Z-Domain Entropy Adaptable Flex for Semi-supervised Action Recognition in the Dark

Zhi Chen* 1142154569@qq.com
Zijun Fan* 1395869798@qq.com
Yongjie Li 1183474278@qq.com
Huaien Gao 1496609@qq.com

Shan Lin† alice333.happy@163.com

Guangzhou Xi Ma Information Technology Company
101 Waihuan Xi Road, Da Xue Cheng, Guangzhou, Guangdong, China 510006

Abstract

The subtask of Human Action Recognition (AR) in the dark is gaining a lot of traction nowadays, which takes a significant place in the field of computer vision. The implementation of its application includes self-driving at night, human-pose estimation, night surveillance, etc. Currently, solutions such as DLN for AR have emerged. However, due to the poor accuracy even when leveraging on large amounts of datasets and complex architectures, the development of AR in the dark has been slow to progress. In this paper, we propose a novel and straightforward method: Z-Domain Entropy Adaptable Flex. This constructs a neural network architecture R(2+1)D, including (i) a self-attention mechanism, which combines and extracts corresponding and complementary features from the dual pathways; (ii) Zero-DCE low light image enhancement, which improves enhanced quality; and (iii) FlexMatch method, which can generates the pseudo-labels flexibly. With the help of pseudo-labels from FlexMatch, our proposed Z-DEAF method facilitates the process of gaining desired classification boundaries. This works by repeating Expanding Entropy and Shrinking Entropy. It aims to solve the problem of unclear classification boundaries between the categories. Our model obtains superior performance in experiments, and achieves state-of-the-art results on ARID.

1. Introduction

Action Recognition under dark conditions has gained widespread attention and more practical applications in real life such as self-driving in dim environment [4]. Nevertheless, there is still a lack of relevant and effective methods for Semi-Supervised Action Recognition in the dark (SS-ARID), because it requires efficiency and high accuracy. There are two main reasons for this: (i) inadequately labeled datasets in the dark which can be costly if it needs manual annotation; (ii) improper and unreasonable enhancement methods which could likely cause corruption of the datasets. Due to the reasons above, Semi-Supervised AR in the dark has gradually taken a place in solving the model degradation caused by the adverse visual condition [16, 27, 34].

As mentioned, the main question is how do we use a small amount of labeled data, plus a large amount of unlabeled data for our training. The strategy for combining two types of data becomes the key to solving the problem. Since the two types of data perform disharmony of features in domain, we can take the task as unsupervised. In unsupervised learning, the lack of category information leads to poor features, that is, poor extraction performance. Thanks to the R(2+1)D method [32] in the 3D-CNN model, we succeeded in extracting and getting abundant features to ensure the final classification.

Recently, the Semi-supervised Domain Adaptation via Minimax Entropy [8, 9, 23] (MME). It has been proposed to adversely optimize an adaptive few-shot model [2, 3]. The adaptation is achieved by alternately maximizing the conditional entropy of unlabeled target data with respect to the classifier and minimizing it with respect to the feature extractor.

Based on the MME [23] method, we proposed a novel and straightforward method, named Z-Domain Entropy Adaptable Flex for SS-ARID (Fig.1). It combines the concept of the K-means clustering architecture [21]. This model utilizes normalization to screen out the significant features from the feature extractor so that we can focus on

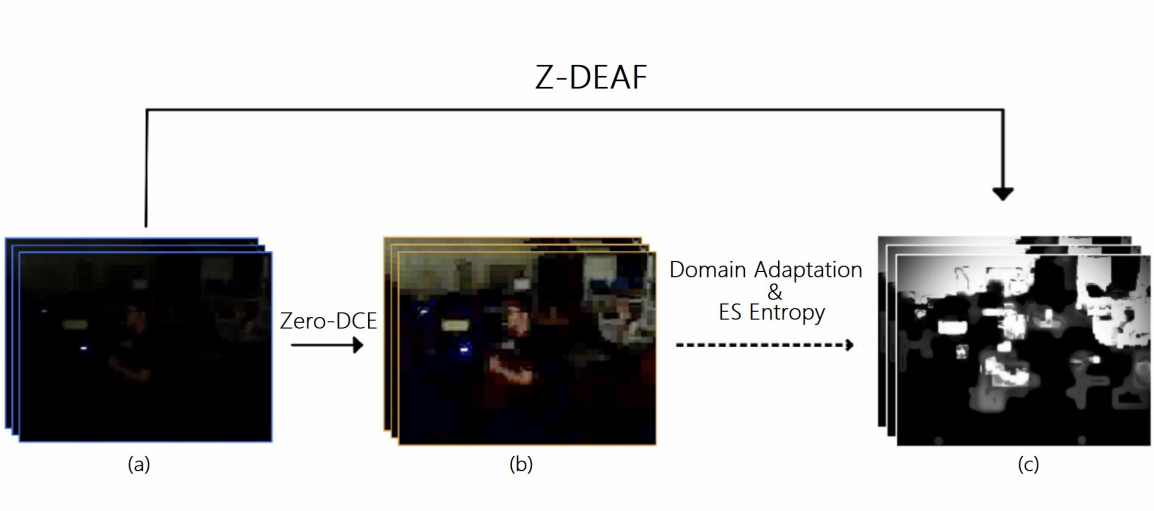

Figure 1. The architecture of Z-Domain Entropy Adaptable Flex method. The input is a sequence of dark frames (a). By applying Zero-DCE low light image enhancement method, we gain another input for the domain adaptation section (b). Then through multiple trainings and iterations that repeating the process of Expanding Entropy and Shrinking Entropy yields desired classification boundaries for the task (c).

enhancing the suitable feature information. Meanwhile, this can further reduce divergence of the data so that the subsequent entropy process can distinguish the boundaries between different classes.

To sum up, our main work here is to (i) extend the MME [23] method from general datasets to dark datasets; and (ii) solve the problem that MME [23] does not perform well under semi-supervised conditions. On one hand, we combine the FlexMatch [39] method with the traditional MME [23] method to generate pseudo-labels in order to improve the domain adaptable performance. On the other hand, we also use the Zero-DCE enhancement [11] method to replace traditional enhancement methods [12, 30, 38, 40] to better protect dark data from too much damage. From the results of our ablation experiments, we found that Z-DEAF has proved to improve the performance which has reflected in the accuracy of the classification.

2. Related Works

R(2+1)D Based 3D ConvNet Architecture 3D ConvNet [13] is developed from 2D ConvNet by rising dimension. 2D convolution applied on an image will output an image, 2D convolution applied on multiple images also results in an image. Hence, 2D ConvNets lose temporal information of the input signal right after every convolution operation. Only 3D convolution preserves the temporal information of the input signals resulting in an output volume.

Since the task is based on video action recognition which touch on temporal dimension, we believe that 3D ConvNet is well-suited for spatiotemporal feature extraction. In 3D convolution, filters are designed in a 3D fashion, where channels and temporal information are represented in different dimensions. Compared to 2D ConvNet, 3D ConvNet has the ability to model temporal information better owing to 3D convolution and 3D pooling operations. They are performed spatio-temporally while in 2D ConvNets they are done only spatially [31].

To cut down complexity of the network and gain a better accuracy on feature extraction, a ResNet version of 3D convolution, the R(2+1)D convolutional neural network is introduced in [32]. It is a network for action recognition that employs R(2+1)D convolutions in a ResNet inspired architecture. The use of these convolutions over regular 3D Convolutions reduces computational complexity, prevents overfitting, and introduces more non-linearities that allow for a better functional relationship to be modeled [31].

Domain Adaptation Concept Deep convolutional neural networks have significantly improved image classification accuracy with the help of large quantities of labeled training data, but often generalize poorly to new domains. Recent transfer learning method [1, 17, 29], in which domain adaptation (DA) methods [10, 18, 19, 24, 33, 35, 36] improve generalization on unlabeled target data by aligning distributions. And it has been applied to various applications such as image classification [22], semantic segmentation [26], and object detection [7, 25]. However, it fails to learn discriminative class boundaries on target domains. We show that in the Semi-Supervised Domain Adaptation (SSDA) setting where a few target labels are available, such

methods often do not improve performance relative to just training on labeled source and target examples and can even make it worse [10, 19, 24].

We propose a novel approach for SSDA that overcomes the limitations of previous methods and significantly improves the accuracy of deep classifiers on novel domains with only a few labels per class.

3. Method

Our method is based on K-means clustering algorithm [21] and inspired by the essence of concept of entropy. By optimizing standard cross-entropy loss for training feature extractor and classifier for classification, reduces the distribution gap while learning classification boundaries for the task. The classifier (top layer) predicts a K-way class probability vector by computing cosine similarity between K class-specific weight vectors and the output of a feature extractor (lower layers), followed by a softmax. Each class weight vector is defined as the "base point", that can be regarded as a representative point of that class. Through multiple trainings and iterations that repeating the process of Expanding Entropy and Shrinking Entropy yields desired classification boundaries for the task.

3.1. K-means clustering architecture

K-means algorithm is an unsupervised clustering algorithm. It is widely used due to good clustering effect and easy to implement. If the category of the data is not known before classification, like the unlabeled data in this task, we can use K-means to classify the data. The idea of k-means algorithm is intuitive. For a given sample set, it is divided into K clusters according to the distance between samples. Make the base points in the cluster as close together as possible. Meanwhile make the gap between clusters as large as possible. Through numerous iterations and training, it can effectively classify different types of data. If the gap between clusters is larger, the classification boundary between the categories will be clearer and the classification will be more accurate.

If expressed by mathematical expression, suppose the cluster (C) is divided into $C = \{C_1, C_2, \dots C_k\}$, the input is the sample set: $D = \{x_1, x_2, \dots x_m\}$, our goal is to minimize the squared error (E), which can be expressed as:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|_2^2 \tag{1}$$

Where μ_i is the mean vector of cluster C_i , also called the mass center, can be expressed as:

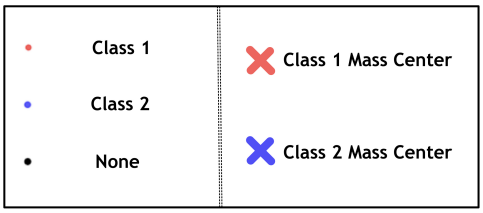

Figure 2. Notation Sample

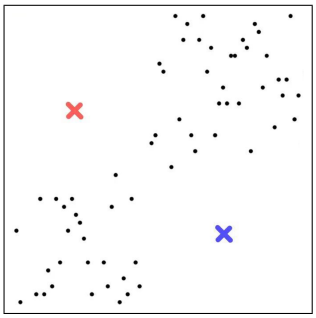

Figure 3. Initial Data Set

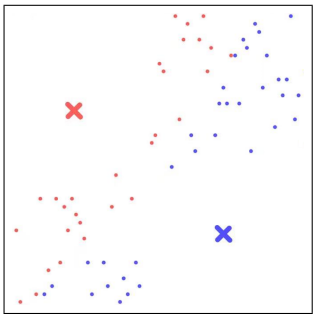

Figure 4. Initial Mass Center

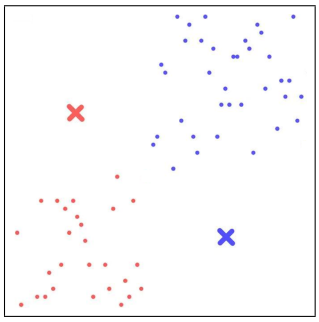

Figure 5. Update Category

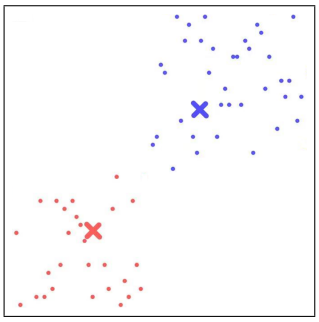

Figure 6. Update Mass Center

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{2}$$

Since it is not easy to directly calculate the minimum value of the above formula, we can consider using heuristic iterative method. The heuristic method adopted by K-means is very simple, which can be vividly described by the follow group of figures.

Fig.3 shows the initial data set, assuming k=2. In Fig.4, we randomly chose 2 mass centers correspond to the category, the red cross and blue cross in this picture. And then respectively measure the distance between each of the points in the sample and the mass centers. Then the category of the mass center with the smallest distance to the sample is used as the category of this sample, as shown in Fig.5. After calculating the distance between the samples and red cross and blue cross, we obtain the categories of all sample points after the first iteration. At this point, we re-

calculate 2 new mass centers of the points currently marked red and blue respectively. The positions of the new red cross and blue cross have changed. Then repeat the process that we did in Fig.5, that is, marking the category of all points as the nearest cross and finding the new mass center. Iterate until the result of this step is the same as that of the previous step. The two categories we end up with are shown in Fig.6.

3.2. Expanding Entropy

In semi-supervised domain adaptation, we are given source images and the corresponding labels in the source domain $\mathcal{D}_l = \{(x_i^l, y_i^l)\}_{i=1}^{m_l}$. We also given a number of unlabeled target images in the target domainz $\mathcal{D}_u = \{(x_i^u)\}_{i=1}^{m_u}$. By training the model on $\mathcal{D}_l$, we can evaluate the model on $\mathcal{D}_u$.

We apply a R(2+1)D method of 3D convolutional neural network and perform ℓ_2 normalization on the output of the network, which is the extracted feature. And then the normalized feature vector is used as an input to the classifier, which consists of weight vectors $W = [w_1, w_2, \ldots, w_K]$, where K represents the number of classes. $\frac{F(x)}{\|F(x)\|}$ serve as the input of classifier and outputs $\frac{1}{T}\frac{W^T F(x)}{\|F(x)\|}$. Then the output is read into a softmax layer to attain the probabilistic output $P \in R^n$, which can be expressed as: $P(x) = \mathcal{S}\left(\frac{1}{T}\frac{W^T F(x)}{\|F(x)\|}\right)$, where $\mathcal{S}$ indicates a softmax function. In order to classify examples accurately, the direction of a weight vector should be representative to the normalized features of the corresponding class. In this respect, the weight vectors can be regarded as the representative "base points" for each class, likewise the mass centers purposed in the K-mean clustering algorithm for each class.

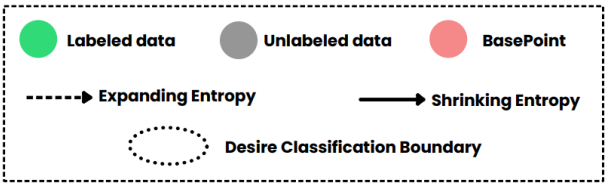

Figure 7. Graphic Sample

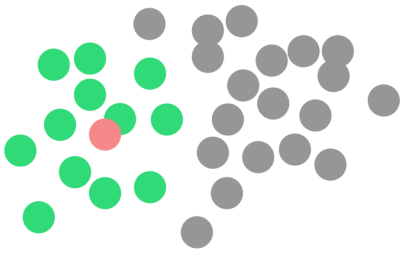

Figure 8. Established Base Point

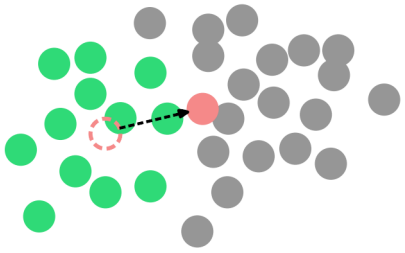

Figure 9. Expanding Entropy

The "base points" are parameterized by the weight vector of the last linear layer. As Fig.8 shows, the first "base point" will be near source distributions because source labels are dominant. Then, we propose to estimate the position of next "base point" by moving each w_i toward target features using unlabeled data in the target domain. By operating entropy expanding with respect to the "base point" established by the previous iteration, and implementing multiple iterations until we generate a relatively stable and invariant "base point" for each class. This step we call it Expanding Entropy, and it can be graphically described as "base point" moves from near source domain to target domain (Fig.9), that is, from order to disorder, which is the essence of entropy. To achieve this, we increase the entropy measured by the similarity between W and unlabeled target features. Entropy is calculated as follows,

$$H = -\mathbb{E}_{(x,y)\in\mathcal{D}_u} \sum_{i=1}^{K} p(y = i \mid x) \log_2 p(y = i \mid x) \quad (3)$$

Where K is the number of classes and $p(y = i \mid x)$ represents the probability of prediction to class i, namely i th dimension of $P(x) = \mathcal{S}\left(\frac{1}{T}\frac{W^T F(x)}{\|F(x)\|}\right)$. To have higher entropy, that is, to have uniform output probability, each w_i should be similar to all target features. Hence, expanding the entropy advance the model to generate the next "base point" for each class.

3.3. Shrinking Entropy

In order to obtain discernible and clear classification boundaries, we need to cluster unlabeled target features around the "base points". We propose to decrease the entropy on unlabeled target examples by the feature extractor.

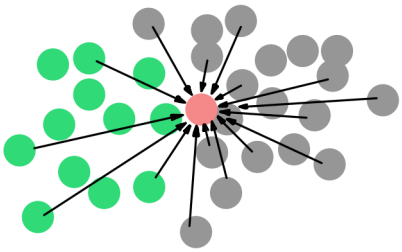

Figure 10. Shrinking Entropy

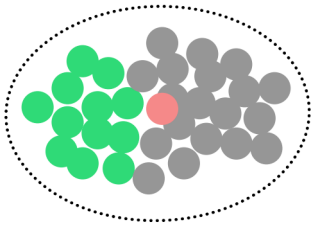

Figure 11. Desire Classification Boundary

As Fig.10 shows, this step is like the reverse of the Expanding Entropy, we call it Shrinking Entropy. We train feature extractor and classifier to classify labeled source correctly and perform a Shrinking Entropy process to cluster unlabeled features into the source domain, that is, aggregating unlabeled features from a divergent state into a more orderly state, with a closer distance to the source domain, and they should be assigned to one of the "base point" to decrease the entropy, resulting in the desired category classification. We use a standard cross-entropy loss $\mathcal{L}_{ce}$ to train feature extractor and classifier for classification:

$$\mathcal{L} = \mathbb{E}_{(x,y)\in\mathcal{D}_l} \mathcal{L}_{ce} \left[\mathcal{S}\left(\frac{1}{T} \frac{W^T F(x)}{\|F(x)\|} \right), y \right] \tag{4}$$

Through multiple trainings and iterations that repeating the process of Expanding Entropy and Shrinking Entropy yields desired classification boundaries for the task (Fig.11).

3.4. FlexMatch

However, due to the lack of sufficient labeled data for this task, the direct training effect of the above method is not ideal, which would affect the shift of "base point" and the final position it will get to. And it would be great if there goes an effective method that could increase the number of the labeled data, to be more specific, by this method, we can convert our trained unlabeled data into trained pseudo-labeled data and consider pseudo-labeled data as labeled data, thus we adopt the method named FlexMatch which is based on FixMatch improvement [39].

The proposed FlexMatch introduced a concept named Curriculum Pseudo Labeling (CPL) [6] and it has been improved that it can be easily adapted to some of the SSL algorithms and remarkably improve their performances [39]. To be more specific, the method includes two main ideas, (i) Lower the threshold value of the classes with low classification accuracy and give these classes more opportunities to be learnt for improving their value of Highest-Confident. (ii) Maintain threshold for classes that already rank high accurate to ensure high accuracy.

To this end, calculating evaluation accuracies for each class and use them to scale the threshold, as:

$$\mathcal{T}_t(c) = a_t(c) \cdot \tau \tag{5}$$

Where $\mathcal{T}_t\left(c\right)$ is the flexible threshold for class c at time step t and $a_t\left(c\right)$ is the corresponding evaluation accuracy. When the threshold is high, the number of samples whose predictions fall into this class and above the threshold can reflect the learning effect of a class. Namely, the class with fewer samples having their prediction confidence reach the threshold is considered to have a greater learning difficulty or a worse learning status, formulated as:

$$\mathcal{L}_{u,t} = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbb{1}\left(\max\left(q_b\right) > \mathcal{T}_t\left(argmax\left(q_b\right) H\left(\hat{q}_b, \mathcal{P}_m\left(y \mid \Omega\left(u_b\right)\right)\right)\right)\right) \tag{6}$$

Where $\sigma_t\left(c\right)$ reflects the learning effect of class c at time step t. $\mathcal{P}_{m,t}\left(y|u_n\right)$ is the model's prediction for unlabeled data u_n at time step t, and N is the total number of unlabeled data. When the unlabeled dataset is balanced, larger $\sigma_t\left(c\right)$ indicates a better estimated learning effect. By applying the following normalization to $\sigma_t\left(c\right)$ to make its range between 0 to 1, it can then be used to scale the fixed threshold τ:

$$\beta_t(c) = \frac{\sigma_t(c)}{\max_c \sigma_t} \tag{7}$$

$$\mathcal{T}_c(c) = \beta_t(c) \cdot \tau \tag{8}$$

The best-learned class has its $\beta_t\left(c\right)$ equal to 1, causing its flexible threshold equal to τ. As learning proceeds, the threshold of a well-learned class is raised higher to selectively pick up higher-quality samples. Eventually, when all classes have reached reliable accuracies, the thresholds will all approach τ. This new threshold is used for calculating the unsupervised loss in FlexMatch, which can be formulated as:

$$\mathcal{L}_{u,t} = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbb{1} \left(\max \left(q_b \; > \right. \right. \\ \left. \left. \mathcal{T}_t \left(\operatorname{argmax} \left(q_b \, H \left(\hat{q}_b, \mathcal{P}_m \left(y \mid \Omega \left(u_b \right) \right) \right) \right) \right) \right) \right) \tag{9}$$

Where $q_b = \mathcal{P}_m \left(y \mid \omega \left(u_b \right) \right)$. The flexible thresholds are updated at each iteration. Finally, we can formulate the loss in FlexMatch as the weighted combination $(by\ \lambda)$ of supervised and unsupervised loss:

$$\mathcal{L}_t = \mathcal{L}_s + \lambda \mathcal{L}_{u,t} \tag{10}$$

Where $\mathcal{L}_s$ is the supervised loss on labeled data:

$$\mathcal{L}_s = \frac{1}{B} \sum_{b=1}^{B} H \left(y_b, \mathcal{P}_m \left(y \mid \omega \left(u_b \right) \right) \right) \tag{11}$$

Practically, every time the prediction confidence of an unlabeled data u_n is above the fixed threshold τ, the data, and its predicted class are marked and will be used for calculating $\beta_t \left(c \right)$ at the next time step. So far, we have successfully converted our unlabeled data into trained pseudo-labeled data and consider pseudo-labeled data as labeled data.

4. Experiments

4.1. Experimental Details

In this section, we conduct experiments on ARID datasets which ranks the first benchmark datasets for Action Recognition in the dark [37]. To be more specific, we take the ARID datasets as unlabeled target domain, while we takes the HMDB51 [15], UCF101 [28], Kinetics-600 [5,14] and Moments in Time datasets [20] as labeled source domain. As a comparison, the former contains 3088 dark videos and the latter contains 2625 clear videos. All the videos are divided into 11 categories: drink, jump, pick, pour, push, run, sit, stand, turn, walk and wave. (The goal of dataset is to achieve satisfactory accuracy on a set of 330 dark videos.) The goal of our experiment is to improve the accuracy on a set of 330 dark videos. Meanwhile, our strategy is using the FixMatch method to generate pseudo-labels by a certain value of step, in order to change the distribution of data over a domain. As an improvement strategy, in changing the way of generating the pseudo-labels, we take FlexMatch method for better adaptable domain, and create the pseudo-labels with the same way and the same value of step. We make comparison for top-1 results of various situations through ablation experiments, and take 0.6 for threshold as the optimal scheme under FlexMarch strategy. It is well noted that the unlabeled samples from our target source do not contain our test dataset.

Method	Convert Ratio	Top-1
FixMatch	30%	41.21%
FlexMatch	30%	45.75%
FlexMatch	60%	47.27%
FixMatch	**90%**	**46.67%**
FlexMatch	**90%**	**49.39%**

Table 1. The Convert Ratio and Top-1 accuracy results of a few competitive models and ours

4.2 Results and Comparisons

As a result of different scales between videos, we first zoom out the video in the fomat of (-1,256). Extract 32 frames respectively from each video. Then we clip and normalize these frames uniformly. After preprocessing, we process the video into a series of frames of size $3 \times 32 \times 112 \times 112$. As for the feature extractor, we adopt R(2+1)D-34 which has pretrained on IG65M to accelerate our training. In the stage of feature extraction, we would obtain the output feature of size $512 \times 4 \times 7 \times 7$. Noted that we classify the feature into $unlabel\_strong$ and $unlabel\_weak$, and we enhance the $unlabel\_strong$ by the method of Zero-DCE while we do nothing to $unlabel\_weak$.

The backbone of classification network is based on our proposed Z-DEAF method for its training, which takes used of linear network as our classifier. From Table.1, it is worth noting that our ablation experiment is to adjust the number of the pseudo-labels by the setting of the threshold value in the method of FlexMatch, referring to the accuracy of about 50% when we set the threshold to 0.9 with FixMatch.

Our model is optimized by AdamW optimizer, letting learning rate be 1×10^{-3}. The number of training epochs are 300 to make sure well-train as possible. To improve the model generality, a parameter $\alpha = 1 \times 10^{-3}$ is set in weight-decay. For efficiency, the batch size of labeled and unlabeled is 64 and 75 for each considering the actual situation.

5. Conclude

In conclusion, we propose a new method for Semi-Supervised Action Recognition in the Dark, named Z-DEAF. It aims to solve the problem of lacking labeled datasets in dimmed environments, and improve adaptable problem of domain adaptation. In order to achieve these goals, it well takes advantage of semi-supervised learning by integrating methods such as Zero-DCE enhancement, MME method and FlexMatch method. The method will result in avoiding any unnecessary extra work, as well as gain better accuracy on the ARID dataset. Especially, we also add self-attention mechanism into the feature extractor

in order to percept the required information of features. We have conducted ablation experiments on the ARID dataset, and the experiments indicate our proposed method is feasible and powerful.

References

[1] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *arXiv preprint arXiv:1812.11941*, 2018. 2

[2] Shuang Ao, Xiang Li, and Charles Ling. Fast generalized distillation for semi-supervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 1

[3] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006. 1

[4] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016. 1

[5] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 6

[6] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. *arXiv preprint arXiv:2001.06001*, 2020. 5

[7] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 2

[8] Abhimanyu Dubey, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. Maximum-entropy fine grained classification. *Advances in neural information processing systems*, 31, 2018. 1

[9] Ayse Erkan and Yasemin Altun. Semi-supervised learning via generalized maximum entropy. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 209–216. JMLR Workshop and Conference Proceedings, 2010. 1

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 2, 3

[11] 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. 2

[12] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on image processing*, 26(2):982–993, 2016. 2

[13] M Kalfaoglu, Sinan Kalkan, and A Aydin Alatan. Late temporal modeling in 3d cnn architectures with bert for action recognition. In *European Conference on Computer Vision*, pages 731–747. Springer, 2020. 2

[14] 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. 6

[15] 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. 6

[16] Jingyuan Li and Eli Shlizerman. Sparse semi-supervised action recognition with active learning. *arXiv preprint arXiv:2012.01740*, 2020. 1

[17] Zixi Liang, Ming Yin, Junli Gao, Yicheng He, and Weitian Huang. View knowledge transfer network for multi-view action recognition. *Image and Vision Computing*, 118:104357, 2022. 2

[18] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 2

[19] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018. 2, 3

[20] 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. 6

[21] Michal Moshkovitz, Sanjoy Dasgupta, Cyrus Rashtchian, and Nave Frost. Explainable k-means and k-medians clustering. In *International conference on*

machine learning, pages 7055–7065. PMLR, 2020. 1, 3

[22] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 2

[23] 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. 1, 2

[24] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. *arXiv preprint arXiv:1711.01575*, 2017. 2, 3

[25] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019. 2

[26] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3752–3761, 2018. 2

[27] Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, and Abir Das. Semi-supervised action recognition with temporal contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10389–10399, 2021. 1

[28] 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. 6

[29] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 32–42, 2021. 2

[30] Panos E Trahanias and Anastasios N Venetsanopoulos. Color image enhancement through 3-d histogram equalization. In *11th IAPR International Conference on Pattern Recognition. Vol. III. Conference C: Image, Speech and Signal Analysis,*, volume 1, pages 545–548. IEEE Computer Society, 1992. 2

[31] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

[32] 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. 1, 2

[33] 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

[34] Junfei Xiao, Longlong Jing, Lin Zhang, Ju He, Qi She, Zongwei Zhou, Alan Yuille, and Yingwei Li. Learning from temporal gradient for semi-supervised action recognition. *arXiv preprint arXiv:2111.13241*, 2021. 1

[35] 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. 2

[36] Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. Aligning correlation information for domain adaptation in action recognition. *arXiv preprint arXiv:2107.04932*, 2021. 2

[37] 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. 6

[38] Zhenqiang Ying, Ge Li, Yurui Ren, Ronggang Wang, and Wenmin Wang. A new image contrast enhancement algorithm using exposure fusion framework. In *International Conference on Computer Analysis of Images and Patterns*, pages 36–46. Springer, 2017. 2

[39] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 5

[40] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1632–1640, 2019. 2