PromptonomyViT: Multi-Task Prompt Learning Improves Video Transformers Using Synthetic Scene Data

Roei Herzig*[1,4], Ofir Abramovich*[2], Elad Ben-Avraham[1], Assaf Arbelle[4], Leonid Karlinsky[5], Ariel Shamir[2], Trevor Darrell[3], Amir Globerson[1]

[1]Tel-Aviv University, [2]Reichman University, [3]UC Berkeley, [4]IBM Research, [5]MIT-IBM Watson AI Lab

Motivation

- Instance level distinction and 3D perception are crucial for understanding complex scenes
- Collecting and annotating real large-scale video datasets require an extensive amount of effort and a high budget
- We exploit the availability of synthetic data and transformer-based video architectures to fill these gaps

Contributions

- We present a method for exploiting synthetically generated labels for several tasks to improve video understanding models
- We propose the concept of special "multi-task prompts" to capture task-related information through multi-task supervision
- We demonstrate improved performance on five video understanding benchmarks

Real Video Dataset

- Expensive to produce/collect/annotate

Synthetic Image/Video Datasets

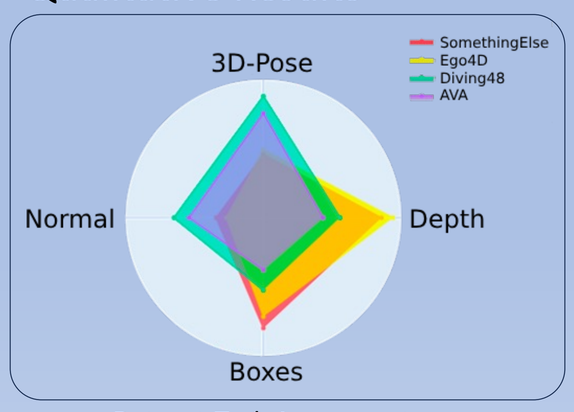
- Largely available
- Easy to scale
- High Quality Annotated

Main Dataset (for Action Recognition/Classification etc.)

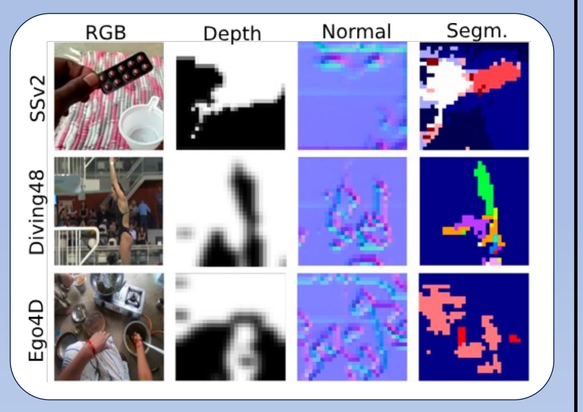
Auxiliary data with scene-level annotations

PViT Model

Qualitative Results

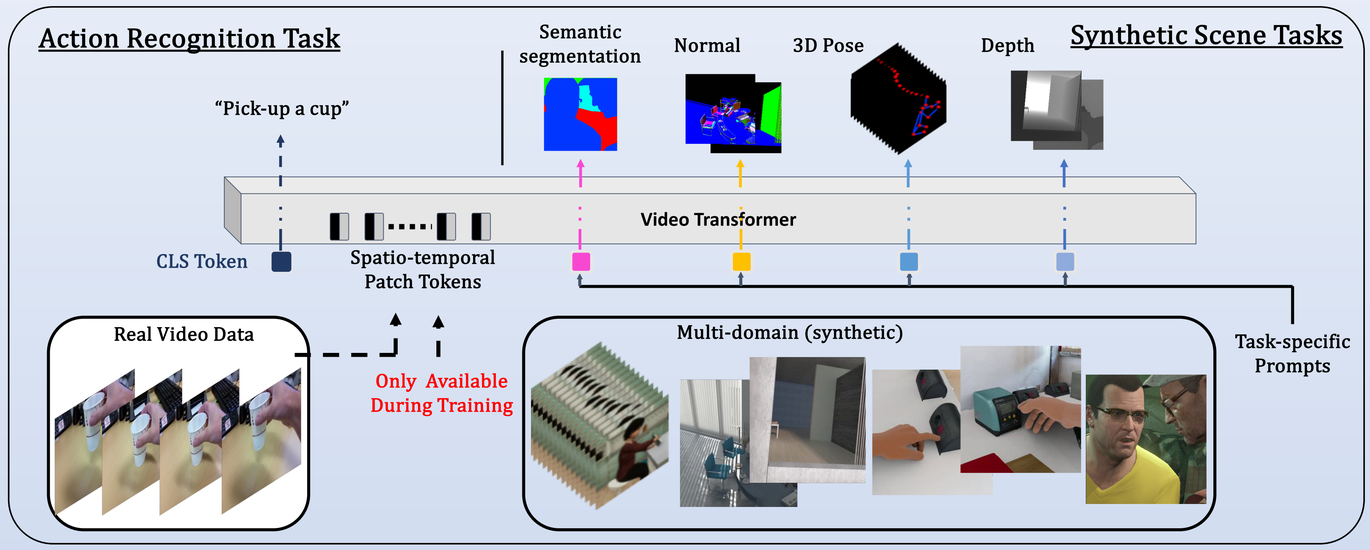


Dataset-Task Agreement



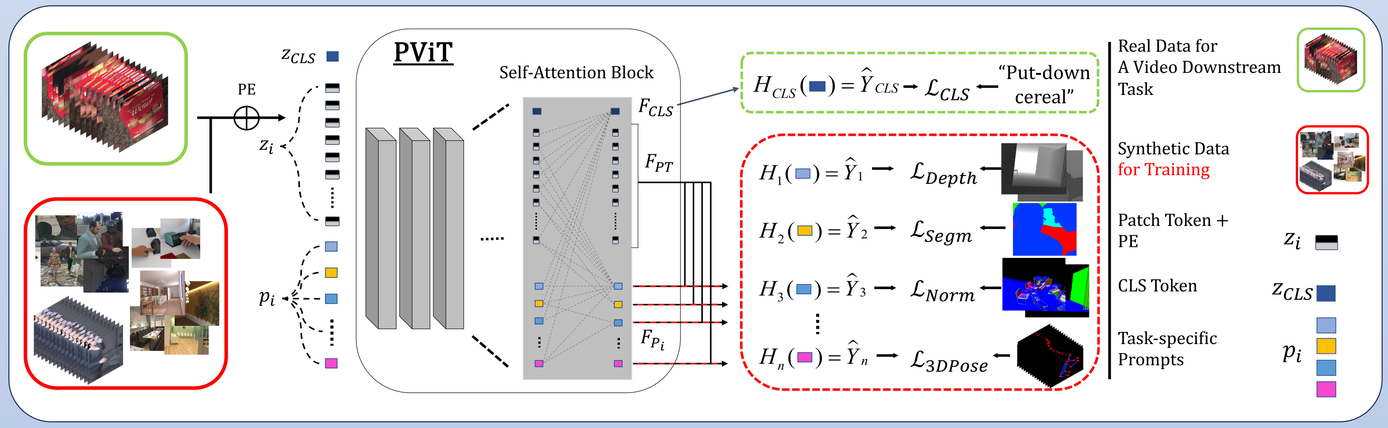
Visualization of the output of the "task prompts" prediction heads on real data

PViT Approach



PViT adds a set of multiple prompts to a video transformer to capture inter-task structure and solve a downstream task. It utilizes a multi-task prompt learning approach for video transformers, where a shared transformer backbone is enhanced with task-specific prompts (colored squares).

Method



Model Architecture. We extend a transformer with a set of "task prompts", pi, that are designed to capture information regarding each task, as well as capture the inter-task structure. The prompts are supervised by synthetic scene auxiliary tasks (depth, segmentation, normal, and 3D pose) available only during training.

Quantitative Results

(a) **Something–Something V2**

Model	Pretrain	Top-1	Top-5
SlowFast [28], R101	K400	63.1	87.6
MViTv1 [27]	K400	64.7	89.2
ViViT-L [2]	IN+K400	65.4	89.8
UniFormer-S [62]	IN+K600	67.9	92.1
ORViT Mformer [40]	K400	67.9	90.5
VideoMAE (ViT-S)	K400	66.8	90.3
MViTv2 [66]	K400	68.2	91.4
MViTv2 MT	K400	68.4	91.3
MViTv2 VPT	K400	61.5	87.5
PViT (Ours)	K400	**69.6** (+1.2)	**91.6** (+0.2)

(b) **Diving48**

Model	Pretrain	Frames	Top-1
SlowFast [28], R101	K400	16	77.6
TimeSformer [10]	IN	16	74.9
TimeSformer-L [10]	IN	96	81.0
SViT [5]	K400	16	79.8
MViTv2 [66]	K400	16	73.1
MViTv2 MT	K400	16	75.6
MViTv2 VPT	K400	16	69.8
PViT (Ours)	K400	16	**85.8** (+6.0)

(c) **AVA-V2.2**

Model	Pretrain	mAP
SlowFast [28], R50	K400	22.7
SlowFast [28], R101	K400	23.8
ORViT MViT-B [40]	K400	26.6
VideoMAE (ViT-S) [99]	K400	22.5
VideoMAE (ViT-B) [99]	K400	26.7
MViTv1 [27]	K400	25.5
MViTv2 [66]	K400	26.8
MViTv2 MT	K400	27.2
MViTv2 VPT	K400	19.0
PViT (Ours)	K400	**28.4** (+1.6)

(d) **SomethingElse**

Model	Compositional		Base		Few-Shot	
	Top-1	Top-5	Top-1	Top-5	5-Shot	10-Shot
I3D [14]	42.8	71.3	73.6	92.2	21.8	26.7
SlowFast [28]	45.2	73.4	76.1	93.4	22.4	29.2
TimeSformer [10]	44.2	76.8	79.5	95.6	24.6	33.8
STIN [77]	48.2	72.6	-	-	-	-
TSM [68]	52.3	78.0	-	-	-	-
Mformer [81]	60.2	85.8	82.8	96.2	28.9	33.8
SAFCAR [57]	60.7	84.2	-	-	-	-
MViTv2 [66]	63.3	87.5	83.7	96.8	32.7	40.2
MViTv2 MT	62.7	87.6	81.4	96.2	34.0	40.9
MViTv2 VPT	53.0	81.8	76.8	94.8	31.8	39.0
PViT (Ours)	**65.5** (+2.2)	**89.0** (+2.5)	**85.0** (+1.3)	**97.4** (+0.6)	**34.3** (+1.6)	**41.3** (+1.1)

(e) **Ego4D**

Model	Temporal Localization Error	PNR Classification Top-1
Bi-LSTM	0.790	65.3
BMN [70]	0.780	-
I3D ResNet-50 [14]	0.739	68.7
EgoVLP (TimeSformer) [69]	0.666	73.9
Video Swin Transformer [75]	0.660	69.5
MViTv2 [66]	0.702	71.6
MViTv2 MT	0.640	73.6
MViTv2 OP	0.652	73.7
PViT (Ours)	0.637 (-0.065)	**74.8** (+3.2)

Results on SSv2, Diving48, AVA-V2.2, SomethingElse, and Ego4D datasets. We report top-1 and top-5 accuracy on SSv2 and SomethingElse. On AVA, we report the mAP metric. On Diving48, we report top-1. On Ego4D we report classification error. IN refers to ImageNet-21K.

JAN 4-8 WACV 2024 WAIKOLOA HAWAII