that is, action order is strictly monotonic and there are no background frames in the videos. We use these

two datasets to benchmark our results against TCN [47] and LAV [28], which assume strict monotonic alignment. On the IKEA ASM dataset, [28] removes background frames for model training and evaluation. Since we aim to align unconstrained sequences, we instead keep the background frames by treating it as an additional category. In addition, in another evaluation setting, we remove the background frames to be able to compare against previous work [28].

Implementation Details. Following [16, 28], we use ResNet-50 [30] as the encoder network. The input videos are resized to 224×224. The embeddings are extracted from the output of $Conv4c$ layer and are of size $14 \times 14 \times 1024$. We initialized our networks from ImageNet pre-trained models as in [16, 28]. We set weighting of the regularization term, γ, in Eq. 17 as 0.5. We provide further details and ablation studies for the parameters we used in our Sup. Mat..

Evaluation Metrics. Following [16,28], we use three different metrics for our evaluation. We first train our encoder network on the training set without using any labels, and then evaluate the performance of our approach with the frozen embeddings. The first metric is *Phase Classification Accuracy*, which is the per frame classification accuracy for fine-grained action recognition. The second one is *Phase Progression(Progress) [16]*, which measures how well the *progress* of a process or action is captured by the embeddings. This metric assumes that actions are strictly consistent, thus is only suitable for monotonic datasets, that is Pouring and Penn Action, in our case. The last one is *Kendall's Tau (τ) [16]*, which is a statistical measure that can determine how well-aligned two sequences are in time. Since this metric assumes strictly monotonic order of actions, it is only suitable for Pouring and Penn Action datasets. For all measures a higher score implies a better model.

4.1. Comparison to the State-of-the-Art

We evaluate the accuracy of our learned representation in the action phase classification task with an SVM classifier trained on a fraction 0.1, 0.5 and 1.0 of the ground truth labels. We compare against the accuracy numbers reported in [16, 28] on the Pouring, Penn Action and IKEA ASM datasets. Previous approaches do not report results on the unconstrained COIN dataset. Therefore we reproduce the results of these baselines on this dataset, to be able benchmark our results against them. To do so, we follow the implementation details of [16, 28] and also validate the accuracy of our reproduced implementation on the Pouring, Penn Action and IKEA ASM datasets. We denote our Variation-Aware Video Alignment approach as **VAVA** and report results on the COIN, IKEA ASM, Pouring and Penn Action datasets in Table 1.

Our model clearly outperforms earlier work on the COIN and IKEA ASM datasets which feature temporal variations that are exhibited by many real world applications. Particularly, the improvement over state-of-the-art methods is around 7% (with a relative improvement of 20%) on the COIN dataset, which demonstrates the effectiveness of our approach for aligning sequential actions across unlabeled videos from in-the-wild settings. Similarly, **VAVA** achieves 5% improvement (with a relative increase of 25%) over existing approaches on the IKEA ASM dataset that shows the benefits of our approach in aligning videos that feature temporal variations.

For Pouring and Penn Action datasets that do not involve temporal variations, our approach still outperforms previous work in phase progression, Kendall's τ and most of the phase classification accuracies, which demonstrates the representation power of our framework in modeling the progress of actions and their temporal structure. Note also that Pouring dataset contains videos that follow a strict monotonic temporal order, and therefore methods that rely on the monotonicity assumption [28] are more likely to overfit to this dataset.

Dataset	Model	Fraction of Labels 0.1	0.5	1.0	Progress	τ
COIN	Supervised Learning	37.11	40.73	49.18	-	-
	Random Features	29.50	30.29	30.38	-	-
	Imagenet Features	31.32	34.74	37.43	-	-
	SAL [40]	34.69	39.23	40.32	-	-
	TCN [47]	34.87	39.73	40.51	-	-
	TCC [16]	35.87	39.56	40.66	-	-
	LAV [28]	36.79	38.85	39.81	-	-
	VAVA(ours)	**43.77**	**46.18**	**47.26**	-	-
IKEA ASM No Background	Supervised Learning	21.76	30.26	33.81	-	-
	Random Features	17.89	17.89	17.89	-	-
	Imagenet Features	18.05	19.27	19.50	-	-
	SAL [40]	21.68	21.72	22.14	-	-
	TCN [47]	25.17	25.70	26.80	-	-
	TCC [16]	24.74	25.22	26.46	-	-
	LAV [28]	29.78	29.85	30.43	-	-
	VAVA(ours)	**31.66**	**33.79**	**32.91**	-	-
IKEA ASM Background	Supervised Learning	20.74	25.61	31.92	-	-
	Random Features	17.03	17.41	17.61	-	-
	Imagenet Features	17.27	18.02	18.64	-	-
	SAL [40]	22.94	23.43	25.46	-	-
	TCN [47]	22.51	25.47	25.88	-	-
	TCC [16]	22.70	25.04	25.63	-	-
	LAV [28]	23.19	25.47	25.54	-	-
	VAVA(ours)	**29.12**	**29.95**	**29.10**	-	-
Pouring	Supervised Learning	75.43	86.14	91.55	-	-
	Random Features	42.73	45.94	46.08	-	-
	Imagenet Features	43.85	46.06	51.13	-	-
	SAL [40]	85.68	87.84	88.02	0.7451	0.7331
	TCN [47]	89.19	90.39	90.35	0.8057	0.8669
	TCC [16]	89.23	91.43	91.82	0.8030	0.8516
	LAV [28]	91.61	**92.82**	**92.84**	0.8054	0.8561
	VAVA(ours)	**91.65**	91.79	92.45	**0.8361**	**0.8755**
Penn Action	Supervised Learning	67.10	82.78	86.05	-	-
	Random Features	44.18	46.19	46.81	-	-
	Imagenet Features	44.96	50.91	52.86	-	-
	SAL [40]	74.87	78.26	79.96	0.5943	0.6336
	TCN [47]	81.99	83.67	84.04	0.6762	0.7328
	TCC [16]	81.26	83.35	84.45	0.6726	0.7353
	GTA [25]	-	-	-	-	0.7829
	LAV [28]	83.56	83.95	84.25	0.6613	0.8047
	VAVA(ours)	**83.89**	**84.23**	**84.48**	**0.7091**	**0.8053**

Table 1. **Benchmark Evaluation.**

Figure 7. **Frame Retrieval.** VAVA can precisely reason about fine grained actions and background frames. While we capture the fine-grained action of *opening laptop cover*, [28] retrieves images where laptop cover is already open (top). We recover background frames more consistently in comparison to [28] (bottom).

Intra-Video	Inter-Video	KL	Consistency Prior	Optimality Prior	Virtual Frame	Threshold	Fraction of Labels 0.1	0.5	1.0
✓							20.38	23.09	23.27
	✓						19.46	22.58	22.94
✓	✓						22.80	24.67	24.96
✓	✓	✓	✓				24.75	27.35	27.03
✓	✓	✓	✓	✓			27.81	28.03	28.59
		✓	✓	✓	✓		21.72	22.47	23.60
✓	✓	✓	✓	✓		✓	26.49	27.25	27.63
✓	✓	✓			✓		24.21	27.64	27.29
✓	✓		✓	✓	✓		28.03	28.65	28.37
✓	✓	✓	✓	✓	✓		**29.12**	**29.95**	**29.10**

Table 2. **Ablation.** We ablate each proposed term on IKEA ASM [5]. All proposed terms consistently improve performance.

4.2. Ablation Studies

In Table 2, we provide an ablation study to demonstrate the influence of each design choice of **VAVA** on the accuracy of action phase classification. *Intra-Video* and *Inter-Video* denote the effect of the contrastive loss terms we introduced in Eq. 14 and Eq. 15 to regularize the training process. *KL* shows the effect of KL divergence regularization term. While *Consistency Prior* denotes the temporal prior, introduced in Eq. 4, that enforces time consistency across videos during alignment, *Optimality Prior* denotes the temporal prior introduced in Eq. 6 that favors optimal matching of frames across videos. *Virtual Frame* shows the effect of extra virtual frame we incorporated in the optimal transport formulation to address background and redundant frames. We further compare our *Virtual Frame* strategy to a *Threshold* approach, in which alignments with a low matching score are removed based on a tuned threshold.

As shown in Table 2, all of our design choices consistently improve the accuracy of our algorithm. *Optimality Prior* tackles variations in the sequence order, whereas *Consistency Prior* allows for respecting the coarse-level temporal structure and consistency of videos. While they both individually improve the performance, the Gaussian Mixture Model that combines the two further boosts the accuracy, which demonstrates the complementary nature of each prior. We further demonstrate that *Virtual Frame* strategy significantly improves performance as compared to a model that does not include it and a model that uses a simpler thresholding based approach (*Threshold*) to handle background frames. We also evaluate the influence of the *Intra-Video* and *Inter-Video* contrastive loss terms and demonstrate that they result in superior performance, by regularizing the self-supervised learning process. Besides, the KL divergence loss that encourages smooth alignment further improves the performance. We present qualitative results of frame retrieval, in which we match the most similar frame with a given query frame, in Fig. 7. As shown on this example, **VAVA** is able to reliably align both regular action frames and background frames.

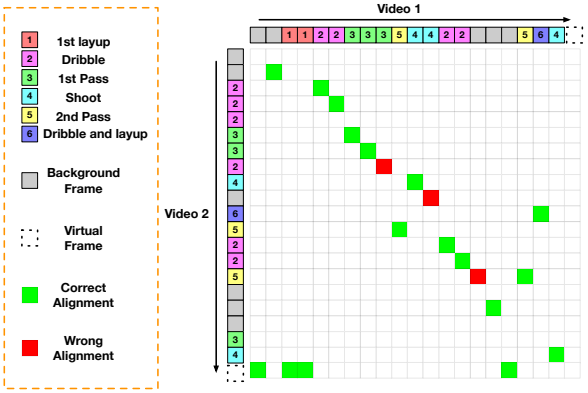


Figure 8. **Example Alignment.** We align two videos from the *Attend NBA Skills Challenge* task of the COIN dataset. For each frame in video 1, we align it with the optimal match in the other sequence. Correct alignment means that the action frame is aligned with another frame with the same action. Redundant frames are aligned to the virtual frame and the background frames are aligned to either another background frame or the virtual frame. As can be seen by the high number of correct matches, our model can reliably align two sequences with temporal variations.

To demonstrate that our approach is able to align sequential actions in unconstrained environments, we visualize the assignment matrix for a representative example on the COIN dataset, that feature different temporal variations involving *background*, *redundant* and *non-monotonic* frames. As shown in Fig. 8, our model is able to align such sequences with high accuracy and brings in robustness against temporal variations, which makes it suitable for aligning sequential actions in-the-wild.

5. Conclusion

In this paper, we propose a self-supervised learning framework that uses video alignment as a proxy task. The proposed VAVA approach is able to align sequential actions in-the-wild with an optimal transport based sequence alignment formulation. We further propose to enforce adaptive temporal priors on optimal transport, which efficiently handles temporal variations. Our experiments show that VAVA outperforms the state-of-the-art on the Pouring, Penn Action, IKEA ASM and COIN dataset. Our future work will explore applications of video alignment for AR-based task guidance and procedure learning.

Acknowledgments This work was funded by Microsoft and in part by the Swiss National Science Foundation.

References

[1] Fritz Albregtsen. Statistical texture measures computed from gray level coocurrence matrices. *Image processing laboratory, department of informatics, university of oslo*, 2008. 6

[2] Theodore Wilbur Anderson. An introduction to multivariate statistical analysis. *Wiley New York*, 1958. 3

[3] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, 2015. 3

[4] Nadine Behrmann, Mohsen Fayyaz, Juergen Gall, and Mehdi Noroozi. Long Short View Feature Decomposition via Contrastive Video Representation Learning. In *International Conference on Computer Vision*, 2021. 3

[5] Yizhak Ben-Shabat, Xin Yu, Fatemeh Sadat 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 *arXiv Preprint*, 2020. 6, 8

[6] Yoshua Bengio and James Bergstra. Slow, Decorrelated Features for Pretraining Complex Cell-Like Networks. In *Advances in Neural Information Processing Systems*, 2009. 3

[7] Donald J. Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, 1994. 3

[8] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Conference on Computer Vision and Pattern Recognition*, 2017. 1

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Learning Representations*, 2020. 3

[10] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and Attend: Video Domain Adaptation. In *European Conference on Computer Vision*, 2020. 3

[11] Huseyin Coskun, Zeeshan Zia, Bugra Tekin, Federica Bogo, Nassir Navab, Federico Tombari, and Harpreet Sawhney. Domain-Specific Priors and Meta Learning for Low-shot First-Person Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[12] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013. 4, 6

[13] Ali Diba, Vivek Sharma, Luc Van Gool, and Rainer Stiefelhagen. Dynamonet: Dynamic Action and Motion Network. In *International Conference on Computer Vision*, 2019. 3

[14] Ali Diba, Vivek Sharma, Reza Safdari, Dariush Lotfi, M. Saquib Sarfraz, Rainer Stiefelhagen, and Luc Van Gool. Vi2CLR: Video and Image for Visual Contrastive Learning of Representation. In *International Conference on Computer Vision*, 2021. 3

[15] Pelin Dogan, Boyang Li, Leonid Sigal, and Markus Gross. A Neural Multi-sequence Alignment TeCHnique (NeuMATCH). In *Conference on Computer Vision and Pattern Recognition*, 2018. 1

[16] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal Cycle-Consistency Learning. In *Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 3, 7

[17] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning of Visual Representations. In *International Conference on Computer Vision*, 2021. 3

[18] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning. In *Conference on Computer Vision and Pattern Recognition*, 2021. 3

[19] Zeyu Feng, Chang Xu, and Dacheng Tao. Self-Supervised Representation Learning by Rotation Feature Decoupling. In *Conference on Computer Vision and Pattern Recognition*, 2019. 3

[20] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-Supervised Video Representation Learning with Odd-One-Out Networks. In *Conference on Computer Vision and Pattern Recognition*, 2017. 3

[21] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Predicting the Future: A Jointly Learnt Model for Action Anticipation. In *International Conference on Computer Vision*, 2019. 3

[22] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. OTA: Optimal Transport Assignment for Object Detection. In *Conference on Computer Vision and Pattern Recognition*, 2021. 3

[23] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised Representation Learning by Predicting Image Rotations. In *International Conference on Learning Representations*, 2018. 3

[24] Ross Goroshin, Joan Bruna, Jonathan Tompson, David Eigen, and Yann LeCun. Unsupervised Learning of Spatiotemporally Coherent Metrics. In *International Conference on Computer Vision*, 2015. 3

[25] Isma Hadji, Konstantinos G. Derpanis, and Allan D. Jepson. Representation Learning via Global Temporal Alignment and Cycle-Consistency. In *Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2, 3, 7

[26] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality Reduction by Learning an Invariant Mapping. In *Conference on Computer Vision and Pattern Recognition*, 2006. 3

[27] Tengda Han, Weidi Xie, and Andrew Zisserman. Video Representation Learning by Dense Predictive Coding. In *International Conference on Computer Vision Workshops*, 2019. 3

[28] Sanjay Haresh, Sateesh Kumar, Huseyin Coskun, Shahram Najam Syed, Andrey Konin, Muhammad Zeeshan Zia, and Quoc-Huy Tran. Learning by Aligning Videos in Time. In *Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2, 3, 6, 7, 8

[29] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *Conference on Computer Vision and Pattern Recognition*, 2020. 3

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Conference on Computer Vision and Pattern Recognition*, 2016. 7

[31] Geoffrey E. Hinton and Richard S. Zemel. Autoencoders, Minimum Description Length and Helmholtz Free Energy. In *Advances in Neural Information Processing Systems*, 1994. 3

[32] Klemen Kotar, Gabriel Ilharco, Ludwig Schmidt, Kiana Ehsani, and Roozbeh Mottaghi. Contrasting Contrastive Self-Supervised Representation Learning Pipelines. In *International Conference on Computer Vision*, 2021. 3

[33] Taein Kwon, Bugra Tekin, Jan Stuhmer, Federica Bogo, and Marc Pollefeys. H2O: Two Hands Manipulating Objects for First Person Interaction Recognition. In *International Conference on Computer Vision*, 2021. 1

[34] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning Representations for Automatic Colorization. In *European Conference on Computer Vision*, 2016. 3

[35] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a Proxy Task for Visual Understanding. In *Conference on Computer Vision and Pattern Recognition*, 2017. 3

[36] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised Representation Learning by Sorting Sequences. In *International Conference on Computer Vision*, 2017. 3

[37] Ruibo Li, Guosheng Lin, and Lihua Xie. Self-Point-Flow: Self-Supervised Scene Flow Estimation from Point Clouds with Optimal Transport and Random Walk. In *Conference on Computer Vision and Pattern Recognition*, 2021. 3

[38] Yuanze Lin, Xun Guo, and Yan Lu. Self-Supervised Video Representation Learning with Meta-Contrastive Network. In *International Conference on Computer Vision*, 2021. 3

[39] Weizhe Liu, David Ferstl, Samuel Schulter, Lukas Zebedin, Pascal Fua, and Christian Leistner. Domain Adaptation for Semantic Segmentation via Patch-Wise Contrastive Learning. In *arXiv Preprint*, 2021. 3

[40] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and Learn: Unsupervised Learning Using Temporal Order Verification. In *European Conference on Computer Vision*, 2016. 3, 7

[41] Hossein Mobahi, Ronan Collobert, and Jason Weston. Deep Learning from Temporal Coherence in Video. In *International Conference on Machine Learning*, 2009. 3

[42] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation Learning by Learning to Count. In *International Conference on Computer Vision*, 2017. 3

[43] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. VideoMoCo: Contrastive Video Representation Learning with Temporally Adversarial Examples. In *Conference on Computer Vision and Pattern Recognition*, 2021. 3

[44] Gilles Puy, Alexandre Boulch, and Renaud Marlet. FLOT: Scene Flow on Point Clouds guided by Optimal Transport. In *European Conference on Computer Vision*, 2020. 3

[45] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal Contrastive Video Representation Learning. In *Conference on Computer Vision and Pattern Recognition*, 2021. 3

[46] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning Feature Matching with Graph Neural Networks. In *Conference on Computer Vision and Pattern Recognition*, 2020. 3

[47] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. Time-Contrastive Networks: Self-Supervised Learning from Video. In *International Conference on Robotics and Automation*, 2018. 3, 6, 7

[48] Mathieu Serrurier, Franck Mamalet, Alberto González-Sanz, Thibaut Boissin, Jean-Michel Loubes, and Eustasio del Barrio. Achieving robustness in classification using optimal transport with hinge regularization. In *Conference on Computer Vision and Pattern Recognition*, 2021. 3

[49] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised Learning of Video Representations Using LSTMs. In *International Conference on Machine Learning*, 2015. 3

[50] Bing Su and Gang Hua. Order-preserving Wasserstein Distance for Sequence Matching. In *Conference on Computer Vision and Pattern Recognition*, 2017. 3, 6

[51] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. COIN: A Large-scale Dataset for Comprehensive Instructional Video Analysis . In *Conference on Computer Vision and Pattern Recognition*, 2019. 6

[52] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning Spatiotemporal Features with 3d Convolutional Networks. In *Conference on Computer Vision and Pattern Recognition*, 2018. 1

[53] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *Conference on Computer Vision and Pattern Recognition*, 2018. 1

[54] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating Videos with Scene Dynamics. In *Advances in Neural Information Processing Systems*, 2016. 3

[55] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Conference on Computer Vision and Pattern Recognition*, 2018. 1

[56] Donglai Wei, Joseph Lim, Andrew Zisserman, and William T. Freeman. Learning and Using the Arrow of Time. In *Conference on Computer Vision and Pattern Recognition*, 2018. 3

[57] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-Supervised Spatiotemporal Learning via Video Clip Order Prediction. In *Conference on Computer Vision and Pattern Recognition*, 2019. 3

[58] Renjun Xu, Pelen Liu, Liyan Wang, Chao Chen, and Jindong Wang. Reliable Weighted Optimal Transport for Unsupervised Domain Adaptation. In *Conference on Computer Vision and Pattern Recognition*, 2020. 3

[59] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta.

Multimodal Contrastive Training for Visual Representation Learning. In *Conference on Computer Vision and Pattern Recognition*, 2021. 3

[60] Weiyu Zhang, Menglong Zhu, and Konstantinos G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *International Conference on Computer Vision*, 2013. 6

[61] Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Weakly Supervised Contrastive Learning. In *International Conference on Computer Vision*, 2021. 3

[62] Will Zou, Shenghuo Zhu, Kai Yu, and Andrew Y. Ng. Deep Learning of Invariant Features via Simulated Fixations in Video. In *Advances in Neural Information Processing Systems*, 2012. 3

[63] Will Y. Zou, Andrew Y. Ng, and Kai Yu. Unsupervised Learning of Visual Invariance with Temporal Coherence. In *Advances in Neural Information Processing Systems Workshops*, 2012. 3