# Domain Adaptable Normalization for Semi-Supervised Action Recognition in the Dark

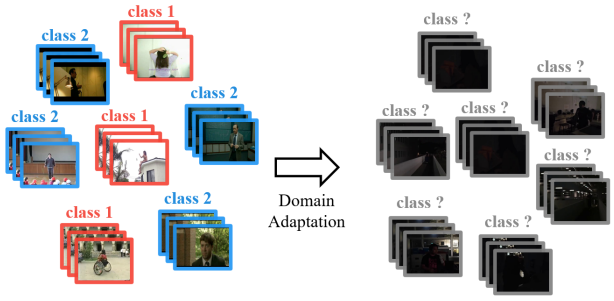
Zixi Liang* zixiliang@qq.com
Jiajun Chen* Ticuby@qq.com
Rui Chen 2584775140@qq.com
Bingbing Zheng zhengbingbingjuli@163.com
Mingyue Zhou raphaelzhou0725@gmail.com
Huaien Gao gao.huaien@hotmail.com
Shan Lin† alice333.happy@163.com
Guangzhou Xi Ma Information Technology Company, Guangzhou, China

## Abstract

*Action recognition in the dark is gaining more and more attention with the rapid development of intelligent recognition applications in real-world applications,* e.g. *self-driving at night and night surveillance. However, limited by the expensive labeling cost, it is impractical to produce a large-scale labeled dataset only for dark environments. Therefore, a practical solution adopted is to transfer models trained from clear environments to dark environments through semi-supervised learning. However, prior works rely heavily on additional efforts such as extra annotations, or extra sensors. To this end, we proposed a novel and simple Domain Adaptable Normalization (DANorm) method to align different domains directly, which consists of feature normalization, angle constraint and the Pseudo-Label. Specifically, the proposed DANorm method enables the model automatically learning the associated features between labeled source domain and unlabeled target domain by constraining the feature subspace vectors. Experimental results show that our model achieves superiority performance on Semi-supervised ARID dataset. Code is available at:* https://github.com/NikkiElwin/DANorm.

## 1. Introduction

Action Recognition (AR) has received notable attention because of its great success in rich real-world applications such as video surveillance [5,8,35] and human computer interaction [1, 11, 21]. However, existing methods often generalize poorly to dark environments, partly due to the fact the labeled dark data is limited and costly. Although multi-view methods [15, 17, 20, 31] by leveraging on extra sensors can effectively relieve the performance degradation caused by dark environments, the high cost restricts them for large-scale use in the real-world. Thanks to unlabeled dark data is easily available, Semi-Supervised Action Recognition in the Dark (SS-ARID) methods have been proposed to solve the model degradation caused by the adverse visual condition (see Fig. 1).




(a) Labeled Source Domain (b) Unlabeled Target Domain

Figure 1. Semi-Supervised Action Recognition in the Dark (SS-ARID). For the given labeled set in source domain and the unlabeled set in target domain, the main task of SS-ARID is to obtain satisfactory classification performance in the target domain when only the source domain sample ground true is available.

Technically speaking, SS-ARID can be viewed as a Video-based Unsupervised Domain Adaptation (VUDA) [33] for Action Recognition in the Dark task from the labeled source domain (clear videos) to the unlabeled target domain (dark videos). This is a challenging task due to the unknown discriminative class boundaries on target domain, and the low brightness and contrast characteristic in dark videos [34]. Most current works for SS-ARID can be simply divided into two types: (i) Data Domain Adaptation [3, 9, 34], and (ii) Feature Domain Adaptation [2, 4, 22, 33]. Data domain adaptation aims to produce visually clearer video frames through frame enhancement methods [7, 10, 32], achieving adaptation from dark data domain to clear data domain. However, because the data domain in-

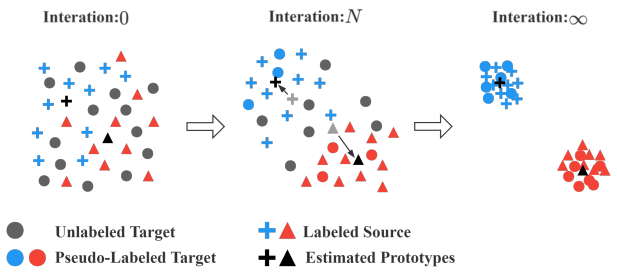



Figure 2. Example of Domain Adaptable Normalization method. For the given source domain and target domain, each iteration can be divided into two steps: (i) minimize the task loss and the feature normalization loss on the training set, and (ii) add target domain data with high classification scores to the training set. All labeled data in the source domain is added to the training set at the beginning.

formation is noisy and surplus, inappropriate data augmentation methods will destroy the original data domain distribution, resulting in the inferior performance [34]. Feature domain adaptation focuses on aligning the two domains on the feature space, but most of their performances are not satisfactory. The lack of category information in the target domain during supervised training will cause the feature subspace to be ineffective, resulting in poor model performance.

Recently, feature subspace constraint methods [24, 25, 28, 36] have been proposed to align different domains statistically. Most approaches are based on optimizing a constraint loss (*e.g*. Conditional Entropy [24] and Mean Discrepancy [28]) on the feature subspace of source and target domain, as well as the task loss on the labeled data. These methods have achieved good results in experiments, but the premise is there are at least labeled samples in the target domain, which doesn't exist in the task of SS-ARID, where the target domain is unlabeled.

To this end, we proposed a novel and simple method, named Domain Adaptable Normalization (DANorm), to achieve the alignment of the two domains when the target domain is unlabeled. Specifically, the proposed method uses feature normalization, which forces the target domain to align with the source domain. And then the method adds pseudo-labels to unlabeled target data, which enabling the model to learn discriminative class boundaries on the target domain. Experimental results show that DANorm can align the target domain with the source domain well, and achieve superior performance on the target domain.

In summary, the major contributions of the paper are summarized as follows: (i) we proposed a novel and simple Domain Adaptable Normalization (DANorm) method to explore the associated features between labeled source domain and unlabeled target domain, (ii) we explain the effective improvement brought by feature normalization and angle constraint methods, and (iii) the proposed method achieved the state-of-the-art result on Semi-supervised ARID [34] dataset.

## 2. Related Work

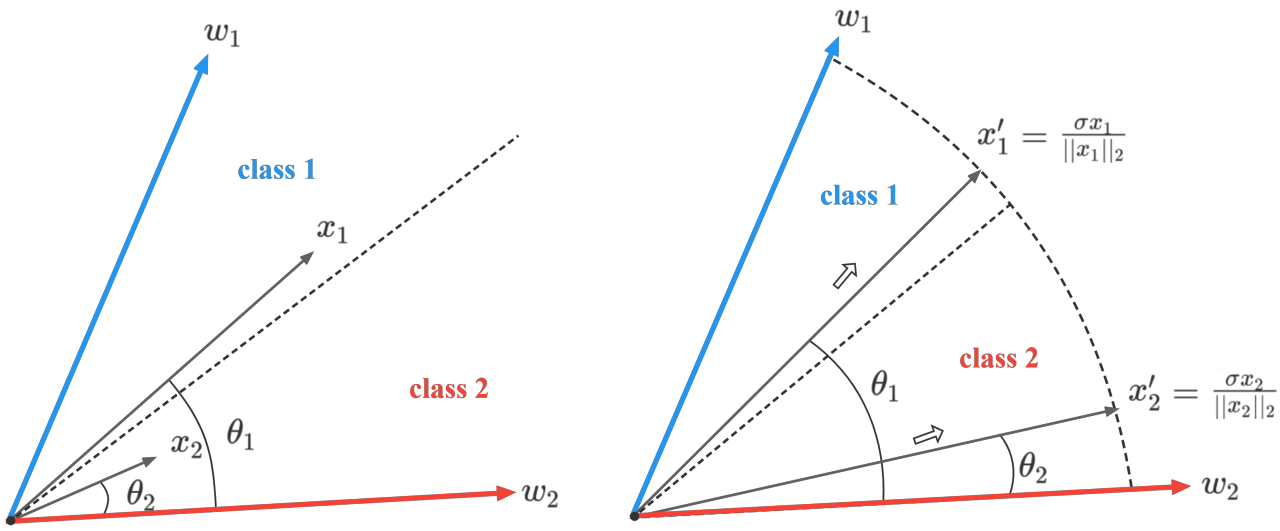
### 2.1. Video-based Unsupervised Domain Adaptation (VUDA)

Unsupervised Domain Adaptation for Action Recognition methods aim to transfer knowledge learned from labeled source domains to target domains with unlabeled data only. Depending on the relationship between source and target domains, this task can be divided into two different types: (i) across different capturing conditions and (ii) across different modalities. Across different capturing conditions means the source and target domain are captured in different environmental conditions, *e.g*. sunny and haze. Chen *et al*. [2] proposed a novel method to efficiently sifting out suitable features to align domains in space and time. Jinwoo *et al*. [4] combined the domain adversarial loss and the clip order prediction loss to encourage learning of representations which focus on the humans and objects involved in the actions. Sanchit *et al*. [9] introduced a delta sampling method, with Zero-DCE [7] enhanced technique to convert dark videos into clear ones. Across different modalities represents the source and target domain are captured in different sensors, and this type of adaptation is generally used to enhance fusion abilities in multi-view methods. Wang *et al*. [31] introduced a generative adversarial network to explore the potential between multi-view features. Liu *et al*. [17] considered the label-level and feature-level fusion simultaneously in a unified framework, to better align the feature distributions across different views. Liang *et al*. [15] proposed a novel network to reproduce each view's latent representation and bridge the semantic gap between two different views.

In this work, we mainly focus on the adaptation across different capturing conditions, proposes the Domain Adaptable Normalization (DANorm), to achieve the alignment of the clear videos and dark videos.

### 2.2. Feature Subspace Constraint

Feature subspace constraint methods focus on how to achieve a more robust semantic representation by regularizing feature vectors. Generally speaking, cross entropy loss can effectively optimize the distance between classes, but it is not satisfactory when optimizing the intra-class distance. Liu *et al*. [16] proposed a loss to guide the network to learn features with small intra-class distances and large inter-class distances. Wang *et al*. [30] studied the effect of feature normalization during training, and proposed a regularization method that aligns training samples and test sam-




(a) Without Feature Normalization

(b) With Feature Normalization

Figure 3. A simple case of binary semi-supervised classification task (a) without feature normalization and (b) with feature normalization. $w_1$ represents the cluster prototype of class 1 and $x_1$ represents the unlabeled sample with the ground true category is class 1, vice versa for towards the class 2 cluster prototype $w_2$ and the unlabeled sample $x_2$ in figure. Feature normalization can push the unlabeled data feature closer the cluster prototype, reducing the distance between the source and target domains. Note that the dotted line represents the decision boundary between the two classes and $\sigma$ is a scale hyperparameter.

ples. Ranjan *et al*. [23] proposed an $\mathcal{L}_2$-constrained method to make the model pay the same attention to samples of different quality. Zheng *et al*. [36] proposed a simple convex feature regularization method to get more robust features.

Compared to the previous methods, our proposed DANorm combined the L2 normalization and the novel Dot Product Ring Loss (DP-Ring Loss). The proposed DP-Ring Loss which can align the target domain with the source domain well, and achieve the superior performance on the target domain.

## 3. Method

### 3.1. Motivation

Consider a simple case of binary semi-supervised classification task with categories class 1 and class 2, two cluster prototypes $w_1$, $w_2$ and two unlabeled features $x_1$, $x_2$. Among them the ground true of $x_1$ is class 1 and the ground true of $x_2$ is class 2 (see Fig. 3a). And the example is trained with cross entropy loss. Generally, the dot product $w^T x$ of cluster prototype $w$ and feature $x$ represents the score for that category. Therefore, for $w_2$ pick the correct unlabeled feature $x_2$, it is necessary to require $w_2^T x_2 > w_2^T x_1 \Longrightarrow ||x_2||_2 \cos\theta_2 > ||x_1||_2 \cos\theta_1$, where $\theta_1$, $\theta_2$ are the angles between $x_1$, $x_2$ and $w_2$. However, due to the unclear features because of different causes such as unclear outlines of actors [34] in the dark videos, the target features would be ineffective with a low L2-norm [23]. The ineffective representation may result in $||x_2||_2 << ||x_1||_2 \Longrightarrow ||x_2||_2 \cos\theta_2 < ||x_1||_2 \cos\theta_1$, which would lead to the mis-classification of $x_1$ as $w_2$. The error data will lead to semi-supervised algorithms get worse results, such as dot product similarity based Metric Learning. Therefore, the key to the solution is changing the feature vectors length $||x||_2$ or its corresponding angle.

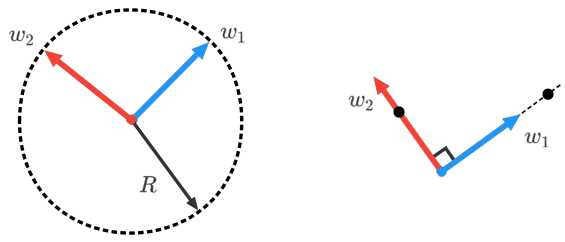
### 3.2. Feature Normalization

To avoid the above mis-classification, our proposed DANorm method simply forces $||x_1||'_2 \equiv ||x_2||'_2$ by $\mathcal{L}_2$ normalization $||x||' = \frac{\sigma x}{||x||_2}$ (see Fig. 3b). Among them $\sigma$ is a scale hyperparameter to ensure the normal convergence of the network. $\mathcal{L}_2$ normalization is used as a constraint on the features to strict their $\mathcal{L}_2$-norm to a constant. Under the $\mathcal{L}_2$-norm constraint, the category score $w^T x$ can be further reduced to the cosine similarity of the feature and the cluster prototype.

On the other hand, $\mathcal{L}_2$ normalization also balances the low $\mathcal{L}_2$-norm target domain features and the high $\mathcal{L}_2$-norm source domain ones. Specifically, this normalization method constrains the source and target domains to the same hypersphere. This constraint effectively solves the mis-classification caused by ineffective representations as in Fig. 3a), thereby improving the classification performance of semi-supervised algorithms such as Pseudo-

Label. [14].

### 3.3. Angle Constraint




(a) Ring Loss (b) DP-Ring Loss

Figure 4. Constraints on (a) Ring loss [36] and (b) our DP-Ring loss. Ring loss constraints the feature a hypersphere with learnable radius $R$, while DP-Ring loss constraints them to a point at the cluster prototype direction as in the black point in Fig. 4b. Note that to let $||Wx||_2 \equiv R$, DP-Ring loss also forces different clustering prototypes orthogonal to each other, where $R$ is also a learnable parameter.

Besides the $\mathcal{L}_2$ normalization, another way to prevent mis-classification is to add constraints on corresponding angles between feature $x$ and prototype $w$. Inspired by [36], the proposed DANorm method uses the novel Dot Product Ring (DP-Ring) loss as an angle constraint criteria formulated as:

$$\mathcal{L}_{DP-Ring} = \frac{\lambda}{2}\mathbb{E}_{x\sim p_x}(||Wx||_2 - R)^2 \tag{1}$$

where the labeled feature vector $x$ is sampled from training feature distribution $p_x$, $R$ represents the learn-able parameter and $\lambda$ is defined to balance the terms of difference losses. $W = [w_1, w_2, \cdots, w_n]^T$ is the $n \times m$ matrix combined with different category prototypes $w_1, w_2, \cdots, w_n$, where $m$ is the dimension of $x$.

Different from the original Ring loss [36], the proposed DP-Ring loss focuses more on the relationship between features and clustering prototypes. In semi-supervised learning, since part of training samples ground-truth are unlabeled, it is necessary to explore the potential relationship between sample features and cluster prototypes. In particular, the DP-Ring loss result in $||Wx||_2 \equiv R$, indicate that the $\mathcal{L}_2$-norm of the model decision result must be constant. Noted that $||Wx||_2 = \sqrt{\sum_{i=1}^{m}(w_i^T x)^2} = R \implies \sum_{i=1}^{m}(w_i^T x)^2 = R^2 \geq 0$, so the dot products of every feature $x$ and all cluster prototypes are non-negative. Meanwhile, the cross entropy loss will try to narrow the angle between the feature and the cluster prototype [30]. With the joint optimization of DP-Ring loss and cross entropy loss, the network will promote $w_i^T x \to R$ and $w_j^T x \to 0$, where $w_i$ represents the correct cluster prototype and $i \neq j$. In other words, the DP-Ring loss constraints features to a point at their correct cluster prototype direction, and leads to different clustering prototypes orthogonal to each other (see Fig. 4b). This constraint will result in the source and target domains mix at several points, resulting in a better classification performance.

### 3.4. Pseudo-Label

The lack of category information in the target domain will cause the feature subspace to be ineffective, resulting in poor model performance. To overcome this problem, we can apply the Pseudo-Label [14] to produce pseudo-labels to unlabeled target data. The Pseudo-Label enable the model to learn discriminative class boundaries on the target domain, therefore improve the generalization performance of the network in the unlabeled domain.

In detail, we can initialize a threshold for the classification score before each iteration, and add the pseudo-label to all unlabeled data which classification score higher than threshold. To avoid the mis-classification situation, all target domain data will be tested after each iteration. And the previous pseudo-label data will be overwritten by the same and new ones.

### 3.5. Domain Adaptable Normalization (DANorm)

The proposed Domain Adaptable Normalization (DANorm) method is combined from feature $\mathcal{L}_2$ normalization, the DP-Ring loss and the Pseudo-Label. The total loss is formulated as Eq. (2).

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda\mathcal{L}_{DP-Ring} \tag{2}$$

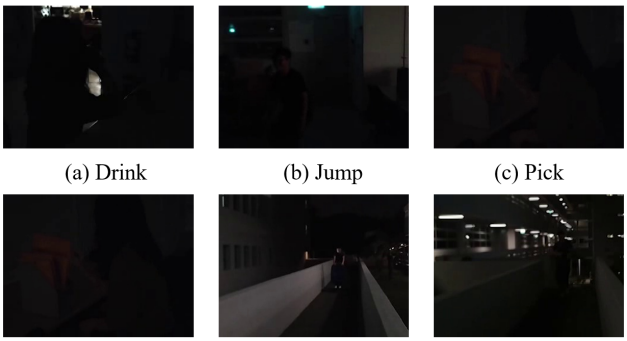
where $\mathcal{L}_{CE}$ means the cross entropy loss and $\lambda$ is a hyper-parameter.

Under the two constraints, the pseudo-label method will largely avoid the sample classified as the wrong label. Specifically, For the given source domain and target domain, in each iteration, the DANorm can be divided into two steps: (i) minimize the cross entropy loss and the DP-Ring loss on $N$ sample from training set, and (ii) add target domain data with high classification scores to the training set. All labeled data in the source domain is added to the training set at the beginning, and the previous pseudo-label data will be overwritten by the same and new pseudo-label data.

## 4. Experiment

### 4.1. Datasets

In this section, the ARID [34] dataset is used as unlabeled target domain, and HMDB51 [13], UCF101 [26], Kinetics-600 [12], and Moments in Time [19] datasets is used as labeled source domain to evaluate our method. All videos are divided into 3 splits and 11 categories: drink,



(a) Drink (b) Jump (c) Pick

(d) Pour (e) Push (f) Run

Figure 5. Sample frames from ARID dataset.

jump, pick, pour, push, run, sit, stand, turn, walk, and wave. Specifically, all split settings are displayed in Tab. 1.

| Split | 1 | 2 | 3 |
|---|---|---|---|
| Labeled Sample | 2625 | | |
| Unlabeled Sample | 2289 | 2253 | 2260 |
| Test Sample | 824 | 860 | 853 |

Table 1. Split settings in the experiment. Labeled source domains are the same and target domains are different of each split.

### 4.2. Implementation

We process the video into a series of frames of size $3 \times 48 \times 112 \times 112$, which are normalized and regularized sequentially. For the training set sample, we randomly drop the frame from sequence and finally normalize the sample into 48 frames.

The backbone of experiment network is the R(2+1)d-34 [27] that pretrained in IG65M [6] dataset and the classifier is a linear layer. Between them, the experiment network uses $\mathcal{L}_2$ normalization to constraint the feature subspace. The training algorithm is the proposed DANorm method, where letting $N = 400$, $\lambda = 1$ and the classification score threshold is $0.8$. Our model is optimized by AdamW [18] optimizer, letting learning rate be $1 \times 10^{-5}$. To improve the model generality, a parameter $\alpha = 1 \times 10^{-4}$ is used to weight-decay.

Specifically, we add all labeled data in the source domain to the training set at the beginning. Then in each iteration, $N$ sample will be taken in 100 batches for network to train, and the training losses is described as Eq. (2). After training step in iteration, we randomly sample $N$ sample in target domain for estimation, and add the high classification score data as the pseudo-label sample to the training set. All the previous pseudo-label data will be overwritten by the same and new pseudo-label data. The network will be trained with 25 iterations, and all experiments will be performed in each split independently.

### 4.3. Results

| Method | Top-1(%) |
|---|---|
| Baseline [27] | 75.85±2.71 |
| Pseudo-Label [14] | 72.50±0.58 |
| Zero-DCE + BERT [9] | 78.02 ±1.61 |
| **DANorm(ours)** | **80.73±1.45** |

Table 2. The Top-1 classification results of related methods and DANorm. All methods except Zero-DCE + BERT use the R(2+1)D-34 as backbone and linear layer as classifier.

Tab. 2 presents comparisons to other method, using IG65M-pretrained R(2+1)d as backbone. Except of the Zero-DCE + BERT, other methods apply single linear layer as their classifier. Note that Baseline represents not any semi-supervised method is applied during training.

Compared to previous methods, the DANorm achieves significant performance improvements: +4.88% for Baseline, +8.23% for the Pseduo-Label and +2.71% for the R(2+1)d + BERT architecture with Zero-DCE [7] enhancement. Besides, note that the performance for Pseudo-Label is lower than the Baseline, which indicate that the misclassification presented in Sec. 3.1 does affect the performance of semi-supervised methods, and the DANorm can ease this well.

### 4.4. Ablation study

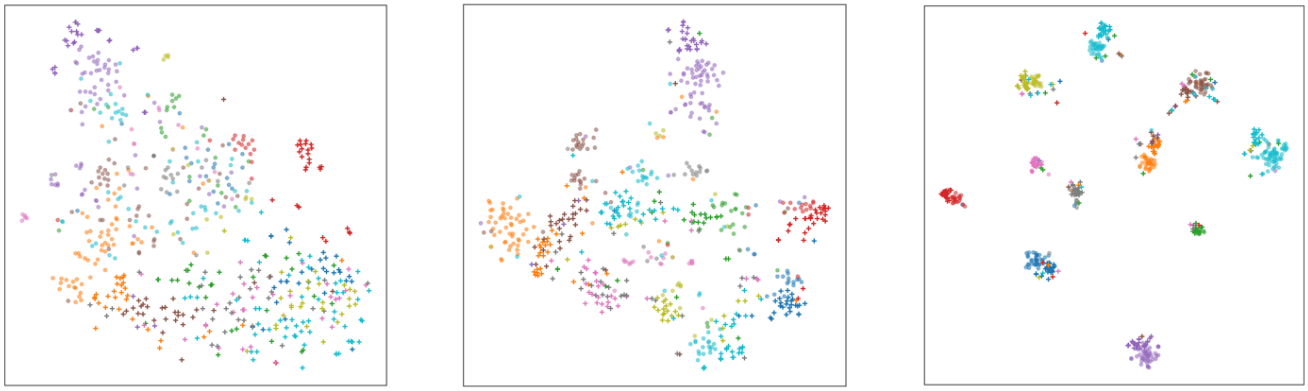
In this section, we will verify the effectiveness of each part in our method. Specifically, the feature normalization, the DP-Ring loss and the Pseudo-Label are tested successively.

Firstly, the $\mathcal{L}_2$ feature normalization method is evaluated for its effectiveness. The results are reported in Tab. 3. Under $\mathcal{L}_2$ regularization, the network average performance and stability both better than which without normalization. And the average 1.70% improvement strongly proves the effectiveness of the normalization method in SS-ARID task.

| Method | Top-1(%) |
|---|---|
| **With $\mathcal{L}_2$ Normlization** | **77.55±3.71** |
| Without $\mathcal{L}_2$ Normlization | 75.85±2.71 |

Table 3. Evaluation of the DANorm method with/without $\mathcal{L}_2$ normalization.

Next, the DP-Ring loss is tested for the performance of our network, and the results can be seen in Tab. 4. Compared with the other constraint losses, the DP-Ring Loss shows better average performance and smaller performance swings. This performance difference may be due to the similarity of Ring Loss to $\mathcal{L}_2$ normalization. And other losses



(a) Iteration 0 (b) Iteration 5 (c) Iteration 45

Figure 6. t-SNE [29] visualization results of the feature subspace of different iterations. The solid dots represent the labeled source domain feature, and the crosses represent the features of the unlabeled target domain. With the iteration increase, the distribution gap between the two domains decreases, and the distribution range of each category is gradually reduced to a point, while the distance of different category distribution increases.

do not have the ability to constrain features to a point in the feature subspace.

| Method | Top-1(%) |
|---|---|
| None | 77.55±3.71 |
| Triplet Loss [25] | 78.29±1.84 |
| Ring Loss [36] | 73.84±3.79 |
| **DP-Ring Loss (ours)** | **80.73±1.45** |

Table 4. Evaluation of the subspace constraint loss. None means there is not any subspace constraint loss during training.

In addition, we also test the effect of the Pseudo-Label, and the results are shown in Tab. 5, where without Pseudo-Label means there only $\mathcal{L}_2$ normalization and the DP-Ring Loss. Overall, the DANorm brings a 8.23% and 1.35% average improvement on only Pseudo-Label and without Pseudo-Label, respectively. The results point to the effectiveness of using the Pseudo-Label, $\mathcal{L}_2$ normalization and the DP-Ring Loss to train the network.

| Method | Top-1(%) |
|---|---|
| Only Pseudo-Label | 72.50±0.58 |
| Without Pseudo-Label | 79.38±1.45 |
| **DANorm (ours)** | **80.73±1.45** |

Table 5. Evaluation of the DANorm with/without Pseudo-Label. Without Pseudo-Label indicates that only $\mathcal{L}_2$ normalization and the DP-Ring Loss are applied on the network.

Further, we used t-SNE [29] to visualize the feature subspace at different iterations, and the results are shown in Fig. 6. The results from Fig. 6a to Fig. 6c show that feature normalization can align different domain features into one distribution, while angle constraint can force them distribute at several points. The transformation from Fig. 6a to Fig. 6c conforms to the theoretical analysis of feature normalization and angle constraint, displaying the effectiveness of the proposed DANorm method.

## 5. Conclusion

In this paper, to achieve better alignment across different video domains, a novel Domain Adaptable Normalization (DANorm) method is proposed for semi-supervised action recognition in the dark. The proposed DANorm method consists of feature normalization, angle constraint and the Pseudo-Label, which achieves the alignment of the two domains when the target domain is unlabeled. Besides, we analyzed how to achieve better alignment by constraining features under the unsupervised domain adaptation task. We found that feature normalization can align the low $\mathcal{L}_2$-norm feature and the high $\mathcal{L}_2$-norm ones, while angle constraint can mix different domains at several points. Experimental results show that our model can effectively improve the performance for semi-supervised action recognition in the dark.

## References


[1] Alphonse Chapanis. Interactive human communication. *Scientific American*, 232(3):36–46, 1975. 1

[2] 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/CVF International Conference on Computer Vision*, pages 6321–6330, 2019. 1, 2

[3] Rui Chen, Jiajun Chen, Zixi Liang, Huaien Gao, and Shan Lin. Darklight networks for action recognition in the dark.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 846–852, 2021. 1

[4] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *European Conference on Computer Vision*, pages 678–695. Springer, 2020. 1, 2

[5] Robert T Collins, Alan J Lipton, Takeo Kanade, Hironobu Fujiyoshi, David Duggins, Yanghai Tsin, David Tolliver, Nobuyoshi Enomoto, Osamu Hasegawa, Peter Burt, et al. A system for video surveillance and monitoring. *VSAM final report*, 2000(1-68):1, 2000. 1

[6] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12046–12055, 2019. 5

[7] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1780–1789, 2020. 1, 2, 5

[8] Niels Haering, Péter L Venetianer, and Alan Lipton. The evolution of video surveillance: an overview. *Machine Vision and Applications*, 19(5):279–290, 2008. 1

[9] Sanchit Hira, Ritwik Das, Abhinav Modi, and Daniil Pakhomov. Delta sampling r-bert for limited data and low-light action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 853–862, 2021. 1, 2, 5

[10] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021. 1

[11] Victor Kaptelinin. Activity theory: Implications for human-computer interaction. *Context and consciousness: Activity theory and human-computer interaction*, 1:103–116, 1996. 1

[12] 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. 4

[13] 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. 4

[14] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 4, 5

[15] Zixi Liang, Ming Yin, Junli Gao, Yicheng He, and Weitian Huang. View knowledge transfer network for multi-view action recognition. *Image and Vision Computing*, page 104357, 2021. 1, 2

[16] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, volume 2, page 7, 2016. 2

[17] Yunyu Liu, Lichen Wang, Yue Bai, Can Qin, Zhengming Ding, and Yun Fu. Generative view-correlation adaptation for semi-supervised multi-view learning. In *Proceedings of ECCV*, pages 318–334. Springer, 2020. 1, 2

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[19] 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. 4

[20] Feiping Nie, Lai Tian, Rong Wang, and Xuelong Li. Multiview semi-supervised learning model for image classification. *IEEE Transactions on Knowledge and Data Engineering*, 32(12):2389–2400, 2019. 1

[21] Alessandro Ortis, Giovanni M Farinella, Valeria D'Amico, Luca Addesso, Giovanni Torrisi, and Sebastiano Battiato. Organizing egocentric videos of daily living activities. *Pattern Recognition*, 72:207–218, 2017. 1

[22] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11815–11822, 2020. 1

[23] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017. 3

[24] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8050–8058, 2019. 2

[25] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2, 6

[26] 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. 4

[27] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 5

[28] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 2

[29] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6

[30] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017. 2, 4

[31] Lichen Wang, Zhengming Ding, Zhiqiang Tao, Yunyu Liu, and Yun Fu. Generative multi-view human action recognition. In *Proceedings of ICCV*, pages 6212–6221, 2019. 1, 2

[32] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6849–6857, 2019. 1

[33] Yuecong Xu, Jianfei Yang, Haozhi Cao, Zhenghua Chen, Qi Li, and Kezhi Mao. Partial video domain adaptation with partial adversarial temporal attentive network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9332–9341, 2021. 1

[34] 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. 1, 2, 3, 4

[35] Jianfei Yang, Han Zou, Hao Jiang, and Lihua Xie. Carefi: Sedentary behavior monitoring system via commodity wifi infrastructures. *IEEE Transactions on Vehicular Technology*, 67(8):7620–7629, 2018. 1

[36] Yutong Zheng, Dipan K Pal, and Marios Savvides. Ring loss: Convex feature normalization for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5089–5097, 2018. 2, 3, 4, 6