Novelty Assessment Report

Paper: REALIGN: Regularized Procedure Alignment with Matching Video Embeddings via Partial Gromov-Wasserstein Optimal Transport

PDF URL: https://openreview.net/pdf?id=kop52LaSAB

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-27

Abstract

Learning from procedural videos remains a core challenge in self-supervised representation learning, as real-world instructional data often contains background segments, repeated actions, and steps presented out of order. Such variability violates the strong monotonicity assumptions underlying many alignment methods. Prior state-of-the-art approaches, such as OPEL and RGWOT, leverage Kantorovich Optimal Transport (KOT) and Gromov-Wasserstein Optimal Transport (GWOT) to build frame-to-frame correspondences but operate only on local feature similarity and pairwise relational structure, without explicit temporal priors, which limits their ability to capture the higher-order temporal structure of a task. In this paper, we introduce **REALIGN**, an unsupervised framework for procedure learning based on Regularized Fused Partial Gromov-Wasserstein Optimal Transport (R-FPGWOT). In contrast to RGWOT, our formulation jointly models visual correspondences and temporal relations under a partial alignment scheme, enabling robust handling of irrelevant frames, repeated actions, and non-monotonic step orders common in instructional videos. To stabilize training, we integrate FPGWOT distances with inter-sequence contrastive learning, avoiding the need for multiple regularizers and preventing collapse to degenerate solutions. Across egocentric (EgoProceL) and third-person (ProceL, CrossTask) benchmarks, REALIGN achieves up to **18.9\% (7.62pp)** average F1-score improvements and over **30\% (7.74pp)** temporal IoU gains, while producing more interpretable transport maps that preserve key-step orderings and filter out noise.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Unsupervised Procedure Learning from Instructional Videos**

A total of **48 papers** were analyzed and organized into a taxonomy with **24 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Temporal Alignment and Correspondence Learning**
- **Step Discovery and Segmentation**
- **Procedure Representation and Task Modeling**
- **Pretraining and Representation Learning**
- **Procedure Planning and Goal-Directed Reasoning**
- **Weakly Supervised and Cross-Task Learning**
- **Specialized Procedure Understanding Tasks**
- **Auxiliary Methods and Resources**
- **Automatic Procedure Learning from Web Videos**

Complete Taxonomy Tree

- Unsupervised Procedure Learning from Instructional Videos Survey Taxonomy
- Temporal Alignment and Correspondence Learning
  - Optimal Transport-Based Alignment ★ (2 papers)
  - [0] REALIGN: Regularized Procedure Alignment with Matching Video Embeddings via Partial Gromov-Wasserstein Optimal Transport (Anon et al., 2026) View paper
  - [7] Unsupervised procedure learning via joint dynamic summarization (Ehsan Elhamifar, 2019) View paper
  - Embedding-Based Correspondence (3 papers)
  - [8] Learning procedure-aware video representation from instructional videos and their narrations (Yiwu Zhong, 2023) View paper
  - [29] Time-contrastive networks: Self-supervised learning from video (Pierre Sermanet, 2018) View paper
  - [43] AVLnet: Learning Audio-Visual Language Representations from Instructional Videos (Rouditchenko, 2021) View paper
  - Cross-Modal and Multi-Cue Alignment (3 papers)
  - [2] Unsupervised learning from narrated instruction videos (Jean-Baptiste Alayrac, 2016) View paper
  - [14] Achieving procedure-aware instructional video correlation learning under weak supervision from a collaborative perspective (Tianyao He, 2025) View paper
  - [32] Learning to ground instructional articles in videos through narrations (Effrosyni Mavroudi, 2023) View paper
- Step Discovery and Segmentation
  - Autoregressive and Sequential Models (3 papers)
  - [4] Unsupervised discovery of actions in instructional videos (AJ Piergiovanni, 2021) View paper
  - [5] Unsupervised action segmentation for instructional videos (Piergiovanni, 2021) View paper
  - [20] Unsupervised learning and segmentation of complex activities from video (Fadime Å□ener, 2018) View paper
  - Self-Supervised Contrastive Segmentation (3 papers)
  - [6] Stepformer: Self-supervised step discovery and localization in instructional videos (Nikita Dvornik, 2023) View paper
  - [16] Steps: Self-supervised key step extraction and localization from unlabeled procedural videos (Anshul Shah, 2023) View paper

  - [37] STEPs: Self-Supervised Key Step Extraction from Unlabeled Procedural Videos (Shah, 2023) View paper
  - Graph-Based Step Discovery (2 papers)
  - [11] United we stand, divided we fall: Unitygraph for unsupervised procedure learning from videos (Siddhant Bansal, 2024) View paper
  - [39] SVGraph: Learning Semantic Graphs from Instructional Videos (Madeline C. Schiappa, 2022) View paper
- Procedure Representation and Task Modeling
  - Task Graph Learning (3 papers)
  - [12] Video-mined task graphs for keystep recognition in instructional videos (Ashutosh Kumar, 2023) View paper
  - [17] Differentiable task graph learning: Procedural activity representation and online mistake detection from egocentric videos (Farinella, 2024) View paper
  - [18] Unsupervised task graph generation from instructional video transcripts (Logeswaran, 2023) View paper
  - State-Action Dynamics (3 papers)
  - [21] Procedure Planning in Instructional Videos (C. Chang, 2022) View paper
  - [27] PDPP: Projected Diffusion for Procedure Planning in Instructional Videos (Hanlin Wang, 2023) View paper
  - [35] Skip-Plan: Procedure Planning in Instructional Videos via Condensed Action Space Learning (Zhiheng Li, 2023) View paper
  - Hierarchical Procedure Encoding (2 papers)
  - [28] Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation (Yuan Kun, 2024) View paper
  - [46] Hierarchical Modeling for Task Recognition and Action Segmentation in Weakly-Labeled Instructional Videos (Reza Ghoddoosian, 2021) View paper
- Pretraining and Representation Learning
  - Procedure-Aware Pretraining (2 papers)
  - [10] Procedure-aware pretraining for instructional video understanding (Honglu Zhou, 2023) View paper
  - [22] Efficient Pre-training for Localized Instruction Generation of Procedural Videos (Anil Batra, 2024) View paper
  - Multi-Task and Multi-Modal Pretraining (2 papers)
  - [1] Self-supervised multi-task procedure learning from instructional videos (Ehsan Elhamifar, 2020) View paper
  - [31] Eagle: Egocentric aggregated language-video engine (Jing Bi, 2024) View paper
  - General Video Representation Learning (3 papers)
  - [36] Unsupervised Learning Layers for Video Analysis (Zhao Liang, 2022) View paper
  - [41] Dense Unsupervised Learning for Video Segmentation (Nikita Araslanov, 2021) View paper
  - [48] Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning (Lotter, 2016) View paper
- Procedure Planning and Goal-Directed Reasoning
  - Deterministic Sequence Planning (1 papers)
  - [44] Procedure Planning in Instructional Videos via Contextual Modeling and Model-based Policy Learning (Jing Bi, 2021) View paper
  - Knowledge-Augmented Planning (1 papers)
  - [25] Why not use your textbook? knowledge-enhanced procedure planning of instructional videos (Kumaranage Ravindu Yasas Nagasinghe, 2024) View paper
- Weakly Supervised and Cross-Task Learning
  - Narration-Guided Weak Supervision (1 papers)
  - [42] Learning from Narrated Instruction Videos. (Jean-Baptiste Alayrac, 2019) View paper
  - Ordered-List Weak Supervision (1 papers)
  - [45] Cross-task weakly supervised learning from instructional videos (Dimitri Zhukov, 2019) View paper
- Specialized Procedure Understanding Tasks
  - Error Detection and Task Assessment (2 papers)
  - [23] Error detection in egocentric procedural task videos (Shihâ□□Po Lee, 2024) View paper
  - [26] Error recognition in procedural videos using generalized task graph (SP Lee, 2025) View paper
  - Reference Resolution (1 papers)
  - [13] Unsupervised visual-linguistic reference resolution in instructional videos (De-An Huang, 2017) View paper
  - Structured Knowledge Extraction (1 papers)
  - [38] A benchmark for structured procedural knowledge extraction from cooking videos (Frank F. Xu, 2020) View paper
- Auxiliary Methods and Resources
  - Dataset Construction and Alignment (2 papers)
  - [15] Ht-step: Aligning instructional articles with how-to videos (T Afouras, 2023) View paper
  - [40] A recipe for creating multimodal aligned datasets for sequential tasks (Lin, 2020) View paper
  - Fine-Grained Activity Recognition (2 papers)
  - [9] Learning to recognize procedural activities with distant supervision (Xudong Lin, 2022) View paper
  - [30] Procedure completion by learning from partial summaries (Ehsan Elhamifar, 2020) View paper
  - General Unsupervised Video Analysis (4 papers)
  - [19] Unsupervised learning of procedures from demonstration videos (K Goel, 2018) View paper
  - [24] Unsupervised learning of event classes from video (Muralikrishna Sridhar, 2010) View paper
  - [34] Instructional videos for unsupervised harvesting and learning of action examples (Shoou-I Yu, 2014) View paper
  - [47] Joint Visual-Temporal Embedding for Unsupervised Learning of Actions in Untrimmed Sequences (VidalMata, 2021) View paper
  - Unrelated Applications (1 papers)
  - [33] Comparison of Machine Learning algorithms on detecting the confusion of students while watching MOOCs (VASM Anala, 2022) View paper
- Automatic Procedure Learning from Web Videos (1 papers)
  - [3] Towards automatic learning of procedures from web instructional videos (Luowei Zhou, 2018) View paper

Narrative

Core task: unsupervised procedure learning from instructional videos. The field aims to extract structured procedural knowledge—such as step sequences, temporal boundaries, and task dependencies—from large collections of unlabeled how-to videos. The taxonomy reflects a diverse landscape organized around several complementary challenges. Temporal Alignment and Correspondence Learning focuses on matching video segments across demonstrations, often using techniques like optimal transport to align steps without explicit labels. Step Discovery and Segmentation addresses the problem of identifying meaningful action boundaries and clustering them into coherent steps, as seen in works like Action Discovery[4] and Action Segmentation[5]. Procedure Representation and Task Modeling emphasizes building graph-based or hierarchical structures that capture dependencies and ordering constraints, with approaches such as Task Graphs[12] and Differentiable Task Graph[17]. Pretraining and Representation Learning explores self-supervised objectives that yield embeddings sensitive to procedural structure, while Procedure Planning and Goal-Directed Reasoning targets the synthesis of step sequences for novel goals. Weakly Supervised and Cross-Task Learning leverages partial annotations or transfers knowledge across related tasks, and Specialized Procedure Understanding Tasks tackle domain-specific challenges like error detection or localized instruction generation. Auxiliary Methods and Resources provide datasets and supporting techniques, and Automatic Procedure Learning from Web Videos scales these ideas to noisy, in-the-wild data.

Several active lines of work highlight key trade-offs and open questions. One thread pursues robust temporal alignment methods that can handle high variability across demonstrations, balancing computational efficiency with alignment quality. Another explores how to discover and segment steps in a fully unsupervised manner, often debating whether to rely on clustering in learned feature spaces or to impose stronger structural priors. Graph-based representations have gained traction for capturing task dependencies, yet questions remain about how to learn these graphs from raw video without ground-truth annotations. REALIGN[0] sits within the Temporal Alignment and Correspondence Learning branch, specifically under Optimal Transport-Based Alignment, and shares methodological kinship with Dynamic Summarization[7], which also addresses correspondence across video segments. Compared to earlier alignment work like Narrated Instruction Learning[2] or Automatic Procedure Learning[3], REALIGN[0] emphasizes principled transport-based matching to handle diverse procedural variations, positioning it as a recent refinement in the ongoing effort to align instructional content at scale.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Unsupervised procedure learning via joint dynamic summarization

Authors: Ehsan Elhamifar, Zwe Naing | **Year/Venue**: 2019 | **URL**: View paper

Abstract

We address the problem of unsupervised procedure learning from unconstrained instructional videos. Our goal is to produce a summary of the procedure key-steps and their ordering needed to perform a given task, as well as localization of the key-steps in videos. We develop a collaborative sequential subset selection framework, where we build a dynamic model on videos by learning states and transitions between them, where states correspond to different subactivities, including background and proce...

Relationship Analysis

Both papers belong to the Optimal Transport-Based Alignment category, using optimal transport formulations to align procedural sequences in instructional videos. While the original paper (REALIGN) employs Regularized Fused Partial Gromov-Wasserstein Optimal Transport with virtual sink nodes to handle background frames and non-monotonic orderings, the candidate paper uses a joint dynamic summarization framework with Hidden Markov Models and standard subset selection to discover key-steps and their ordering. The key difference is that REALIGN focuses on partial transport with explicit structural priors for frame-level alignment, whereas the candidate paper emphasizes collaborative summarization across multiple videos using HMM-based state transitions without partial matching.

Contributions Analysis

Overall novelty summary. The paper introduces REALIGN, a framework based on Regularized Fused Partial Gromov-Wasserstein Optimal Transport for unsupervised procedure learning from instructional videos. It resides in the Optimal Transport-Based Alignment leaf, which contains only two papers including this one. This leaf sits within the broader Temporal Alignment and Correspondence Learning branch, which encompasses three distinct methodological approaches. The sparse population of this specific leaf suggests that optimal transport formulations for procedural alignment remain relatively underexplored compared to embedding-based or cross-modal methods.

The taxonomy reveals that Temporal Alignment and Correspondence Learning is one of eight major research directions in the field. Neighboring branches include Step Discovery and Segmentation, which focuses on identifying action boundaries without alignment, and Procedure Representation and Task Modeling, which builds structured task graphs. The scope note for Optimal Transport-Based Alignment explicitly excludes contrastive or embedding methods, positioning REALIGN within a methodologically distinct subfield. The broader Temporal Alignment branch contains eleven papers across three leaves, indicating moderate activity in correspondence learning overall but concentration in embedding-based approaches rather than transport-based formulations.

Among ten candidates examined, the core REALIGN framework contribution shows one refutable candidate from six examined, while the unified alignment loss contribution also identifies one refutable candidate from four examined. The partial alignment scheme contribution was not tested against any candidates. The limited search scope—ten total candidates rather than an exhaustive survey—means these statistics reflect only top-K semantic matches and immediate citations. The presence of refutable candidates for two of three contributions suggests some overlap with prior work in the examined sample, though the scale of examination leaves substantial uncertainty about the broader literature landscape.

Given the sparse population of the Optimal Transport-Based Alignment leaf and the limited search scope, the analysis captures a narrow slice of potentially relevant work. The taxonomy structure indicates this is a methodologically specialized area within a diverse field, but the ten-candidate examination cannot definitively characterize novelty relative to the full literature. The refutable candidates identified represent overlaps within the examined sample, not comprehensive prior art assessment.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: REALIGN framework based on Regularized Fused Partial Gromov-Wasserstein Optimal Transport

Description: The authors propose REALIGN, a novel unsupervised procedure learning framework that extends Fused Gromov-Wasserstein Optimal Transport with partial alignment constraints. This formulation jointly models visual correspondences and temporal relations while enabling robust handling of irrelevant frames, repeated actions, and non-monotonic step orders common in instructional videos.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. THESAURUS: Contrastive Graph Clustering by Swapping Fused Gromov-Wasserstein Couplings

URL: View paper

Brief Assessment

THESAURUS[50] applies Fused Gromov-Wasserstein OT to graph node clustering, not to unsupervised procedure learning from instructional videos. The technical domains and problem formulations are fundamentally different.

2. Weakly-Supervised Temporal Action Alignment Driven by Unbalanced Spectral Fused Gromov-Wasserstein Distance

URL: View paper

Brief Assessment

Spectral Fused Distance[51] addresses weakly-supervised temporal action alignment using unbalanced spectral fused Gromov-Wasserstein distance, while REALIGN focuses on unsupervised procedure learning with regularized fused partial Gromov-Wasserstein optimal transport. The candidate requires video-text pairs for training, whereas REALIGN operates purely on visual data without textual supervision.

3. A Fused Gromov-Wasserstein Framework for Unsupervised Knowledge Graph Entity Alignment

URL: View paper

Brief Assessment

Entity Alignment[53] applies Fused Gromov-Wasserstein to knowledge graph entity alignment across different KGs, not to procedure learning from instructional videos. The domains, problem formulations, and technical objectives are fundamentally different.

4. A Fused Gromov-Wasserstein Approach to Subgraph Contrastive Learning

URL: View paper

Brief Assessment

Subgraph Contrastive Learning[52] applies Fused Gromov-Wasserstein to graph representation learning with node-level and subgraph-level contrastive tasks, not to video procedure alignment with partial transport constraints for handling background frames and temporal irregularities.

5. Procedure learning via regularized gromov-wasserstein optimal transport

URL: View paper

Prior Art Analysis

Regularized Gromov Wasserstein[49] demonstrates that a fused Gromov-Wasserstein optimal transport formulation with structural priors for procedure learning was already proposed and published. The candidate paper presents a 'fused gromov-wasserstein optimal transport (fgwot)' framework that 'fuses the kot and gwot objectives' and uses structural priors for temporal video alignment. This directly overlaps with the ORIGINAL paper's claim of introducing REALIGN based on 'Regularized Fused Partial Gromov-Wasserstein Optimal Transport'. While the ORIGINAL adds partial alignment constraints, the core fused GWOT formulation with structural priors was already established in the candidate work.

Evidence

Evidence 1 - **Rationale**: Both papers propose frameworks using fused Gromov-Wasserstein optimal transport with structural priors for procedure learning, indicating the candidate established this approach prior to the original's partial extension. - **Original**: we introducerealign, an unsupervised framework for procedure learning based onregularized fused partial gromov-wasserstein optimal transport(r-fpgwot). in contrast to rgwot, our formulation jointly models visual correspondences and temporal relations under a partial alignment scheme - **Candidate**: we propose a self-supervised framework, which utilizes a fused gromov-wasserstein optimal transport with a structural prior for frame-to-frame mapping

Evidence 2 - **Rationale**: The ORIGINAL paper cites 'rgwot' as prior work, and the candidate paper is titled 'Procedure learning via regularized gromov-wasserstein optimal transport' (Regularized Gromov Wasserstein[49]), establishing that the candidate is the rgwot work referenced as prior art. - **Original**: prior state-of-the-art approaches, such as opel and rgwot, leverage kantorovich optimal transport (kot) and gromov-wasserstein optimal transport (gwot) to build frame-to-frame correspondences - **Candidate**: we introduce an optimal transport-based framework for self-supervised procedure learning, which adopts a fused gromov-wasserstein optimal transport formulation with a structural prior for frame-to-frame correspondences

6. Temporally Consistent Unbalanced Optimal Transport for Unsupervised Action Segmentation

URL: View paper

Brief Assessment

Temporally Consistent Transport[54] focuses on unsupervised action segmentation in videos using unbalanced Gromov-Wasserstein optimal transport, whereas REALIGN addresses procedure learning from instructional videos with partial alignment constraints for handling background frames and non-monotonic step orders. The tasks and technical formulations differ substantially.

Contribution 2: Partial alignment scheme with virtual sink node for handling background and redundant frames

Description: The method introduces a partial transport formulation that relaxes balanced marginal constraints by incorporating a virtual sink node. This allows irrelevant or background frames to be mapped to a null mass instead of being forced into spurious correspondences, addressing a key limitation of prior fully balanced optimal transport methods.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Contribution 3: Unified alignment loss integrating temporal priors and contrastive regularization

Description: The authors develop a unified loss function that combines FPGWOT distances with Laplace-shaped temporal priors, structural regularization, and inter-sequence contrastive learning. This integration stabilizes training by avoiding degenerate solutions and preventing collapse to trivial mappings without requiring multiple separate regularizers.

This contribution was assessed against **4 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Optimal transport guided contrastive video summarization

URL: View paper

Brief Assessment

Optimal Transport Summarization[57] focuses on video summarization tasks using optimal transport for contrastive learning, not on procedure learning with temporal priors and structural regularization for alignment stabilization.

2. Timestamp-supervised wearable-based activity segmentation and recognition with contrastive learning and order-preserving optimal transport

URL: View paper

Brief Assessment

Timestamp Supervised Segmentation[55] focuses on wearable activity segmentation with timestamp supervision, not video procedure alignment. The candidate uses contrastive learning and optimal transport for activity recognition tasks, which differs fundamentally from the original's video alignment framework.

3. Procedure learning via regularized gromov-wasserstein optimal transport

URL: View paper

Prior Art Analysis

Regularized Gromov Wasserstein[49] demonstrates prior work combining temporal priors with contrastive regularization in a unified loss for optimal transport-based alignment. The candidate paper presents 'contrastive inverse difference moment (c-idm) as a regularization' combined with temporal priors in their training objective. The ORIGINAL paper claims novelty in 'integrating temporal smoothness, optimal regularization, and a novel inter-video contrastive term', but the candidate already combined temporal priors with contrastive regularization (c-idm) to prevent degenerate solutions, establishing this integration approach before the ORIGINAL work.

Evidence

Evidence 1 - **Rationale**: Both papers integrate contrastive regularization with optimal transport to prevent degenerate solutions, with the candidate establishing this approach first. - **Original**: we design a unified alignment loss that integrates temporal smoothness, optimal regularization, and a novel inter-video contrastive term, preventing degenerate matches and improving stability in ot-based training - **Candidate**: to avoid trivial solutions, we incorporate contrastive inverse difference moment (c-idm) [35] as a regularization, which is applied separately on frame embeddingsx andy, yielding our regularized gromov-wasserstein optimal transport framework (rgwot)

Evidence 2 - **Rationale**: The candidate paper explicitly describes using a unified loss combining temporal priors and contrastive regularization to prevent collapse, which is the same integration claimed as novel by the ORIGINAL. - **Original**: to further improve stability, we integrate these temporal smoothness priors and the c-idm regularizer into a unified loss, which prevents degenerate collapse of all frames into a single cluster - **Candidate**: our regularized gromov-wasserstein optimal transport (rgwot) approach uses a unified loss with a purpose-aligned regularization, avoiding the difficulty of balancing multiple losses and conflicting regularizations

4. Multi-granularity correspondence learning from long-term noisy videos

URL: View paper

Brief Assessment

Multigranularity Correspondence[56] focuses on video-text alignment using optimal transport for multi-granularity correspondence learning, not on general RL frameworks with temporal priors and contrastive regularization as in the original paper's procedural video alignment context.

Appendix: Text Similarity Detection

Textual similarity detection checked 10 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Procedure learning via regularized gromov-wasserstein optimal transport

Detected in: Contribution: contribution_1, Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] REALIGN: Regularized Procedure Alignment with Matching Video Embeddings via Partial Gromov-Wasserstein Optimal Transport View paper
- [1] Self-supervised multi-task procedure learning from instructional videos View paper
- [2] Unsupervised learning from narrated instruction videos View paper
- [3] Towards automatic learning of procedures from web instructional videos View paper
- [4] Unsupervised discovery of actions in instructional videos View paper
- [5] Unsupervised action segmentation for instructional videos View paper
- [6] Stepformer: Self-supervised step discovery and localization in instructional videos View paper
- [7] Unsupervised procedure learning via joint dynamic summarization View paper
- [8] Learning procedure-aware video representation from instructional videos and their narrations View paper
- [9] Learning to recognize procedural activities with distant supervision View paper
- [10] Procedure-aware pretraining for instructional video understanding View paper
- [11] United we stand, divided we fall: Unitygraph for unsupervised procedure learning from videos View paper
- [12] Video-mined task graphs for keystep recognition in instructional videos View paper
- [13] Unsupervised visual-linguistic reference resolution in instructional videos View paper
- [14] Achieving procedure-aware instructional video correlation learning under weak supervision from a collaborative perspective View paper
- [15] Ht-step: Aligning instructional articles with how-to videos View paper
- [16] Steps: Self-supervised key step extraction and localization from unlabeled procedural videos View paper
- [17] Differentiable task graph learning: Procedural activity representation and online mistake detection from egocentric videos View paper
- [18] Unsupervised task graph generation from instructional video transcripts View paper
- [19] Unsupervised learning of procedures from demonstration videos View paper
- [20] Unsupervised learning and segmentation of complex activities from video View paper
- [21] Procedure Planning in Instructional Videos View paper
- [22] Efficient Pre-training for Localized Instruction Generation of Procedural Videos View paper
- [23] Error detection in egocentric procedural task videos View paper
- [24] Unsupervised learning of event classes from video View paper
- [25] Why not use your textbook? knowledge-enhanced procedure planning of instructional videos View paper
- [26] Error recognition in procedural videos using generalized task graph View paper
- [27] PDPP: Projected Diffusion for Procedure Planning in Instructional Videos View paper
- [28] Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation View paper

- [29] Time-contrastive networks: Self-supervised learning from video View paper
- [30] Procedure completion by learning from partial summaries View paper
- [31] Eagle: Egocentric aggregated language-video engine View paper
- [32] Learning to ground instructional articles in videos through narrations View paper
- [33] Comparison of Machine Learning algorithms on detecting the confusion of students while watching MOOCs View paper
- [34] Instructional videos for unsupervised harvesting and learning of action examples View paper
- [35] Skip-Plan: Procedure Planning in Instructional Videos via Condensed Action Space Learning View paper
- [36] Unsupervised Learning Layers for Video Analysis View paper
- [37] STEPs: Self-Supervised Key Step Extraction from Unlabeled Procedural Videos View paper
- [38] A benchmark for structured procedural knowledge extraction from cooking videos View paper
- [39] SVGraph: Learning Semantic Graphs from Instructional Videos View paper
- [40] A recipe for creating multimodal aligned datasets for sequential tasks View paper
- [41] Dense Unsupervised Learning for Video Segmentation View paper
- [42] Learning from Narrated Instruction Videos. View paper
- [43] AVLnet: Learning Audio-Visual Language Representations from Instructional Videos View paper
- [44] Procedure Planning in Instructional Videos via Contextual Modeling and Model-based Policy Learning View paper
- [45] Cross-task weakly supervised learning from instructional videos View paper
- [46] Hierarchical Modeling for Task Recognition and Action Segmentation in Weakly-Labeled Instructional Videos View paper
- [47] Joint Visual-Temporal Embedding for Unsupervised Learning of Actions in Untrimmed Sequences View paper
- [48] Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning View paper
- [49] Procedure learning via regularized gromov-wasserstein optimal transport View paper
- [50] THESAURUS: Contrastive Graph Clustering by Swapping Fused Gromov-Wasserstein Couplings View paper
- [51] Weakly-Supervised Temporal Action Alignment Driven by Unbalanced Spectral Fused Gromov-Wasserstein Distance View paper
- [52] A Fused Gromov-Wasserstein Approach to Subgraph Contrastive Learning View paper
- [53] A Fused Gromov-Wasserstein Framework for Unsupervised Knowledge Graph Entity Alignment View paper
- [54] Temporally Consistent Unbalanced Optimal Transport for Unsupervised Action Segmentation View paper
- [55] Timestamp-supervised wearable-based activity segmentation and recognition with contrastive learning and order-preserving optimal transport View paper
- [56] Multi-granularity correspondence learning from long-term noisy videos View paper
- [57] Optimal transport guided contrastive video summarization View paper