Alignment-Uniformity aware Representation Learning for Zero-shot Video Classification

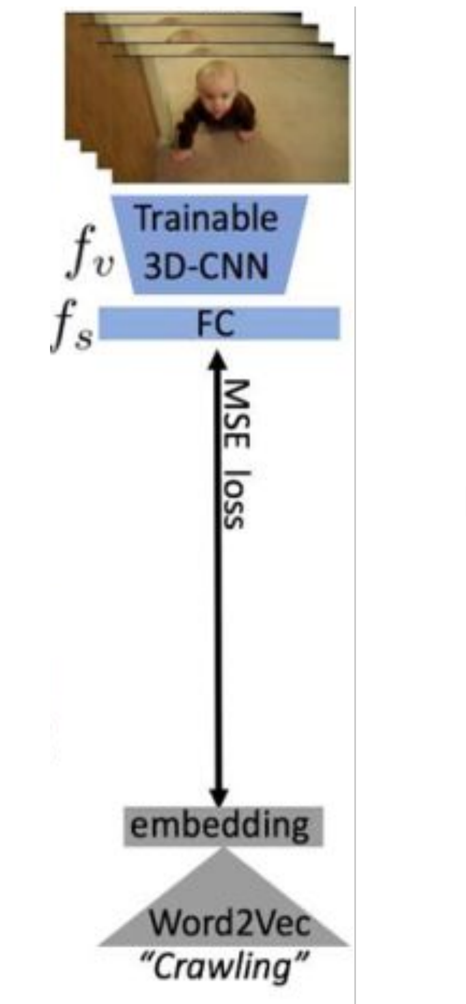
[1]Tencent TEG Machine Learning Platform Department [2]Beijing University of Posts and Telecommunications

CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

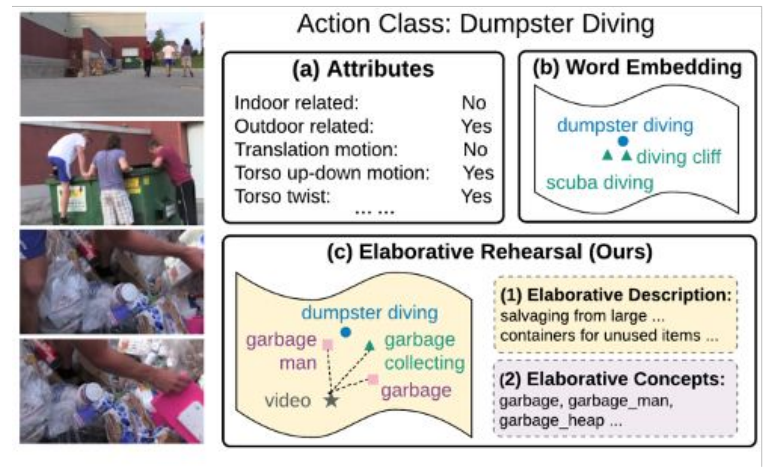
Shi Pu[1*] Kaili Zhao[2*] Mao Zheng[1]

Introduction

➢ SoTA zero-shot video classification (ZSVC)



E2E: Align visual and semantics features [r1].

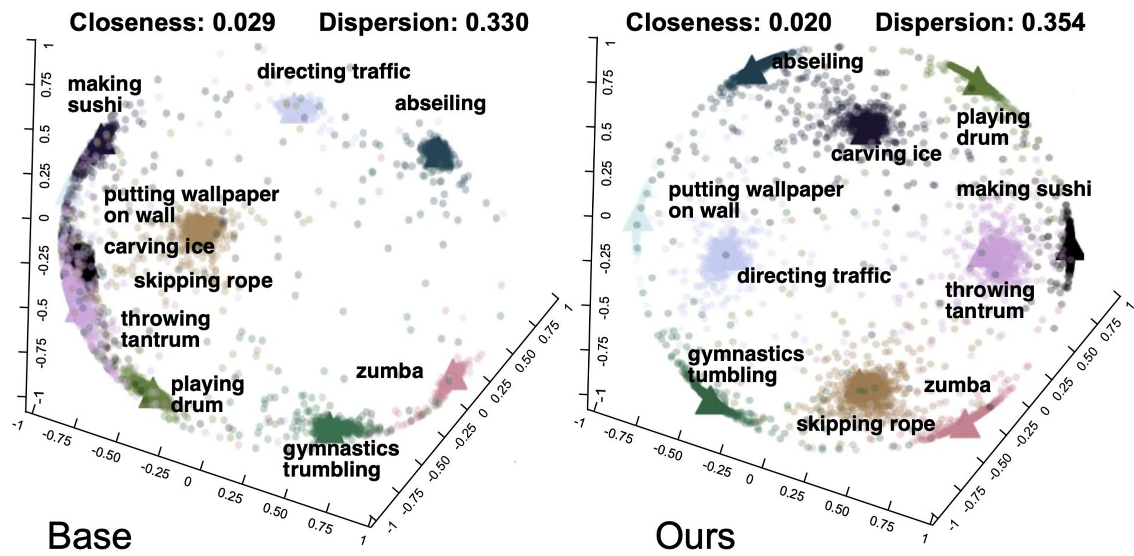


ER: Expand class names by hand-crafted annotations [r2]

✅ Both **E2E** and **ER** learn a unified visual and semantic representation.
❎ But, only alignment property of the representation is considered while uniformity that contributes to generalization is neglected.
✅ **ER** tries to learn unseen classes by expanding existing class names.
❎ But, amount of extra annotations are required.

[r1] Rethinking zero-shot video classification: End-to-end training for realistic applications, CVPR 2020.
[r2] Elaborative rehearsal for zero-shot action recognition, ICCV 2021.

➢ Contributions



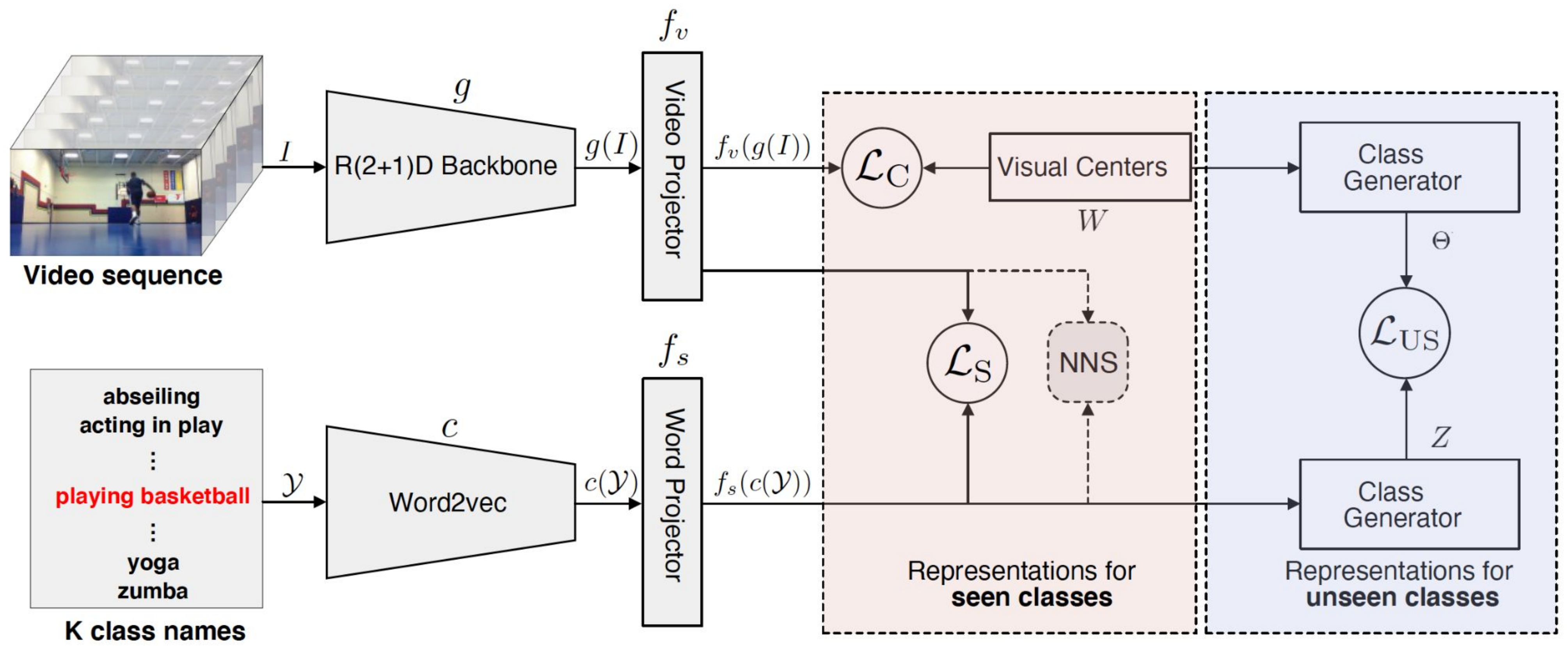
Base Ours

The unified **representations** of visual and semantics.

(1) Propose a unified visual and semantic representation awareness of alignment and uniformity jointly.
(2) Explicitly generate synthetic "unseen" classes via our class generator.
(3) Present closeness and dispersion to quantify alignment and uniformity, serving as new measurements of model generalizability.
(4) Achieve state-of-the-art performance in an end-to-end manner.

Alignment-Uniformity aware Representation Learning (AURL)

➢ AURL architecture



➢ Alignment-uniformity aware loss to regularize the representations

- **Supervised contrastive loss enables alignment and uniformity properties.**

$$\mathcal{L}^{sup} = -\log[\frac{\exp[\lambda \mathrm{sim}(v_{y_i}, s_{y_i})]}{\sum_{y_j \in \mathcal{Y}} \exp[\lambda \mathrm{sim}(v_{y_i}, s_{y_j})]}] = \lambda \mathrm{SP}_\lambda(\underbrace{-\mathrm{sim}(v_{y_i}, s_{y_i})}_{\text{alignment}} + \frac{1}{\lambda}\underbrace{\mathrm{LSE}(\lambda \mathrm{sim}(v_{y_i}, s_{y_j})_{y_j \in \mathcal{Y}\setminus y_i})}_{\text{uniformity}}),$$

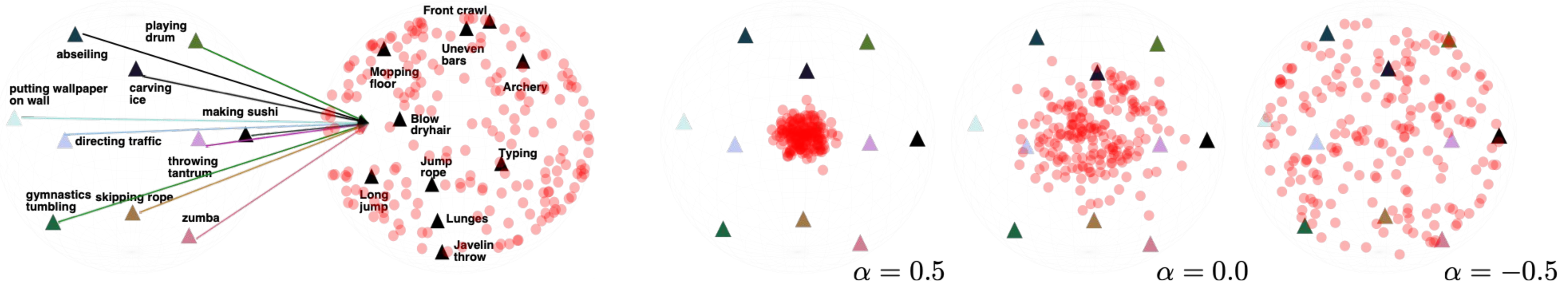
$$\text{where,} \quad \mathrm{SP}_\lambda(x) = \frac{1}{\lambda}\log(1+\exp(\lambda x)), \quad \mathrm{LSE}(x) = \log(\sum_{x \in \mathcal{X}} \exp(x)).$$

- **Apply the supervised contrastive loss for seen and synthetic unseen classes**

$$\mathcal{L}_{\text{contrast}} = \mathcal{L}_{\text{S}} + \mathcal{L}_{\text{US}}$$

$$= -\log[\frac{\exp[\lambda \cos(v_{y_i}, s_{y_i})]}{\sum_{y_j \in \mathcal{Y}} \exp[\lambda \cos(v_{y_i}, s_{y_j})]}] + \frac{1}{K_u}\sum_{u_i \in \mathcal{U}} -\log[\frac{\exp[\lambda \cos(\Theta_{u_i}, Z_{u_i})]}{\sum_{u_j \in \mathcal{U}} \exp[\lambda \cos(\Theta_{u_i}, Z_{u_j})]}].$$

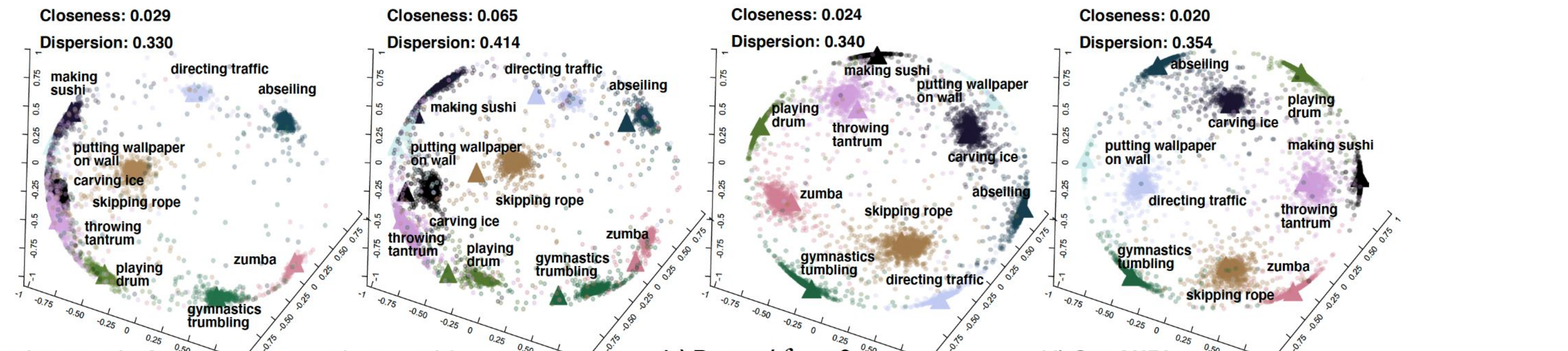
- **Class generator to synthesize unseen classes.**



(a) Synthetic classes can cover test classes. (b) Generate unseen classes w/ different parameters.

Experiments

➢ Ablations w/ crucial modules



(a) Base w/ MSE (b) Base w/ $\mathcal{L}_S$ (c) Base w/ $\mathcal{L}_S$ + e2e (d) Our AURL

Method	$\mathcal{L}_S$	e2e	$\mathcal{L}_{US}$ +CG	Close.	Disper.	UCF top-1	HMDB top-1
Base w/ MSE				0.30	0.09	35.1	21.3
Base w/ $\mathcal{L}_S$	✓			0.45	0.29	40.3	24.2
Base w/ $\mathcal{L}_S$ + e2e	✓	✓		0.30	0.29	43.2	26.2
AURL (ours)	✓	✓	✓	0.29	0.32	44.4	27.4
AURL w/o CG.	✓	✓	✓	0.33	0.32	43.7	25.8

➢ Takeaways

1. Uniformity expands borders of representations as much as possible, and has been validated its effectiveness in model generalizability of ZSVC.
2. Alignment and uniformity jointly benefit ZSVC.
3. Our code is available at https://github.com/ShipuLoveMili/CVPR2022-AURL

➢ Comparisons with SoTA alternatives

Method	Test splits	Train dataset	UCF top-1	Train dataset	HMDB top-1
SoTA* [3]	1	Kinetics	37.6	Kinetics	26.9
AURL*	1	Kinetics	46.8	Kinetics	31.7
Obj2act [16]	3	-	30.3	-	15.6
SAOE [27]	3	-	32.8	-	-
OPCL [13]	3	-	36.3	-	-
MUFI [35]	3	Kinetics+	56.3	Kinetics+	31.0
AURL	3	Kinetics	60.9	Kinetics	40.4
TARN [2]	30	UCF	23.2	HMDB	19.5
Act2Vec [14]	-	UCF	22.1	HMDB	23.5
SAOE [27]	10	-	40.4	-	-
PSGNN [13]	50	UCF	43.0	HMDB	32.6
OPCL [13]	10	-	47.3	-	-
SoTA* [3]	10	Kinetics	48.0	Kinetics	32.7
DASZL [19]	10	-	48.9	-	-
ER [6]	50	UCF	51.8	HMDB	35.3
AURL*	10	Kinetics	58.0	Kinetics	39.0

- **Closeness and dispersion score to quantify the alignment and uniformity.**

$$\text{Closeness} = \frac{1}{K}\sum_{y_i \in \mathcal{Y}}[\frac{1}{N_{y_i}}\sum_{n=1}^{N_{y_i}}(1-\cos(v_{y_i}^n, s_{y_i}^n))] \qquad \text{Dispersion} = \frac{1}{K}\sum_{y_i \in \mathcal{Y}} \min_{y_k \in \mathcal{Y}\setminus y_i}(1-\cos(\bar{v}_{y_i}, \bar{v}_{y_k}))$$

- **Inference with a strict label requirement that excludes highly overlapped classes.**

$$\underset{y^t \in \mathcal{Y}^t}{\mathrm{argmax}} \cos(f_v(g(I^t)), s_{y^t}) \qquad \forall y \in \mathcal{Y}, \ \min_{y^t \in \mathcal{Y}^t}(1-\cos(c_y, c_{y^t})) > \tau$$