Augmenting and Aligning Snippets for Few-Shot Video Domain Adaptation Supplementary Material

Yuecong Xu[1*] Jianfei Yang[2*] Yunjiao Zhou[2] Zhenghua Chen[1†] Min Wu[1] Xiaoli Li[1]
[1]Institute for Infocomm Research, A*STAR, Singapore [2]Nanyang Technological University
{xuyu0014, yang0478, yunjiao001, chen0832}e.ntu.edu.sg {wumin, xlli}@i2r.a-star.edu.sg

This appendix presents more details of the proposed Snippet-attentive Semantic-statistical Alignment with Stochastic Sampling Augmentation (SSA^{2}lign) and is organized as follows: first, we introduce the detailed implementation of SSA^{2}lign with specific hyperparameter settings, supported by additional results of hyperparameter sensitivity analysis to show the robustness of SSA^{2}lign. Subsequently, we present details of the cross-domain action recognition benchmarks for evaluating SSA^{2}lign, including Daily-DA and Sports-DA; lastly, we compare in detail our SSA^{2}lign with related but different FSDA and UDA/VUDA methods to highlight our novelty. Code will be provided at: https://github.com/xuyu0010/SSA2lign.

1. Implementation Details

Brief Review of SSA^{2}lign. In this work, we propose the **S**nippet-attentive **S**emantic-statistical **Align**ment with **S**tochastic **S**ampling **A**ugmentation (**SSA^{2}lign**) to address *Few-Shot Video Domain Adaptation* (FSVDA) by augmenting the source and target domains and performing domain alignment at the snippet level. SSA^{2}lign firstly augments the source and target domain data by a simple yet effective stochastic sampling process that makes full use of the abundance of snippet information and then performs semantic alignment from three perspectives: alignment based on semantic information within each snippet, cross-snippets of each video, and across snippet-level data distribution. To further improve the stability of snippet-level alignment, a statistical alignment strategy is additionally adopted, while snippet attention is proposed to weigh the impact of different target snippets on the domain alignment dynamically. In this section, we present the detailed implementation of SSA^{2}lign, whose pipeline is demonstrated in Fig. 1.

TimeSFormer as Feature Extractor. To obtain features from snippets during training and videos during testing, we instantiate the Transformer-based TimeSFormer [2] as the feature extractor thanks to its capability in obtaining features that include both spatial and temporal information. TimeSFormer extracts spatial and temporal features with separate space-time attention blocks based on self-attention [20] and obtains very competitive results on various action recognition benchmarks [2]. While other Transformer-based video models, such as Swin [13] and ViViT [1], also achieve competitive performances on action recognition, TimeSFormer possesses the least amount of parameters, requiring only 60% parameters of Swin and only 40% parameters of ViViT. The final classifier is implemented as a single fully connected layer. Both the feature extractor and the subsequent classifier are shared across source and target data.

Methods	Sports-DA						Avg.
	$k=10$		$k=5$		$k=3$		
	U→S	KS→S	U→S	KS→S	U→S	KS→S	
UniFormer w/T	68.47	72.21	67.37	70.58	64.37	67.42	68.39
MMD	71.42	73.58	71.63	70.95	65.89	68.89	70.39
ACAN	71.37	74.21	73.28	72.11	67.68	68.26	71.15
d-SNE	71.79	73.74	73.05	74.95	69.32	68.05	71.82
SSA^{2}lign	**78.21**	**79.63**	**75.16**	**79.05**	**73.79**	**74.16**	**76.67**

Table 1. Comparison with the UniFormer-XXS [11] backbone.

It should also be noted that TimeSFormer is not the sole feature extractor available for SSA^{2}lign. To show that the superiority of SSA^{2}lign is model agnostic, we adopted a variant of UniFormer [11] (UniFormer-XXS) as a much more lightweight backbone and compared SSA^{2}lign with 3 other best performing methods (MDD [30], ACAN [25] and d-SNE [24]) under this backbone using the same benchmark settings as in Tables 4-6 of the paper for Sports-DA [28]. The results in Table 1 show that SSA^{2}lign still brings significant improvement under the UniFormer backbone, and confirms that the superiority of SSA^{2}lign is model agnostic.

Training Details and Hyper-parameters. For training, we initialize the TimeSFormer feature extractor from pre-trained weights obtained by pre-training on Kinetics-400 [10]. For more efficient training, we freeze the first 8 blocks of TimeSFormer, leaving the last 4 blocks to be fully trainable, with the learning rate set at 0.005. 5 additional

*Equal contribution.
†Corresponding author.

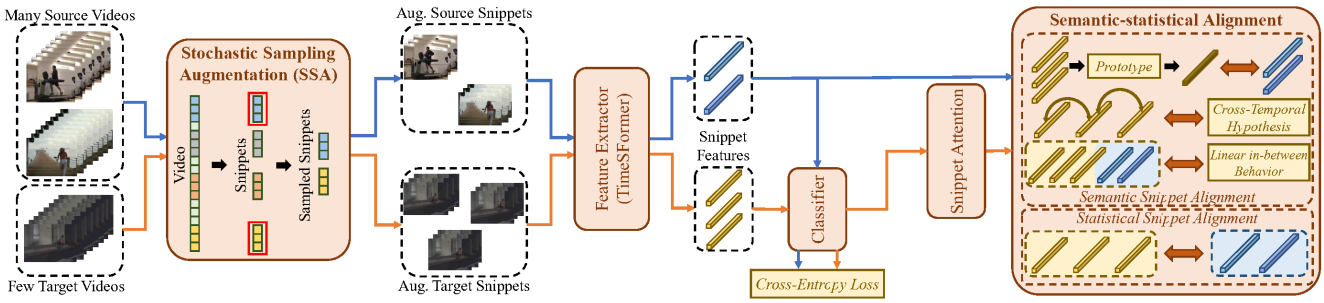

Figure 1. Pipeline of SSA^{2}lign. Source and target snippets are first obtained through the Stochastic Sampling Augmentation, whose features are obtained through the shared feature extractor. SSA^{2}lign then aligns the source and target domains at the snippet level with the Semantic-statistical Alignment, while weighing the impact of different target snippets through snippet attention, whose weight is built based on the output prediction of target snippets, obtained from a shared classifier with source snippets. The blue and orange lines imply the data flow for source and target videos respectively. *Best viewed in color.*

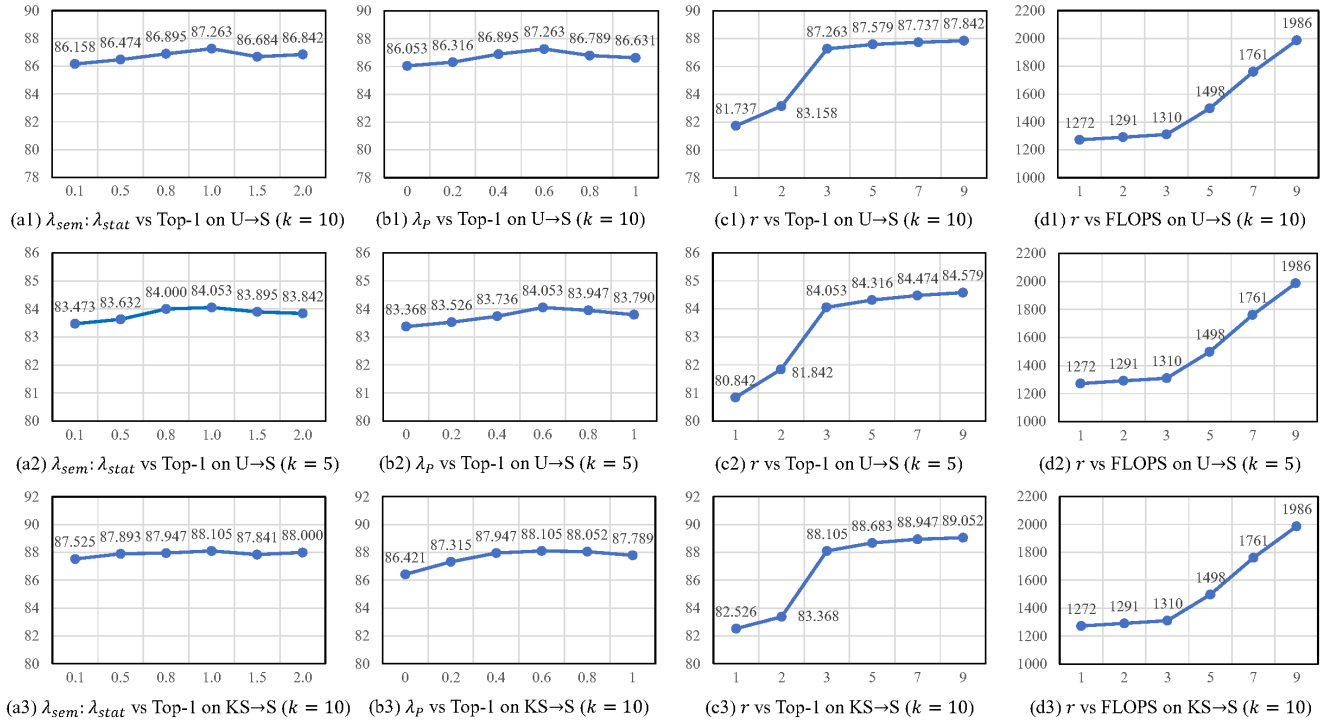

Figure 2. Hyper-parameter sensitivity on the U→S task with $k = 10$ (top), $k = 5$ (mid), and the KS→S task with $k = 10$ (bot).

layers with 10 parameters are trained from scratch, with their learning rates set to be 10 times that of the pretrained-loaded trainable layers (blocks). These are linear and batch-norm layers added to the TimeSFormer backbone, following prior UDA works [12, 27], and are not method-specific. In total, The trainable parameter size for the backbone and SSA^{2}lign is 41.53 M.

For the tasks constructed from the Daily-DA dataset [28], we train a total of 30 epochs, while we train a total of 50 epochs for tasks constructed from the Sports-DA dataset [28]. The stochastic gradient descent (SGD) algorithm [3] is used for optimization, with the weight decay set to 0.0001 and the momentum set to 0.9. During the training phase of SSA^{2}lign, the batch size is set to 24 input snippets per GPU, with 12 source snippets from 12 source videos and 12 target snippets from 4 target videos ($r = 3$ by default). For a fair comparison, the batch size is set to 24 input videos per GPU when training all comparing methods. All experiments are implemented with the PyTorch [16] library and conducted on 2 NVIDIA

A6000 GPUs. We set the length of snippets and the number of snippets per target video via SSA empirically as $m = 8, r = 3$. Hyper-parameters $\lambda_{sem} = 1.0$, $\lambda_{stat} = 1.0$ and $\lambda_P = 0.6$ are empirically set and are fixed. As shown in Section 4.3 and Fig. 2 of the paper, the performance of SSA^{2}lign is robust to hyper-parameters λ_{sem}, λ_{stat} and λ_P as well as r when $r \geqslant 3$, with minimal variations and maintains the best results with high computation efficiency with all the default hyper-parameter settings. To further illustrate the robustness of SSA^{2}lign towards the sensitivity of λ_{sem} and λ_{stat} which control the strength of the semantic and statistical snippet alignment losses, λ_P which relates to the update of target prototypes and r the number of snippets per target video, we present the additional results of hyper-parameter sensitivity analysis under different experimental settings. Specifically, we present the results of the U→S task with $k = 10$, $k = 5$ (the same as presented in Fig. 2 of the paper), and the results of the KS→S task with $k = 10$, as shown in Fig. 2 of this appendix.

The additional results further justify that SSA^{2}lign is robust to hyper-parameters λ_{sem}, λ_{stat} and λ_P as well as r when $r \geqslant 3$ under all examined experimental settings, while achieving the best results with high computation efficiency with the default hyper-parameter settings.

2. Cross-domain Action Recognition Benchmarks

In this paper, to evaluate our proposed SSA^{2}lign, we utilized two cross-domain action recognition benchmarks: the Daily-DA and Sports-DA [28]. In this section, we provide more details on each benchmark.

2.1. Daily-DA

The Daily-DA dataset is a recently proposed cross-domain action recognition dataset for VUDA [28]. It is more comprehensive and challenging compared to prior benchmarks such as UCF-Olympic [18] and UCF-HMDB$_{full}$ [5] which have resulted in saturated performance due to limited domains (only 2 domains in each dataset) and number of videos per domain. Daily-DA includes videos of daily actions from four domains and incorporates both normal videos and low-illumination videos. Specifically, Daily-DA is built from four datasets: the dark dataset ARID (A) [26], as well as HMDB51 (H), Moments-in-Time (M) [14], and Kinetics-600 (KD) [4], which are video datasets widely used for action recognition benchmarking [15]. Compared with other action recognition datasets such as Moments-in-Time and Kinetics, ARID is comprised of videos shot under adverse illumination conditions, characterized by low brightness and low contrast. Statistically, the RGB mean and standard deviation values (std) of videos in ARID are much lower among datasets leveraged in Daily-DA [25], which strongly suggests a larger domain shift between ARID and the other action recognition datasets. The Daily-DA includes a total of 16,295 training videos and 2,654 testing videos from 8 categories as listed in Table 2, with each category corresponding to one or more categories in the original datasets as demonstrated in Table 3.

2.2. Sports-DA

To further demonstrate the efficacy of our proposed SSA^{2}lign on large-scale cross-domain datasets, we further adopt the Sports-DA dataset as another cross-domain action recognition benchmark. Comparatively, Sports-DA contains almost double the amount of training and testing videos of Daily-DA. Specifically, it includes a total of 36,003 training videos and 4,721 testing videos from 23 categories of sports actions, collected from three large-scale datasets: UCF101 (U) [17], Sports-1M (S) [9], and Kinetics-600 (KS) [4], as shown in Table 2. Similar to Daily-DA, each action class corresponds to one or more categories in the original datasets as presented in Table 4. With more than 40,000 training and testing videos, the Sports-DA benchmark is one of the largest cross-domain action recognition benchmarks introduced.

3. Detailed Comparison with Related FS(V)DA and (V)UDA Methods

In this paper, we proposed SSA^{2}lign to address the more realistic and challenging FSVDA task, which achieves state-of-the-art performances with outstanding improvements on both cross-domain action recognition benchmarks (average 13.1% on Daily-DA tasks and average 4.2% on Sports-DA tasks). To further highlight the novelty of SSA^{2}lign, we compare our proposed SSA^{2}lign with prior FSDA/FSVDA and UDA/VUDA methods. Specifically, we compare with d-SNE [24] proposed for FSDA, PASTN [7] designed for FSVDA, ACAN [25] introduced for ACAN, and DM-ADA [23] which is an image-based UDA method that leverages MixUp [29]. These methods are all compared from two perspectives: the tasks they tackle and the techniques leveraged, as displayed in Table 5.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 1

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 1

Statistics	Daily-DA	Sports-DA
Video Classes #	8	23
Training Video #	A:2,776 / H:560 / M:4,000 / KD:8,959	U:2,145 / S:14,754 / KS:19,104
Testing Video #	A:1,289 / H:240 / M:400 / KD:725	U:851 / S:1,900 / KS:1,961

Table 2. Summary of cross-domain action recognition benchmarks statistics.

Class ID	ARID Class	HMDB51 Class	Moments-in-Time Class	Kinetics-600 Class
0	Drink	drink	drinking	drinking shots
1	Jump	jump	jumping	jumping bicycle, jumping into pool, jumping jacks
2	Pick	pick	picking	picking fruit
3	Pour	pour	pouring	pouring beer
4	Push	push	pushing	pushing car, pushing cart, pushing wheelbarrow, pushing wheelchair
5	Run	run	running	running on treadmill
6	Walk	walk	walking	walking the dog, walking through snow
7	Wave	wave	waving	waving hand

Table 3. List of action classes for Daily-DA.

Class ID	UCF101 Class	Sports-1M Class	Kinetics-600 Class
0	Archery	archery	archery
1	Baseball Pitch	baseball	catching or throwing baseball, hitting baseball
2	Basketball Shooting	basketball	playing basketball, shooting basketball
3	Biking	bicycle	riding a bike
4	Bowling	bowling	bowling
5	Breaststroke	breaststroke	swimming breast stroke
6	Diving	diving	springboard diving
7	Fencing	fencing	fencing (sport)
8	Field Hockey Penalty	field hockey	playing field hockey
9	Floor Gymnastics	floor (gymnastics)	gymnastics tumbling
10	Golf Swing	golf	golf chipping, golf driving, golf putting
11	Horse Race	horse racing	riding or walking with horse
12	Kayaking	kayaking	canoeing or kayaking
13	Rock Climbing Indoor	rock climbing	rock climbing
14	Rope Climbing	rope climbing	climbing a rope
15	Skate Boarding	skateboarding	skateboarding
16	Skiing	skiing	skiing crosscountry, skiing mono
17	Sumo Wrestling	sumo	wrestling
18	Surfing	surfing	surfing water
19	Tai Chi	t'ai chi ch'uan	tai chi
20	Tennis Swing	tennis	playing tennis
21	Trampoline Jumping	trampolining	bouncing on trampoline
22	Volleyball Spiking	volleyball	playing volleyball

Table 4. List of action classes for Sports-DA.

[3] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. 2

[4] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 3

[5] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6321–6330, 2019. 3

[6] Carlotta Domeniconi, Dimitrios Gunopulos, and Jing Peng. Large margin nearest neighbor classifiers. *IEEE transactions on neural networks*, 16(4):899–909, 2005. 5

[7] Zan Gao, Leming Guo, Weili Guan, An-An Liu, Tongwei Ren, and Shengyong Chen. A pairwise attentive adversarial spatiotemporal network for cross-domain few-shot action recognition-r2. *IEEE Transactions on Image Processing*, 30:767–782, 2020. 3, 5

[8] Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15, 2002. 5

[9] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 3

[10] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman,

Method	Publication	Task	Techniques
d-SNE [24]	CVPR-19	Few-Shot Domain Adaptation (FSDA): source image data available with labels, **a few** (very limited) target image data available with labels, image-based.	(a) d-SNE learns a latent domain-agnostic space through SNE [8] with large-margin nearest neighborhood [6]; (b) d-SNE conducts FSDA in a min-max formulation with a modified-Hausdorff distance; (c) d-SNE creates sibling target samples with spatial augmentations, and trains feature extractor with the Mean-Teacher technique [19].
PASTN [7]	TIP-20	Few-Shot Video Domain Adaptation (FSVDA): source video data available with labels, **a few** (very limited) target video data available with labels, video-based.	(a) PASTN obtains video features from a frame-based video model; (b) PASTN forms source-target video pairs to address insufficient target video data; (c) PASTN constructs pairwise adversarial networks performed across source-target video pairs optimized by a pairwise margin discrimination loss [22].
DM-ADA [23]	AAAI-20	Unsupervised Domain Adaptation (UDA): source image data available with labels, sufficient target images available without labels, image-based.	(a) DM-ADA augments the target domain with the source domain by domain mixup [29]; (b) DM-ADA improves the feature extractor by leveraging soft domain labels; (c) DM-ADA jointly trains a domain discriminator which judges the samples' differences relative to the two domains with refined scores.
ACAN [25]	TNNLS-22	Video Unsupervised Domain Adaptation (VUDA): source video data available with labels, sufficient target videos available without labels, video-based.	(a) ACAN applies adversarial-based domain adaptation across spatio-temporal video features; (b) ACAN additionally aligns video correlation features in the form of long-range spatiotemporal dependencies [21]; (c) ACAN further aligns the joint distribution of correlation information of different domains by minimizing pixel correlation discrepancy (PCD).
SSA^{2}lign (Ours)	-	Few-Shot Video Domain Adaptation (FSVDA): source video data available with labels, **a few** (very limited) target video data available with labels, video-based.	(a) SSA^{2}lign addresses FSVDA at the **snippet level** instead of the frame or video-levels; (b) SSA^{2}lign augments target domain data and the snippet-level alignments by a simple yet effective stochastic sampling of snippets; (c) SSA^{2}lign performs both semantic and statistical alignments attentively, with the semantic alignments achieved by alignment based on the semantic information within each snippet, cross-snippets of each video, and across snippet-level data distribution.

Table 5. Detailed comparison of SSA^{2}lign with related but different FS(V)DA and (V)UDA methods.

and Andrew Zisserman. The kinetics human action video dataset, 2017. 1

[11] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning. *arXiv preprint arXiv:2201.04676*, 2022. 1

[12] Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[13] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 1

[14] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019. 3

[15] Preksha Pareek and Ankit Thakkar. A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, 54:2259–2322, 2021. 3

[16] 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. 2

[17] 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. 3

[18] Waqas Sultani and Imran Saleemi. Human action recognition across datasets by foreground-weighted histogram decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 764–771, 2014. 3

[19] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 5

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia

Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[21] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 5

[22] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE international conference on computer vision*, pages 2840–2848, 2017. 5

[23] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6502–6509, 2020. 3, 5

[24] Xiang Xu, Xiong Zhou, Ragav Venkatesan, Gurumurthy Swaminathan, and Orchid Majumder. d-sne: Domain adaptation using stochastic neighborhood embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2497–2506, 2019. 1, 3, 5

[25] Yuecong Xu, Haozhi Cao, Kezhi Mao, Zhenghua Chen, Lihua Xie, and Jianfei Yang. Aligning correlation information for domain adaptation in action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 1, 3, 5

[26] Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. Arid: A new dataset for recognizing action in the dark. In *International Workshop on Deep Learning for Human Activity Recognition*, pages 70–84. Springer, 2021. 3

[27] Yuecong Xu, Jianfei Yang, Haozhi Cao, Keyu Wu, Min Wu, and Zhenghua Chen. Source-free video domain adaptation by learning temporal consistency for action recognition. In *European Conference on Computer Vision*, pages 147–164. Springer, 2022. 2

[28] Yuecong Xu, Jianfei Yang, Haozhi Cao, Keyu Wu, Min Wu, Zhengguo Li, and Zhenghua Chen. Multi-source video domain adaptation with temporal attentive moment alignment network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1, 2, 3

[29] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018. 3, 5

[30] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413. PMLR, 2019. 1