Unsupervised Video Domain Adaptation: A Disentanglement Perspective

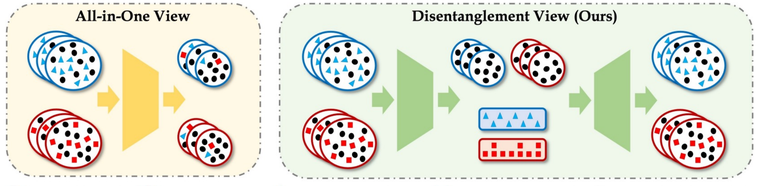
Pengfei Wei, Lingdong Kong,
Xinghua Qu, Yi Ren, Zhiqiang Xu,
Jing Jiang, Xiang Yin

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation & Contribution

TL;DR

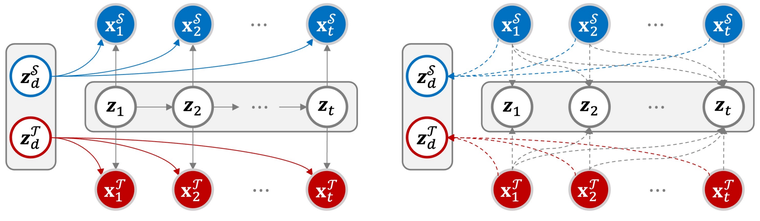
- We introduce the **Transfer Sequential VAE (TranSVAE)** framework to tackle unsupervised video domain adaptation tasks from a **disentanglement view**.
- Our key idea is to handle the **spatial** and **temporal** domain divergence separately through disentanglement; we pursue an explicit decoupling of the domain-specific info from other info via **generative modeling**.



- Our approach achieves new **state-of-the-art** performance on the domain adaptation benchmarks of UCF-HMDB, Jester, Epic-Kitchens, and Sprites.

Domain Generation

- We consider the generation process of **cross-domain videos** from two sets of latent factors: one set consists of a sequence of random variables, which are **dynamic** and inclined to encode the **semantic information**; another set is **static** and introduces some domain-related **spatial information**.

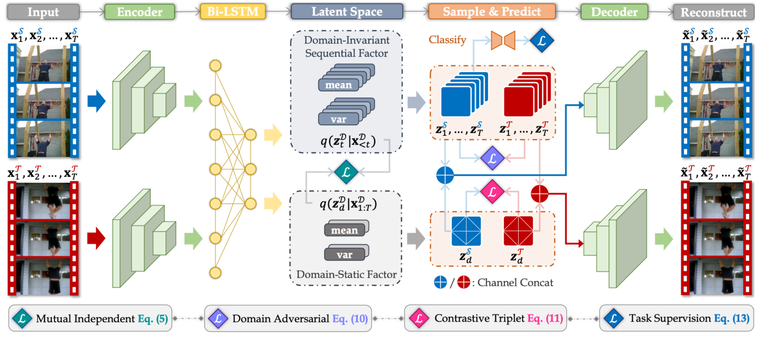


- The **graphical illustrations** of the proposed **generative** (left) and **inference** (right) models for video domain disentanglement are shown above.
- The blue and red **nodes** denote the **observed** source domain and target domain **videos** $\mathbf{x}^S$ and $\mathbf{x}^T$, respectively, over t timestamps.
- The **static latent variables** $\mathbf{z}_d^S$ and $\mathbf{z}_d^T$ follow a **joint distribution** and combining either of them with **dynamic latent variables** $\mathbf{z}_t$ constructs one video data of a domain, with which we developed **TranSVAE**.

Methodology

The TranSVAE Framework

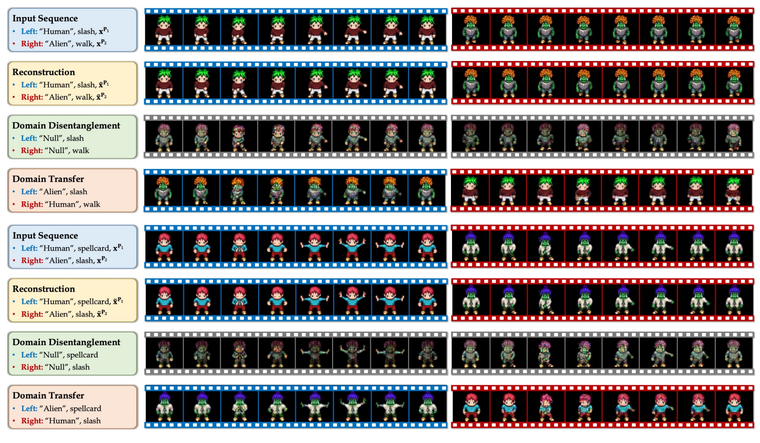
- The input videos are fed into an **encoder** to extract the visual features, followed by an **LSTM** to explore the temporal information.
- Two groups of **mean** and **variance** networks are then applied to model the **posterior** of the **latent factors**, i.e., $q(\mathbf{z}_t^{\mathcal{D}}|\mathbf{x}_{\leq t}^{\mathcal{D}})$ and $q(\mathbf{z}_d^{\mathcal{D}}|\mathbf{x}_{\leq 1:T}^{\mathcal{D}})$.



- The **new representations** $\mathbf{z}_1^{\mathcal{D}}$, $\mathbf{z}_2^{\mathcal{D}}$, ..., $\mathbf{z}_T^{\mathcal{D}}$ and $\mathbf{z}_d^{\mathcal{D}}$ are sampled, and then concatenated and passed to a **decoder** for **reconstruction**. Four constraints are then proposed to regulate the latent factors for adaptation purposes.

Domain Disentanglement

- We show domain **disentanglement** and **transfer** examples using Sprites as follows.



Experiments & Analyses

Comparative Study

- We conduct extensive experiments on popular unsupervised video domain adaptation benchmarks: **UCF-HMDB**, **Jester**, **Epic-Kitchens**, and **Sprites**.

Task	$\mathcal{S}_{\text{only}}$	MM-SADA	STCDA	CMCD	A3R	CleanAdapt	CycDA	MixDANN	CIA	**TranSVAE**
$\mathbf{U} \rightarrow \mathbf{H}$	86.1	84.2	83.1	84.7	/	**89.8**	88.1	82.2	**88.3**	87.8 (+1.7)
$\mathbf{H} \rightarrow \mathbf{U}$	92.5	91.1	92.1	92.8	/	**99.2**	98.0	92.8	94.1	99.0 (+6.5)
Average ↑	89.3	87.7	87.6	88.8	/	**94.5**	93.1	87.5	91.2	93.4 (+4.1)
$\mathbf{D}_1 \rightarrow \mathbf{D}_2$	43.2	49.5	52.0	50.3	**53.2**	52.7	/	56.0	52.5	50.5 (+7.3)
$\mathbf{D}_1 \rightarrow \mathbf{D}_3$	42.5	44.1	45.5	46.3	**52.1**	47.0	/	47.3	47.8	50.3 (+7.8)
$\mathbf{D}_2 \rightarrow \mathbf{D}_1$	43.0	48.2	49.0	49.5	**51.9**	46.2	/	50.3	49.8	50.3 (+7.3)
$\mathbf{D}_2 \rightarrow \mathbf{D}_3$	48.0	52.7	52.5	52.0	55.5	52.7	/	52.4	53.2	**58.6** (+10.6)
$\mathbf{D}_3 \rightarrow \mathbf{D}_1$	43.0	50.9	**52.6**	48.7	51.5	47.8	/	51.0	52.2	48.0 (+5.0)
$\mathbf{D}_3 \rightarrow \mathbf{D}_2$	55.5	56.1	55.6	56.3	**63.2**	54.4	/	54.7	57.6	58.0 (+2.5)
Average ↑	45.9	50.3	51.2	51.0	**54.1**	50.3	/	52.0	52.2	52.6 (+6.7)

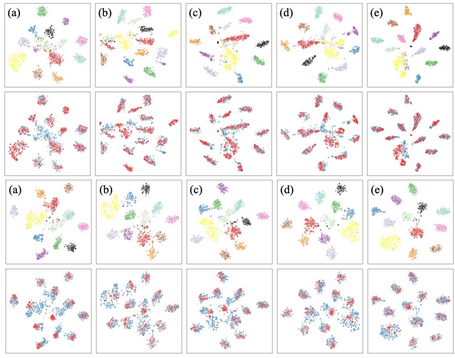
- Across all tasks, **TranSVAE** consistently outperforms **all** previous methods using **single** modality inputs; our approach even achieves better average results than seven out of eight **multi-modal** approaches (RGB + flow).

Ablation Study

- The **loss separation study** on different tasks demonstrates the effectiveness for each of the four constraints designed in the **TranSVAE** framework.

$\mathcal{L}_{\text{svae}}$	$\mathcal{L}_{\text{cls}}$	$\mathcal{L}_{\text{adv}}$	$\mathcal{L}_{\text{mi}}$	$\mathcal{L}_{\text{ctc}}$	PL	$\mathbf{U} \rightarrow \mathbf{H}$	$\mathbf{H} \rightarrow \mathbf{U}$	$\mathbf{J}_S \rightarrow \mathbf{J}_T$	$\mathbf{D}_1 \rightarrow \mathbf{D}_2$	$\mathbf{D}_1 \rightarrow \mathbf{D}_3$	$\mathbf{D}_2 \rightarrow \mathbf{D}_1$	$\mathbf{D}_2 \rightarrow \mathbf{D}_3$	$\mathbf{D}_3 \rightarrow \mathbf{D}_1$	$\mathbf{D}_3 \rightarrow \mathbf{D}_2$	**Avg.**
✓		✓	✓	✓	✓	18.61	26.62	22.92	34.00	30.29	33.79	30.49	28.51	34.27	28.83
✓	✓		✓	✓	✓	83.06	93.52	48.07	40.93	43.33	43.91	51.13	41.84	52.67	55.38
✓	✓	✓		✓	✓	85.83	93.52	65.12	46.67	48.56	49.43	55.34	45.52	54.53	60.60
✓	✓	✓	✓		✓	83.89	95.80	64.89	48.53	48.25	48.96	54.21	45.52	55.73	60.64
✓	✓	✓	✓	✓		87.22	94.40	64.64	49.87	48.25	49.66	56.47	47.59	55.07	61.46
✓	✓	✓	✓	✓	✓	**87.78**	**98.95**	**66.10**	**50.53**	**50.31**	**50.34**	**58.62**	**48.04**	**58.00**	**63.19**

- The **loss integration study** and t-SNE plots verify that all constraints serve to achieve good **disentanglement effect** with domain divergence minimization.



#	$\mathcal{L}_{\text{svae}}$	$\mathcal{L}_{\text{cls}}$	$\mathcal{L}_{\text{adv}}$	$\mathcal{L}_{\text{mi}}$	$\mathcal{L}_{\text{ctc}}$	PL	Accuracy (%)
$\mathcal{S}_{\text{only}}$							80.27
(a)	✓	✓					82.50 (+2.23)
(b)	✓	✓	✓				84.44 (+4.17)
(c)	✓	✓	✓	✓			85.56 (+5.29)
(d)	✓	✓	✓	✓	✓		87.22 (+6.95)
(e)	✓	✓	✓	✓	✓	✓	87.78 (+7.61)
$\mathcal{T}_{\text{sup}}$							95.00

#	$\mathcal{L}_{\text{svae}}$	$\mathcal{L}_{\text{cls}}$	$\mathcal{L}_{\text{adv}}$	$\mathcal{L}_{\text{mi}}$	$\mathcal{L}_{\text{ctc}}$	PL	Accuracy (%)
$\mathcal{S}_{\text{only}}$							88.79
(a)	✓	✓					92.29 (+3.50)
(b)	✓	✓	✓				93.52 (+4.73)
(c)	✓	✓	✓	✓			93.70 (+4.91)
(d)	✓	✓	✓	✓	✓		94.40 (+5.61)
(e)	✓	✓	✓	✓	✓	✓	98.95 (+10.16)
$\mathcal{T}_{\text{sup}}$							96.85

ByteDance

NUS National University of Singapore

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

UTS UNIVERSITY OF TECHNOLOGY, SYDNEY

Paper

Code