CycDA: Unsupervised Cycle Domain Adaptation to Learn from Image to Video

Wei Lin[1,4] Anna Kukleva[2] Kunyang Sun[1,5] Horst Possegger[1] Hilde Kuehne[3] Horst Bischof[1]

[1]Graz University of Technology [2]Max-Planck-Institute for Informatics [3]Goethe University Frankfurt

[4]Christian Doppler Laboratory for Semantic 3D Computer Vision [5]Southeast University

ECCV TEL AVIV 2022

Motivation

Image-to-video adaptation

- Web images are a labeling-free data source for action recognition
- Webly-labeled images to unlabeled videos

But ...

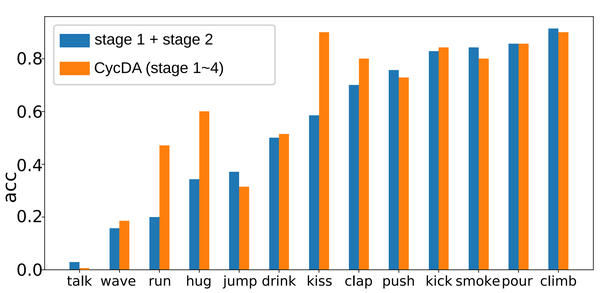
- **Spatial domain shift** between web images and video frames
- **Modality gap** between images and videos

Overview

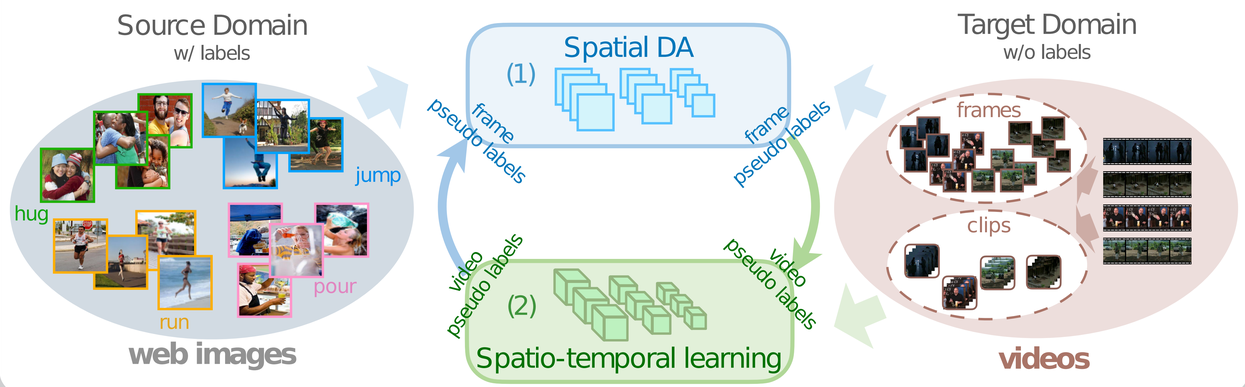
CycDA

- solves image-to-video adaptation by decoupling a) **domain alignment** b) **spatio-temporal learning**
- is a cyclic alternation of 4 stages between spatial and spatio-temporal learning
- has extensive evaluations with SOTA results on **image-to-video adaptation** and **mixed-source (image+video) to video adaptation**

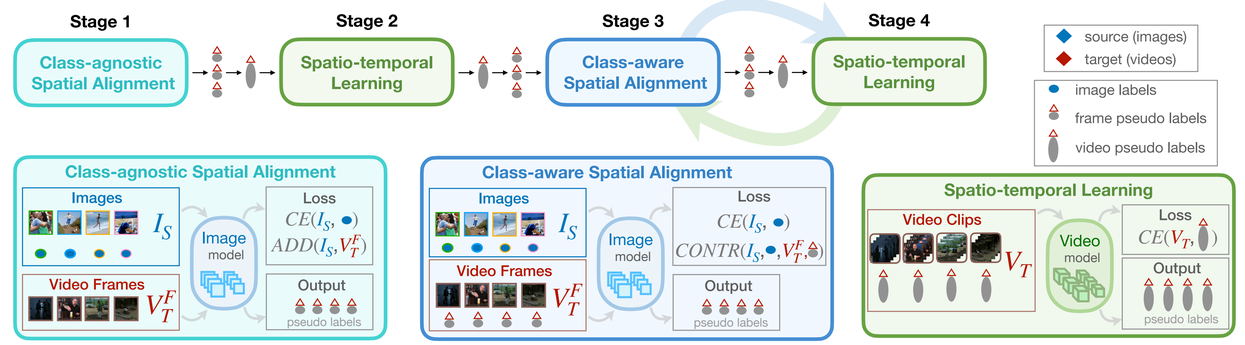
Category-level Pseudo Labels



CycDA



Stages



Stage 1: Class-agnostic domain alignment

- Task: domain alignment without category information. The classes of source and target data could be misaligned.
- Model: image model ϕ^I
- Objective: supervised cross entropy loss and adversarial domain discrimination loss

$\min_{\theta^I_E,\theta^I_C} \mathcal{L}_{CE}(I_S) +$
$\beta \max_{\theta^I_E} \min_{\theta^I_D} \mathcal{L}_{ADD}(I_S, V^F_T)$

Stage 2 & Stage 4: Spatio-temporal learning

- Task: spatio-temporal learning with pseudo labeld target data
- Pseudo labeling: use image model ϕ^I to predict pseudo labels on target frames, and temporally aggregate frame-level labels into video-level labels
- Model: video model ϕ^V
- Objective: supervised cross entropy loss

$\min_{\theta^V_E,\theta^V_C} \hat{\mathcal{L}}_{CE}(\hat{V}_T)$

Stage 3: Class-aware domain alignment

- Task: Domain alignment on the category level
- Pseudo labeling: use the video model ϕ^V to generate video-level labels, and disseminate into frame-level labels
- Model: image model ϕ^I
- Objective: cross-domain contrastive learning

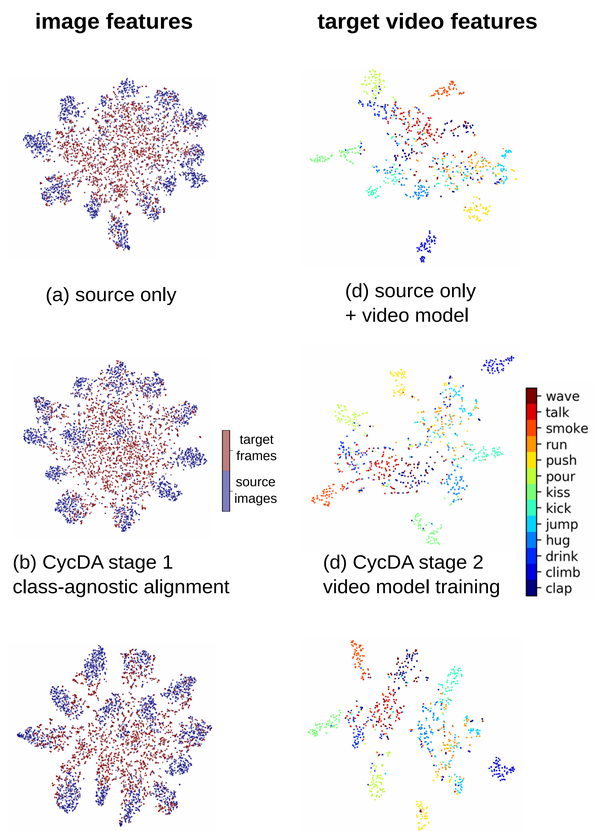
$\mathcal{L}_{CONTR}(I_S, V^F_T) =$
$-\sum_{z^F_j \in Z^F_T} \log \frac{h(z^F_j, z^I_{j+})}{h(z^F_j, z^I_{j+}) + h(z^F_j, z^I_{j-})}$

Pseudo labeled target sample $z^F_j \in Z^F_T$
Positive sample from source $z^I_{j+} \in I_S$
Negative sample from source $z^I_{j-} \in I_S$

Cycling of the Stages

Stage 3 and stage 4 can be performed iteratively.

Feature Distribution



Nearest Neighbor Search