Partial Video Domain Adaptation with Partial Adversarial Temporal Attentive Network Supplementary Material

Yuecong Xu[1,2,*] xuyu0014@e.ntu.edu.sg

Jianfei Yang[1,*] yang0478@ntu.edu.sg

Haozhi Cao[1] haozhi001@e.ntu.edu.sg

Zhenghua Chen[2,†] chen0832@e.ntu.edu.sg

Qi Li[1] liqi0024@e.ntu.edu.sg

Kezhi Mao[1] ekzmao@ntu.edu.sg

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[2]Institute for Infocomm Research, A*STAR, Singapore

1. PVDA Benchmarks

In this work, we propose three sets of benchmarks, UCF-HMDB$_{partial}$, MiniKinetics-UCF, and HMDB-ARID$_{partial}$, which cover a wide range of *Partial Video Domain Adaptation* (PVDA) scenarios and provide adequate baseline environment with distinct domain shift to facilitate PVDA research. Here we provide more detail on each benchmark. All benchmarks could be downloaded at https://xuyu0010.github.io/pvda.html

UCF-HMDB$_{partial}$. UCF-HMDB$_{partial}$ is built from two widely used video datasets: UCF101 (**U**) [7] and HMDB51 (**H**) [5]. The overlapping classes between the two datasets are collected, resulting in 14 classes with 2,780 videos. Among which are 980 training videos and 210 testing videos from HMDB51, 1,324 training videos and 266 testing videos from UCF101. The list of the 14 overlapping classes are listed in Table 1. The first 7 categories in alphabetic order of the target domain are chosen as target categories, and we construct two PVDA tasks: **U-14→H-7** and **H-14→U-7**. We follow the official split for the training and validation sets. Figure 1 shows the comparison of sampled frames from UCF-HMDB$_{partial}$.

MiniKinetics-UCF. MiniKinetics-UCF is built from two large-scale video datasets: MiniKinetics-200 (**M**) [8] and UCF101 (**U**) [7]. MiniKinetics-200 is a subset of the Kinetics [4] dataset, with 200 of its categories. There are 45 overlapping classes between MiniKinetics-200 and UCF101, as shown in Table 2. Similar to the construction of UCF-HMDB$_{partial}$, the first 18 categories in alphabetic order of the target domain are chosen as target categories, resulting in two PVDA tasks: **M-45→U-18** and **U-45→M-18**. In this dataset, there are a total of 22,102 videos, with 4,253

UCF101 Class	HMDB51 Class
RockClimbingIndoor	climb
Diving	dive
Fencing	fencing
GolfSwing	golf
HandstandWalking	handstand
SoccerPenalty	kick_ball
PullUps	pullup
Punch	punch
PushUps	pushup
Biking	ride_bike
HorseRiding	ride_horse
Basketball	shoot_ball
Archery	shoot_bow
WalkingWithDog	walk

Table 1. List of overlapping classes between UCF101 and HMDB51.

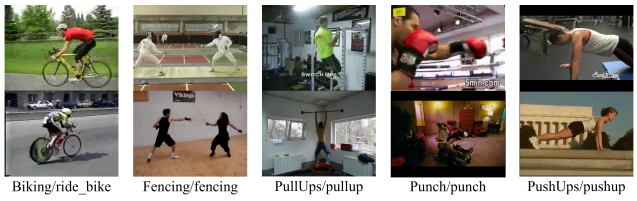

Figure 1. Sampled frames of videos from classes in UCF-HMDB$_{partial}$. Sampled frames from UCF101 are shown in the upper row, and those from HMDB51 are shown in the lower row.

[1]Equal Contribution.
[2]Corresponding Author.

[3]This research is supported by A*STAR under its AME Programmatic Funds (Grant No. A20H6b0151).

MiniKinetics-200 Class	UCF101 Class	MiniKinetics-200 Class	UCF101 Class	MiniKinetics-200 Class	UCF101 Class
archery	Archery	high_jump	HighJump	pole_vault	PoleVault
bench_pressing	BenchPress	hula_hooping	HulaHoop	pull_ups	PullUps
biking_through_snow	Biking	javelin_throw	JavelinThrow	riding_or_walking_with_horse	HorseRiding
blowing_out_candles	BlowingCandles	jetskiing	Skijet	rock_climbing	RockClimbingIndoor
bowling	Bowling	juggling_balls	JugglingBalls	salsa_dancing	SalsaSpin
brushing_teeth	BrushingTeeth	long_jump	LongJump	shaving_head	ShavingBeard
canoeing_or_kayaking	Kayaking	lunge	Lunges	shot_put	Shotput
catching_or_throwing_baseball	BaseballPitch	making_pizza	PizzaTossing	skateboarding	SkateBoarding
catching_or_throwing_frisbee	FrisbeeCatch	marching	BandMarching	skiing	Skiing
clean_and_jerk	CleanAndJerk	playing_basketball	Basketball	squat	BodyWeightSquats
crawling_baby	BabyCrawling	playing_cello	PlayingCello	surfing_water	Surfing
diving_cliff	CliffDiving	playing_guitar	PlayingGuitar	swimming_breast_stroke	BreastStroke
dunking_basketball	BasketballDunk	playing_tennis	TennisSwing	tai_chi	Taichi
golf_driving	GolfSwing	playing_violin	PlayingViolin	throwing_discus	ThrowDiscus
hammer_throw	HammerThrow	playing_volleyball	VolleyballSpiking	walking_the_dog	WalkingWithDog

Table 2. List of overlapping classes between MiniKinetics-200 and UCF101.

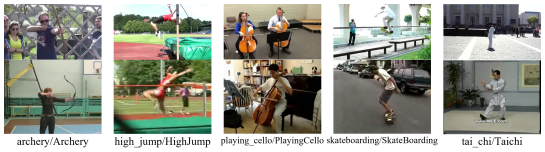

Figure 2. Sampled frames of videos from classes in MiniKinetics-UCF. Sampled frames from MiniKinetics-200 are shown in the upper row, while those from UCF101 are shown in the lower row.

training videos and 683 testing videos from UCF101, along with 16,743 training videos and 423 testing videos from MiniKinetics-200. The number of videos is nearly 8 times larger than that of UCF-HMDB$_{partial}$. Thus this dataset could validate the effectiveness of PVDA approaches on large-scale datasets. Figure 2 shows the comparison of sampled frames from MiniKinetics-UCF.

HMDB-ARID$_{partial}$. HMDB-ARID$_{partial}$ is built with the goal of leveraging current video datasets to boost performance on videos shot in adverse environments. It incorporates both HMDB51 (**H**) [5] and a more recent dark dataset, ARID (**A**) [9], with videos shot under adverse illumination conditions. Compared with current action recognition datasets (e.g. UCF101, HMDB51, MiniKinetics-200), videos in ARID are characterized by low brightness and low contrast. Statistically, videos in ARID possess much lower RGB mean value and standard deviation (std) as presented in Table 4. This leads to larger domain shift between ARID and HMDB51 compared to other cross-domain datasets. The overlapping classes between the two datasets are collected, resulting in 10 classes with 3,252 videos, which includes 2,012 training videos and 390 testing videos from ARID, and 700 training videos and 150 testing videos from HMDB51. The list of the 10 overlapping classes is listed in Table 3. Similar to the other two PVDA benchmarks, the first 5 categories in alphabetic order of the target domain are chosen as target categories, resulting in two PVDA tasks:

HMDB51 Class	ARID Class
RockClimbingIndoor	climb
Diving	dive
Fencing	fencing
GolfSwing	golf
HandstandWalking	handstand
SoccerPenalty	kick_ball
PullUps	pullup
Punch	punch
PushUps	pushup
Biking	ride_bike

Table 3. List of overlapping classes between HMDB51 and ARID.

Dataset	RGB Mean	RGB Std
HMDB51	[0.424,0.364,0.319]	[0.268,0.255,0.260]
UCF101	[0.409,0.397,0.358]	[0.266,0.265,0.270]
MiniKinetics-200	[0.435,0.394,0.381]	[0.225,0.225,0.214]
ARID	[0.079,0.074,0.073]	[0.101,0.098,0.090]

Table 4. Comparison of RGB mean and standard deviation (std) over common action recognition datasets and the ARID dataset.

Figure 3. Sampled frames of videos from classes in HMDB-ARID$_{partial}$. Sampled frames from HMDB51 are shown in the upper row, while those from ARID are shown in the lower row.

H-10→A-5 and **A-10→H-5**. For all the aforementioned benchmarks, the training and validation sets are separated following the official split methods. Figure 3 shows the comparison of sampled frames from HMDB-ARID$_{partial}$.

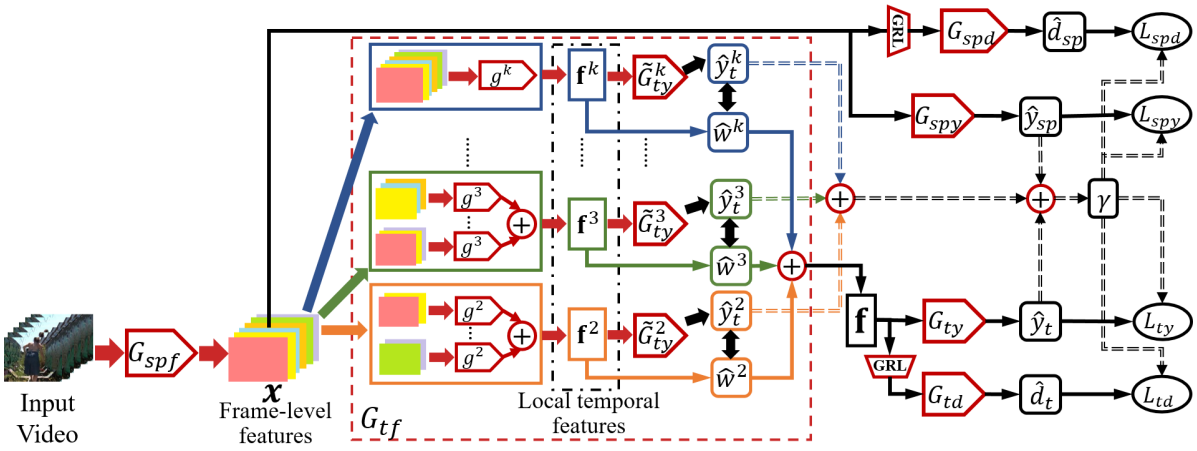

Figure 4. Architecture of the proposed PATAN. Dashed arrows indicate how γ is obtained and used in loss functions. *Best viewed in color and zoomed in.*

2. Detailed Implementation of the Proposed Network

As presented in Section 3, we propose PATAN to tackle the PVDA problem by constructing robust temporal features and utilizing both spatial and temporal features for accurate class filtration. The structure of our proposed PATAN is as shown in Figure 4. In this section, we further describe the implementation of PATAN in detail.

Our networks and experiments are implemented using the PyTorch [6] library. To obtain video features, we instantiate Temporal Relation Network [10] as the backbone for video feature extraction for both source domain videos and target domain videos, with the model pretrained on ImageNet [2]. The source and target feature extractors share parameters. New layers are trained from scratch, and their learning rates are set to be 10 times that of the pretrained-loaded layers.

The stochastic gradient descent algorithm [1] is used for optimization, with the weight decay set to 0.0001 and the momentum to 0.9. The batch size is set to 8 per GPU. Our initial learning rate is set to 0.005 and is divided by 10 for two times during the training process. We train our networks with a total of 50 epochs for UCF-HMDB$_{partial}$ and HMDB-ARID$_{partial}$, while for MiniKinetics-UCF we train for 30 epochs. The flip-coefficient of the Gradient Reverse Layer (GRL) is increased gradually from 0 to 1 as in DANN [3]. All experiments are conducted using two NVIDIA RTX 2080 GPUs.

References

[1] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. 3

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[3] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 3

[4] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[5] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011. 1, 2

[6] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019. 3

[7] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1

[8] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning for video understanding. *arXiv preprint arXiv:1712.04851*, 1(2):5, 2017. 1

[9] Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. Arid: A new dataset for recognizing action in the dark. *arXiv preprint arXiv:2006.03876*, 2020. 2

[10] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 3