Appendix of the Paper (Differentiated Learning for Multi-Modal Domain Adaptation)

Jianming Lv*
South China University of Technology
Guangzhou, Guangdong, China
jmlv@scut.edu.cn

Kaijie Liu
South China University of Technology
Guangzhou, Guangdong, China
cskaijieliu@gmail.com

Shengfeng He
South China University of Technology
Guangzhou, Guangdong, China
shengfenghe7@gmail.com

ACM Reference Format:
Jianming Lv, Kaijie Liu, and Shengfeng He . 2021. Appendix of the Paper (Differentiated Learning for Multi-Modal Domain Adaptation). In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3474085.3475660

1 MORE DETAILS ABOUT DATASETS

(Extension of Section 4.1)

Event Recognition Dataset (AVE) [13]. AVE dataset contains 4,143 videos covering 28 event categories, and image modal and audio modal are aligned in time. The AVE dataset covers a wide range of audio-visual events (e.g., man speaking, dog barking, playing guitar, and frying food etc.), and each video contains at least one 2s long audio-visual event. We use the division method introduced in Section 4.1 to gain the source domain and the target domain. Specifically, the Resnet-50 network [5] pre-trained on Imagenet [3] is used to extract 1024-dimensional features of the image of each sample. Then the feature vectors of each category are clustered into two clusters by the K-Means algorithm [6]. The samples of each category are divided into the source domain and the target domain at a ratio of 2:8 according to the distance from their image feature vectors to the cluster centers. In the end, we obtained 83,458 source domain samples and 333,829 target domain samples. Examples of 12 different categories of images in the source domain and the target domain are shown in Fig 2. It can be seen from Fig 2 that the images in source domain are clearly distinguishable, while images in target domain are more difficult to distinguish due to the poor lighting conditions or occlusion. This shows the obvious domain shift between the source domain and the target domain.

Fatigue Detection Dataset (CogBeacon) [10]. CogBeacon is a multi-modal dataset designed to research the effects of cognitive fatigue in human performance. The dataset consists of 76 sessions collected from 19 male and female users performing different versions of a cognitive task inspired by the principles of the Wisconsin Card Sorting Test. During each session, the users' EEG functionality and facial keypoints are recorded and labeled. Specifically, each user performed three versions (namely V1,V2,V3) of cognitive task tests. Different versions of cognitive task will produce different stimuli to users. For example, the EEG signals when facing text-based stimuli and sound-based stimuli are different [10]. Therefore, the data collected under different versions of cognitive task can be regarded as cross-domain data. In this paper, we choose one version of cognitive task as the source domain, and the other version as the target domain. The number of samples corresponding to the cognitive task V1,V2,V3 are 2,259, 2,221 and 2,389 respectively. In the experiment, we regard one of the domains as the source domain and the other domain as the target domain. This setting method can obtain six domain migration combinations.

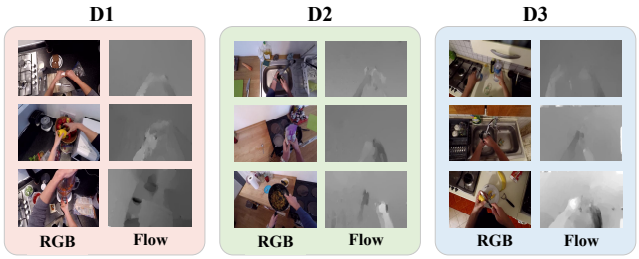

Figure 1: Examples of the instances in the domain D1, D2 and D3 in EPIC Kitchens.

Action Recognition Dataset (EPIC Kitchens) [2]. EPIC Kitchens is a multi-modal dataset designed to test domain adaptation for action recognition. It is recorded in 32 environments and contains two modal forms of RGB image and Optical Flow. In this paper, we use the same division method as the previous work [9], considering the domain adaptation problem among the three domains D1, D2 and D3 in EPIC Kitchens. Some scenes of image and optical flow modals in the three domains are shown in Fig 1, which reflects the shift between domains. Eight types of actions are analyzed: ('put', 'take', 'open', 'close', 'wash', 'cut', 'mix', and 'pour'). The number of action segments in the three domains D1, D2, and D3 are 1,978, 3,245 and 4,871 respectively. Even this ensures sufficient examples per domain and class, EPIC-Kitchens has a large class imbalance offering additional challenges for domain adaptation. Six domain migration combinations can be obtained through combining different domains.

2 MORE RESULTS OF MIGRATION BETWEEN DIFFERENT DOMAINS

(Extension of Section 4.3)

*Corresponding Author

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475660

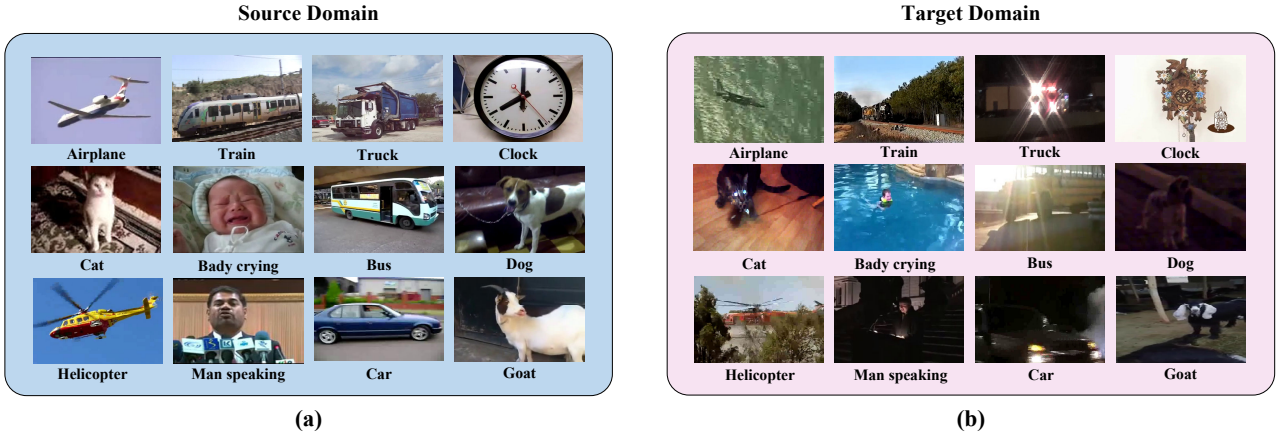

Figure 2: Examples of the images in AVE . (a) Examples in the source domain. (b) Examples in the target domain.

Table 1: Performance comparison on the CogBeacon Dataset. $V1$,$V2$,and $V3$ indicate three different domains.

Method	V1→V2			V2→V3			V3→V1			Mean		
	FK	EEG	Fusion	FK	EEG	Fusion	FK	EEG	Fusion	FK	EEG	Fusion
Direct-Transfer	56.27	59.03	59.31	61.46	63.15	64.09	62.41	65.41	66.41	60.05	62.53	63.27
DANN[4]	57.81	60.12	61.52	63.02	64.87	66.16	64.95	67.77	69.81	61.93	63.31	65.83
CBST[15]	59.61	62.19	63.56	64.05	64.98	66.35	65.82	70.18	71.16	63.16	65.78	67.46
MCD[12]	58.04	61.75	62.83	65.34	66.71	67.65	65.67	69.13	70.25	63.02	65.83	66.91
CT[1]	57.98	60.42	61.93	62.97	64.47	65.13	65.33	67.01	69.34	62.09	63.97	65.47
eCT[14]	58.45	61.23	62.29	62.99	64.50	65.56	65.72	69.51	70.57	62.38	65.08	66.14
MDANN[11]	-	-	62.45	-	-	66.72	-	-	69.92	-	-	66.36
MM-SADA[9]	58.74	62.83	63.47	65.14	66.57	67.11	67.92	69.72	70.69	63.93	66.37	67.09
DLMM-prob	58.02	61.53	62.53	65.09	66.52	67.14	67.51	69.59	70.28	63.54	65.88	66.65
DLMM-entropy	58.17	61.86	63.07	65.26	67.07	68.32	68.47	71.26	72.47	63.97	66.73	67.95
DLMM-margin	58.06	61.62	62.86	65.37	69.17	68.84	68.36	70.89	72.03	63.93	67.23	67.91
DLMM-Separate	59.77	62.93	64.15	67.28	67.61	69.15	69.82	71.85	72.82	65.62	67.46	68.71
DLMM	**59.91**	**64.82**	**66.21**	**67.36**	**70.02**	**71.80**	**70.15**	**73.67**	**74.89**	**65.91**	**69.50**	**70.97**
Supervised	79.34	85.16	87.34	82.10	85.67	86.59	83.46	87.82	89.12	81.63	86.15	87.68

Method	V2→V1			V3→V2			V1→V3			Mean		
	FK	EEG	Fusion	FK	EEG	Fusion	FK	EEG	Fusion	FK	EEG	Fusion
Direct-Transfer	60.13	62.55	63.28	61.77	63.21	63.63	62.21	63.75	64.13	61.37	62.92	64.01
DANN[4]	63.02	65.54	66.27	64.32	66.55	67.23	65.49	67.13	68.02	64.28	67.01	67.17
CBST[15]	64.37	66.09	66.63	65.86	67.79	68.81	66.23	67.76	68.45	65.49	67.21	67.96
MCD[12]	63.36	64.76	65.29	65.15	66.33	66.67	65.12	66.34	67.83	64.55	65.81	66.60
CT[1]	63.44	65.13	65.75	65.39	66.54	66.88	65.78	67.12	68.37	64.87	66.26	67.01
eCT[14]	63.51	65.21	66.05	66.02	67.64	68.78	66.31	67.54	68.85	65.28	66.79	67.89
MDANN[11]	-	-	65.86	-	-	67.51	-	-	68.53	-	-	67.30
MM-SADA[9]	64.65	66.33	67.04	66.17	67.35	68.40	66.56	68.21	69.76	65.79	67.31	68.40
DLMM-p	63.23	65.15	65.87	65.34	66.86	67.68	66.11	67.53	69.41	64.89	66.51	67.65
DLMM-e	64.15	65.65	66.08	65.82	67.08	67.76	66.07	67.47	69.23	65.34	66.73	67.69
DLMM-m	63.34	65.17	65.95	65.55	67.13	67.83	66.34	67.85	69.55	65.07	66.68	67.77
DLMM-separate	64.01	66.34	66.86	66.42	68.27	69.73	67.05	68.01	70.21	65.83	67.54	68.93
DLMM	**64.78**	**67.21**	**67.63**	**66.53**	**68.76**	**70.24**	**68.53**	**69.17**	**70.13**	**66.62**	**68.38**	**69.67**
Supervised	81.12	83.16	85.43	82.54	84.89	86.72	79.72	82.55	84.12	81.13	83.53	85.42

Since there are three domains in both CogBeacon dataset and EPIC Kitchens dataset, which constitutes six different domain migration combinations. The results of the three domain migration combinations have been shown in the main text, and the complete results of all combinations are shown in Table 1 and Table 2. It can be observed that *DLMM* still achieves the highest accuracy among all indicators, which also confirms our analysis of the advantages of *DLMM* in the main text.

REFERENCES

[1] Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*. 92–100.

[2] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 720–736.

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

[4] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17, 1 (2016), 2096–2030.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

Table 2: Performance comparison on the EPIC Kitchens Dataset. $D1$,$D2$,**and** $D3$ **indicate three different domain.**

Method	D1→D2			D2→D3			D3→D1			Mean		
	RGB	Flow	Fusion	RGB	Flow	Fusion	RGB	Flow	Fusion	RGB	Flow	Fusion
Direct-Transfer	36.1	45.6	43.7	33.6	46.0	46.5	36.3	44.2	44.5	35.3	45.3	44.9
AdaBN[7]	44.1	46.5	47.0	44.8	48.3	48.7	41.5	45.0	47.8	43.5	46.6	47.8
MMD[8]	43.7	46.3	46.5	44.5	48.2	48.5	41.7	45.4	48.3	43.3	46.6	47.7
MCD[12]	43.5	46.3	46.4	45.8	50.7	51.0	42.0	45.0	47.9	43.8	47.3	48.4
CT[1]	43.7	46.0	46.1	45.3	50.2	50.7	41.6	44.8	47.3	43.5	47.0	48.0
eCT[14]	43.9	46.2	46.3	45.5	50.2	50.8	41.8	44.9	47.6	43.7	47.1	48.2
MDANN[11]	-	-	45.7	-	-	48.6	-	-	48.2	-	-	47.5
MM-SADA[9]	45.0	49.0	49.5	46.2	52.1	52.7	42.1	45.7	50.9	44.5	48.9	51.1
DLMM-prob	45.8	48.7	49.2	45.3	50.4	50.8	42.0	45.1	48.7	44.4	48.1	49.6
DLMM-entropy	46.0	48.8	49.6	45.7	51.2	51.6	42.2	45.3	49.3	44.6	48.4	50.2
DLMM-margin	45.6	48.7	49.1	45.3	50.7	51.2	42.1	46.1	49.5	44.3	48.5	49.9
DLMM-Separate	46.2	49.8	50.1	47.1	50.6	51.5	42.2	48.8	51.3	45.2	49.7	51.1
DLMM	**48.3**	**52.0**	**52.7**	**49.7**	**54.6**	**55.8**	**46.9**	**51.3**	**53.5**	**48.3**	**52.6**	**54.0**
Supervised	63.1	67.6	71.7	62.2	68.2	74.0	60.9	62.4	62.8	62.1	66.1	69.5

Method	D2→D1			D3→D2			D1→D3			Mean		
	RGB	Flow	Fusion	RGB	Flow	Fusion	RGB	Flow	Fusion	RGB	Flow	Fusion
Direct-Transfer	37.0	44.6	42.5	44.8	54.0	55.1	36.6	41.1	41.2	39.5	46.6	46.3
AdaBN[7]	39.1	43.9	44.6	44.9	53.2	54.7	36.7	40.2	40.3	40.2	45.8	46.5
MMD[8]	38.7	42.8	43.1	45.3	54.6	55.2	36.1	39.0	39.5	40.0	45.5	45.9
MCD[12]	37.5	41.6	42.1	45.0	53.1	53.9	38.4	42.8	43.5	40.3	45.8	46.5
CT[1]	38.8	44.7	43.8	45.6	54.4	55.3	37.3	41.3	41.6	40.6	46.8	46.9
eCT[14]	39.5	45.2	44.7	46.1	55.0	55.7	38.0	42.1	42.4	41.2	47.4	47.6
MDANN[11]	-	-	47.5	-	-	54.9	-	-	42.3	-	-	48.2
MM-SADA[9]	41.7	45.0	48.2	48.4	55.9	56.1	39.7	44.8	44.1	43.3	48.6	49.5
DLMM-p	40.7	44.8	46.5	46.7	53.8	54.7	38.1	41.9	42.2	41.8	46.8	47.8
DLMM-e	41.3	44.5	47.7	47.3	54.2	55.8	39.2	42.7	43.1	42.6	47.1	48.8
DLMM-m	41.2	44.6	47.5	47.2	54.1	55.2	39.5	42.6	43.2	42.6	47.1	48.6
DLMM-separate	42.7	45.1	48.3	48.2	54.4	55.4	40.6	43.1	43.7	43.8	47.5	49.1
DLMM	**44.9**	**47.3**	**50.4**	**49.2**	**56.6**	**57.0**	**42.2**	**45.0**	**44.9**	**45.4**	**49.6**	**50.8**
Supervised	59.7	61.3	62.8	65.7	68.9	71.7	70.2	72.4	74.0	65.2	67.5	69.5

[6] Anil K Jain. 2010. Data clustering: 50 years beyond K-means. *Pattern recognition letters* 31, 8 (2010), 651–666.

[7] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. 2018. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition* 80 (2018), 109–117.

[8] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*. PMLR, 97–105.

[9] Jonathan Munro and Dima Damen. 2020. Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 122–132.

[10] Michalis Papakostas, Akilesh Rajavenkatanarayanan, and Fillia Makedon. 2019. CogBeacon: A Multi-Modal Dataset and Data-Collection Platform for Modeling Cognitive Fatigue. *Technologies* 7, 2 (2019), 46.

[11] Fan Qi, Xiaoshan Yang, and Changsheng Xu. 2018. A unified framework for multimodal domain adaptation. In *Proceedings of the 26th ACM international conference on Multimedia*. 429–437.

[12] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3723–3732.

[13] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 247–263.

[14] Zixing Zhang, Fabien Ringeval, Bin Dong, Eduardo Coutinho, Erik Marchi, and Björn Schüller. 2016. Enhanced semi-supervised learning for multimodal emotion recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5185–5189.

[15] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*. 289–305.