

Zero-Augmentation Pseudo Entropy for Video Unsupervised Domain Adaptation

Zhi Chen

Guangdong University of Technology

Zijun Fan

Guangdong University of Technology

Yongjie Li

Guangdong University of Technology

Shan Lin (✉ alice@gdut.edu.cn)

Guangdong University of Technology

Research Article

Keywords: Video Unsupervised Domain Adaptation, Action Recognition, Image Enhancement, Dark Video

Posted Date: June 30th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1758449/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Zero-Augmentation Pseudo Entropy for Video Unsupervised Domain Adaptation

Zhi Chen^{1†}, Zijun Fan^{1†}, Yongjie Li¹ and Shan Lin^{1*}

¹Guangdong University of Technology, Guangzhou, 510006, China.

*Corresponding author(s). E-mail(s): alice@gdut.edu.cn;

Contributing authors: i_justcz@163.com; Faan1008@163.com; Li_yongjie1183@163.com;

[†]These authors contributed equally to this work.

Abstract

This work focuses on the problem of video unsupervised domain adaptation (VUDA) for the task of human action recognition. Although there exist some methods for VUDA, however, most current methods exhibit ambient light bias. Transferring models from light environment to the dark will lead to a decline in performance due to inevitable domain shift. Unsupervised domain adaptation (UDA) methods are able to alleviate the performance decline by aligning source and target domains utilizing methods such as adversarial training. However, these methods do not enhance images for the dark environment, leading to the loss of information and weak contribution to model performance in the dark environment. In this work, we enhance images by specific approach as an auxiliary task for VUDA in addition to unsupervised adversarial alignment as our proposed method, Zero-Augmentation Pseudo Entropy (Z-APE). We test our method on the HMDB-ARID datasets and empirically show that our VUDA method improves the performance by average **6.3%** compared to other typical VUDA baseline models.

Keywords: Video Unsupervised Domain Adaptation, Action Recognition, Image Enhancement, Dark Video

1 Introduction

Recent vision based action recognition methods have developed and achieved high performances. Many real-world applications have been made possible, such as self-driving [1, 2], real-time face detection [3], human pose estimation [4, 5] and fake license plate recognition [6]. The two main ingredients for successful real-world applications of the various models are (i) the emergence of large datasets on action recognition such as HMDB51 [7], UCF101 [8], Kinetics-600 [9, 10]; and (ii) the establishment of transfer learning methods [11–13] such as unsupervised domain

adaptation based on adversarial learning framework [14–20].

However, the poor quality of video data in real-world scenarios has become a bottleneck constrain in the application of action recognition models in real-world applications. Although unsupervised domain adaptation can alleviate the performance decline caused by the transferring from labeled data to unlabeled data [16], the unlabeled data are not well utilized, and it will hinder the development as well. Further, there is still a lack of relevant research on domain adaptation in action recognition in the dark.

To solve the problem caused by domain shift, the existing methods’ training strategies of scaling data in the domain may mis-classify the noise points and do harm to the process of iteration. For example, the action of a human hip resting in the mid-air, which is falling when humans sit down and rising when humans stand up. In brief, it is an action between “sit down” and “stand up”, and defined as “hesitation”. Such action may join the process of training, resulting in an error classification. Similar problems have been found in the process of transfer learning, but few studies have explicitly solved the problem of classifying fuzzy actions.

The MME domain adaptation method [21–23] is applied to the task of video domain adaptation, in addition to training the feature extractor to obtain more discriminative features, the classifier is trained to make more accurate classification. Performing domain adaptation through the weight deviation between source and target domain. But the drawback of this method is that the proposed weight matrix contains all the unlabeled information of target domain, which degrade method’s performance at a certain extent.

Therefore, we proposed a new method of domain adaptation for videos, named Zero-Augmentation Pseudo Entropy (Z-APE). We work on a setting of unsupervised domain adaptation, domain target is labeled data in light environment, while target domain is only unlabeled data in dark. Due to the process being self-supervised, it does not need any annotations, while our data in target domain is unlabeled.

Based on the above insights, we proposed “Expanding & Shrinking Entropy (ESE)” and “Dynamic Pseudo Label (DPL)” with two novel components. First, we propose ESE to diminishes the distribution discrepancy of source and target domain. Then using DPL to correct the disorganized and reorganized classification process helped improve our model’s error-tolerance rate in action classification and discrimination between categories. The model learns two similar human actions by proposing a shuffling and regrouping task coupled with dynamic generation of pseudo labels. The attention mechanism led to the suppression of temporal background clips, which helps us only focus on aligning the important clips. We also use auxiliary light enhancement approach

Zero-DCA to optimize the images without losing important information.

We provide an extensive empirical evaluation to demonstrate the benefits of the proposed method on a challenging VUDA benchmark setting. We also give qualitative results to highlight the benefits of our approach. To sum up, our contributions are as follows:

- We propose a VUDA approach named Z-APE, which enable more flexible classification by encouraging models to acquire more discriminative features, adapting the model to dark environment target domain without labels.
- We propose ESE, Zero-DCA and DPL components to achieve the representation of important information in the target domain under adverse conditions.
- We test our method on the HMDB-ARID datasets [24], and the results show that our method is superior to other domain adaptation strategies, obtaining the latest result of HMDB-ARID dataset benchmark.

2 Related Works

Self-supervised action recognition.

Video-based self-supervised methods work with spatial context and have been applied to prediction work such as self-supervised domain sequence prediction [25], self-supervised RGB/flow correspondence prediction [26]. The sub-task includes GAN, whose adversarial loss function is also used as a framework in unsupervised domain adaptation action recognition tasks [27]. Combined with the loss of adversarial domain adaptation, the domain-related background is suppressed, and the task-related information is enhanced in the representation, resulting in better domain adaptation performance.

With the rise of deep convolutional neural networks (CNNs), recent video classification work mainly uses CNNs to extract accurate video features and representations, and action recognition has made great progress. CNN was expanded to 3D convolution [28]. These works mainly proposed and adopted dual-flow structure, which including feature extraction of two paths, and the late fusion of dual-flow method combining RGB and Flow [26]. Although this kind of work has made great progress in the development of action

recognition, considering the balance between performance and computational cost, 3D convolution and other processing for complete video clips is still a cost consuming work. Instead, in our work, we focus on the actions of the human torso rather than the rest of the content of the fragment. We add other modules such as attention modules to focus on the valid information and complement our approach.

Video unsupervised domain adaptation.

Video unsupervised domain adaptation (VUDA) commonly uses an adversarial learning framework [15, 18, 19]. Although domain adaptation (especially unsupervised domain adaptation) and semi-supervised domain adaptation (SSDA) seem to be two distinct tasks, semi-supervised task can be regarded as a special case of unsupervised domain adaptation. SSDA works well on domain adaptation tasks, achieving (or exceeding) high accuracy results [22, 29–31].

In recent years, there has been a growing interest in domain adaptation, which aims to extract cross-domain shared knowledge and improve the mobility of models. Recent transfer learning methods [11–13], in which the main adaptation method [5, 14] improves generalization of unlabeled target data by aligning distribution. It has been applied in various applications, such as image classification [32], semantic segmentation [25, 33], and object detection [4, 5]. However, it cannot learn discriminant features in the target domain. We focus on video unsupervised domain adaptation (VUDA), where there is no annotated data in the target domain. We find that unsupervised domain adaptation settings generally do not improve performance, and may even worsen, compared to training with source and target data with tags. Although semi-supervised domain adaptation (SSDA) has made progress in image-based tasks, there are few studies focusing on video-based semi-supervised domain adaptation (VSSDA), such as action recognition and action segmentation. Most of the works cannot effectively use the small amount of labeled data in the target domain while they treat SSDA as UDA for adversarial learning. Therefore, we propose a new VUDA method, which overcomes

the limitations of previous methods and significantly improves the accuracy of depth classifier in different domains to better align video features.

Pseudo label.

Deep learning is the study of machine learning based on data and labels. Tasks that lack labels are usually defined as semi-supervised or unlabeled tasks. In recent years of work, due to most models rely on the fine-tune between source data and target data, models cannot be applied to different domains or scenarios without adequate labels or annotations. Training for unlabeled data is still rare. In a closely related work, Zhang et al. [34] focused on the mass production of pseudo-labels and thus derived methods such as FlexMatch [34] to generate pseudo-label for unlabeled data in tests. However, FlexMatch method is sub-optimal. FlexMatch is only a weak data enhancement method, which cannot provide accurate pseudo-labels and may produce noise correspondingly, and even lead to the degradation of model performance during domain adaptation. Therefore, we recommend aligning relevant information in SSDA tasks for better video feature alignment. In this work, we proposed DPL approach, which is different from the FlexMatch approach, to dynamically generate pseudo-labels and correct misinformation in SSDA during multiple alignments.

3 Method

3.1 Zero-DCA

In the Zero-DCA enhancement method, the method is a framework composed of augmentation curve, DCA-Net, and zero-reference loss function. The framework obtains the enhanced images by repeatedly applying all pixels of the RGB channel input from the curve map. Therefore, we design a curve that can automatically map the image to the enhanced version, and the parameters of the augmentation curve are completely dependent on the input image. Inspired by Zero-DCE [35], we hope that curve can be applied to a larger dynamic range, and the enhanced range tends to enhance or weaken local areas by global mapping. In addition to that, we hope the curve to be repeatedly used to deal with more challenging low light conditions.

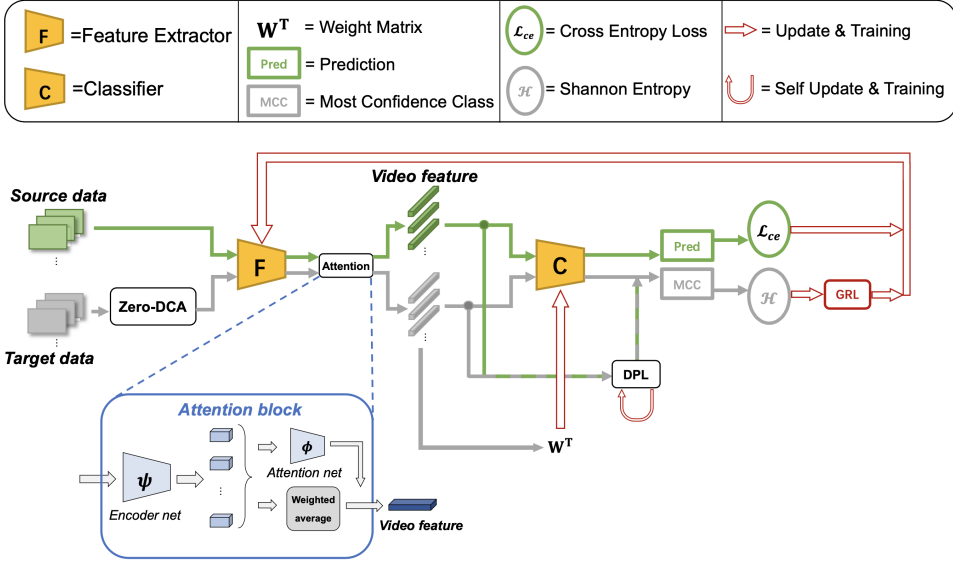


Fig. 1 The architecture of the Z-APE. At first the target domain videos are enhanced by the Zero-DCA component. Then video features are generated by a shared feature extractor for both source and target domain videos. Video-level features are obtained through an attention network. Prediction (Pred) and most confidence class (MCC) are the output of the features of two domains through the classifier respectively. A cross entropy loss and a Shannon entropy are applied to feature extractor for updating and training to reduce domain discrepancy, while a weight matrix from target data is used to train classifier. Besides, the DPL component is proposed to improve data utilization.

Therefore, we designed the curve as follows:

$$E_n(I(x), \gamma) = E_{n-1}(I(x), \gamma) + A_n(x)E_{n-1}(I(x), \gamma)(1 - E_{n-1}(I(x), \gamma)) \quad (1)$$

where x represents pixel coordinates, $E_n(I(x), \gamma)$ is the enhanced version of the given input $I(x)$, and $\gamma \in [-1, 1]$ is the trainable curve parameter, which adjusts the size of the enhanced curve and controls the exposure level. n is the times of iterations, which changes the curvature of the curve. When $n = 1$, the input is the original image. $A_n(x)$ is a parameter mapping of the same size as the given image.

In order to learn the mapping between the input image and its parameter of the best fitting curve, we use the DEC-Net from Zero-DCE method [35]. The input of DCE-Net is a low-light image, and the output is a set of pixel-level curve parameter maps corresponding to higher-order curves. In order to enable zero-reference learning in DCE-Net, we refer to four sets of training loss functions in Zero-DCE [35] method. They are spatial consistency loss, exposure control loss, color constancy loss and illumination smoothness loss. Here is a brief introduction to the above loss

functions:

$$\mathcal{L}_{spa} = \frac{1}{K} \sum_{i=1}^K \sum_{j \in \Omega(i)} (\|Y_i - Y_j\| - \|I_i - I_j\|)^2 \quad (2)$$

where K is the number of local regions, and $\Omega(i)$ is the four adjacent regions centered at region i . Y and I are respectively the average intensity value of the enhanced version and the local area in the input image. $\mathcal{L}_{spa}$ preserves the difference between the adjacent areas of the input image and the enhanced image.

$$\mathcal{L}_{exp} = \frac{1}{M} \sum_{k=1}^M \|Y_k - E\| \quad (3)$$

where M is the number of non-overlapping local regions, and Y is the average intensity value of a local region in the enhanced image. Exposure control loss inhibits underexposure/overexposure areas.

$$\mathcal{L}_{col} = \sum_{\forall (p,q) \in \varepsilon} (J^p - J^q)^2 \quad (4)$$

where J^p is the average intensity value of p channels in the enhanced image, and (p, q) is a pair of channels. The color constancy loss corrects the potential color bias in the enhanced image and establishes the relationship between the three

adjustment channels.

$$\mathcal{L}_{tv_A} = \frac{1}{N} \sum_{n=1}^N \sum_{c \in \xi} (\|\nabla_x A_n^c\| + \|\nabla_y A_n^c\|)^2 \quad (5)$$

where N is the number of iterations, ∇_x and ∇_y respectively represent the horizontal and vertical gradient operations. The illumination smoothness loss preserves the monotonic relationship between adjacent pixels. The total loss can be expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_{spa} + \mathcal{L}_{exp} + W_{col} \mathcal{L}_{col} + W_{tv_A} \mathcal{L}_{tv_A} \quad (6)$$

3.2 Expanding & Shrinking Entropy (ESE)

The following outlines our proposed domain adaptation approach. We define the source domain data and the corresponding labels $S = (x_i^s, y_i^s)$, and the unlabeled data in the target domain $T = (x_i^t)$. Through adversarial training and learning between source domain data and target domain data, the large domain gap can be lessened.

After image enhancement (cf., Section 3.1), the data in source domain and target domain are sent to a feature extractor F . The feature extractor uses a three-dimensional convolutional neural network as its backbone and outputs the video features, which are expressed as $F(x)$. It also serves as one of the inputs of classifier C . The other input is the weight matrix, which consists of the weight vector $W = [W_1, W_2, W_3 \dots W_i]$, where i represents the number of classes. The output of the classifier is applied to a fully connected layer with Softmax activation function, to obtain the network output result $S(W^T \bullet F(x))$, S represents the Softmax Function. For source domain data, the output is prediction, and for target domain data, the output is most confidence class.

In order to train the classifier to classify accurately, the weight vector in the weight matrix should point to the feature of the corresponding category. We visualize weight vectors as representations of each class. At the beginning of training, the representation is more likely to point to the source domain distribution because the source labels are dominant at this time. Then, the orientation is influenced by introducing each weight vector in the weight matrix from the target domain. The orientation is updated and trained by employing multiple iterations until a relatively stable and irrelevant orientation is generated for each classification. This procedure we defined as

“expanding entropy”. It can be described in a graphic way as the orientation of weight vector moves from being partial to the source domain to the target domain, that is, from order to disorder, which is also the essence of entropy. To achieve this, we propose a Shannon entropy H and increase it by calculating the similarity between W^T and unlabeled target data. Shannon entropy is calculated as follows,

$$H = -\sum_{(x,y) \in \{S,T\}} p(y = i|x) \log p(y = i|x) \quad (7)$$

$$\text{where : } p(y = i|x) = S(W^T \bullet F(x))$$

where i is the number of classes, F is the feature extractor, W^T is the weight matrix from the target domain and S is defined as the Softmax Function. $p(y = i|x)$ represents the probability prediction to class i

In order to reduce the gap between two domains, we need to obtain discriminative features from the target domain. In the procedure of “expanding entropy”, the classifier is learnt more target domain data by employing iterative training. The learning experience is conducive to the feature extractor to obtain more discriminative features. As a result, we propose reducing the entropy of unlabeled target data by training the feature extractor. By clustering the target domain to the source domain, that is, from a divergent state to a more ordered state, the entropy is decreasing, resulting in obtaining a more accurate classification. This step is just like the reverse of “expanding entropy”, which we define as “shrinking entropy”. We use the standard cross entropy loss $\mathcal{L}_{ce}$ to train feature extractor and classifier for classification. Cross entropy loss is calculated as:

$$\mathcal{L}_{ce} = \sum_{(x,y) \in \{S,T\}} -y \log S(F(x)). \quad (8)$$

By multiple training and iterations, repeating the process of “expanding entropy” and “shrinking entropy” yields the goal of domain adaptation.

3.3 Dynamic Pseudo Label (DPL)

In the training process of domain adaptation approach, the output of the classifier for target domain data is the most confidence class. The closer result to correct category, the better training effect we get. Inspired by the Flexmatch

method [34], we adopted the approach of marking pseudo labels for target domain data. The approach process is as follows.

First, the learning outcome of the classifier is judged by the evaluation accuracy of each class and using them to scale the threshold, as shown below:

$$T(k) = \alpha(k) \cdot \tau \quad (9)$$

where $T(k)$ is the elastic threshold of class K , and $\alpha(k)$ is the corresponding evaluation accuracy. The lower accuracy of the classification indicates a unsatisfactory learning outcome for this class, the lower the threshold should be, thus encouraging to learn more samples of this class. The learning outcome of the classification can be assessed by the sample number that is not less than the threshold value. The formula is as follows:

$$\varepsilon(k) = \sum_{i=1}^N \mathbf{1}(\max(p(y||un_i)) > \tau) \quad (10)$$

$$\bullet \mathbf{1}(\operatorname{argmax}(p(y||un_i)) = k)$$

The $\varepsilon(k)$ reflects the k class the learning outcome, $p(y||un_i)$ is the model’s prediction of unlabeled data un_i , and N is the total number of unlabeled data. The larger the $\varepsilon(k)$ is, the better learning outcome of this classification is. Since the estimated learning outcome is parameterized, its size varies with the number of samples contained in the dataset. Therefore, it is necessary to normalize $\varepsilon(k)$ so that the range of evaluation accuracy is between 0 to 1.

$$\beta(k) = \varepsilon(k) \quad (11)$$

$$T(k) = \beta(k) \cdot \tau$$

The evaluation accuracy $\beta(k)$ of the best learning class is equal to 1. The advantage is that lower thresholds for poorly learned categories, encouraging to learn more samples from those categories, simultaneously improving data utilization. After iterative training, the thresholds of the well-learning categories will be improved. Thus, forcing the network to obtain higher quality samples and output more accurate predictions.

4 Experiment

4.1 Datasets

We evaluate our approach on the HMDB $\leftrightarrow$ ARID dataset. Xu et al. [14] established and published the HMDB-ARID dataset for video domain adaptation studying. This dataset has 770 training

videos and 330 validation videos from HMDB51, and 2288 training videos and 823 validation videos from ARID. All the videos were taken from the original HMDB [7] and ARID [24] datasets. And they sampled 11 overlapping classes respectively from the 51/11 classes of HMDB/ARID. There are two settings to notice: HMDB $\rightarrow$ ARID, and the other is ARID $\rightarrow$ HMDB, where the arrow direction represents the source (arrow start) to the target (arrow head). We show the performance of our method in both settings. We use the official split provided by the author [14].

4.2 Implementation Details

We employ the PyTorch library to implement our method. We use R(2+1)D [36] networks as source and target clip feature encoder architectures. The source and target encoders are different and independent as they do not share parameters. Both of them are initialized with pre-trained model weights from the IG65M dataset. The input to the clip feature encoder is a clip of 3 channels $\times$ 32 frames $\times$ 112 $\times$ 112 pixels.

When it comes to feature extraction stage, we will get $512 \times 4 \times 7 \times 7$ output features. It should be noted that we divide the features into *unlabel_strong* and *unlabel_weak*, and enhance *unlabel_strong* through Zero-DCA method, without doing any operation on *unlabel_weak*. The backbone of the classification network is trained based on our proposed Z-APE method, which uses a linear network as our classifier. We set the batch sizes for labeled data and unlabeled data to 64 and 75, respectively, based on the actual situation. The network is trained at a learning rate of 10^{-3} for 300 iterations to ensure the best possible training results. In order to improve the universality of the model, we set the weight decay to 10^{-3} . Our model is optimized by using AdamW [37] as the optimizer.

4.3 Baselines

We evaluate whether domain shift between source and target will impact the result by testing using methods (I) “source only”: networks trained using only source data (the lower bound of the adaptation method), and (ii) “Target only”: networks trained using only target data (the upper bound of the adaptation method). Additionally,

we compare to 3 typical baselines for unsupervised domain adaptation as follows:

- Maximum Mean Discrepancy (MMD) [17]: The common-used domain discrepancy measurement MMD is used as a baseline. It directly replaces adversarial alignment with individual discrepancy measures.
- Maximum Classifier Discrepancy (MCD) [38]: Alignment by classifier inconsistencies. Classifiers are trained to maximize predicted inconsistencies in the target domain, implemented as L1 loss, finding examples that are not supported in the source domain.
- Deep Adaptation Network (DAN) [16]: By adding a few standard layers and a simple new gradient inversion layer, this adaptation behavior can be achieved in almost any feed-forward model.

Table 1 Results on the two settings for ARID-HMDB

Method	HMDB→ARID	ARID→HMDB
Source Only	47.6%	37.3%
MCD	48.1%	37.8%
MMD	51.6%	40.9%
DAN	52.1%	41.4%
Z-APE(ours)	56.9%	45.5%
Target Only	71.4%	62.5%

Table 1 shows that our proposed Z-APE achieves the best result under the HMDB→ARID setting and very competitive performance under the ARID→HMDB setting. To be more specific, our Z-APE achieves 45.5% accuracy for ARID→HMDB setting, indicating the improvement brought by Z-APE towards the lower bound of the ARID→HMDB setting is 8.2%. This is significantly higher than that brought by MCD(0.5%), MMD(3.6%) and DAN(4.1%). The larger improvement brought by Z-APE enables our network to perform better on HMDB→ARID setting. Under this setting, our Z-APE is also closer to the upper bound, with a gap of 14.5%, comparing to the ARID→HMDB setting(17.0%). By contrast, MCD, MMD and DAN are 23.3%, 19.8% and 19.3% below the upper bound respectively, which is significantly inferior to our proposed method Z-APE. This further demonstrates the superiority of Z-APE with respect to current video domain adaptation methods.

4.4 Ablation Study

Several ablation experiments are performed to compare the individual contributions of different components employed the proposed method.

Zero-DCA.

Table 2 shows the results of the Zero-DCA effect. Source Only is an R(2+1)D network trained on the source dataset, which we test directly on the target dataset without any adjustments. Target Only is a classifier network trained on the target dataset. We employ Zero-DCA to enhance the source data or target data, or both. The Zero-DCA component on its own gives an average 7.4% improvement over no enhancement. This shows that Zero-DCA is conducive to obtain information in dark environment. More interestingly, the results show that using Zero-DCA only at target can even more improve the performance by 1.2% over both at source and target. That is augmenting only the ARID dataset is more effective than augmenting both datasets at the same time, because all of the ARID videos lacked brightness while only some of the HMDB videos do. Unnecessary augmentation to the dataset can result in information corruption. Therefore, application of Zero-DCA on the high-light dataset does not improve results, whereas application of Zero-DCA on sources or targets containing the low-light dataset can significantly improves the performance.

Table 2 Ablation experiment of Zero-DCA, on HMDB → ARID

Method	Source	Target	Top-1 acc(%)
Z-APE(ours)	×	✓	58.1%
Z-APE(ours)	✓	✓	56.9%
Z-APE(ours)	✓	×	51.9%
Z-APE w/o Zero-DCA	×	×	48.2%

DPL.

Table 3 shows the use of the DPL component in the overall approach. Here, all objects contain the source and target and trained after the enhancement of Zero-DCA. We train our approach with two Settings, (I) the Zero-DCA component (our complete model) that includes the DPL. (ii) Zero-DCA component without DPL. The results show that DPL component on its own gives a 4.8% improvement. Therefore, the dynamic

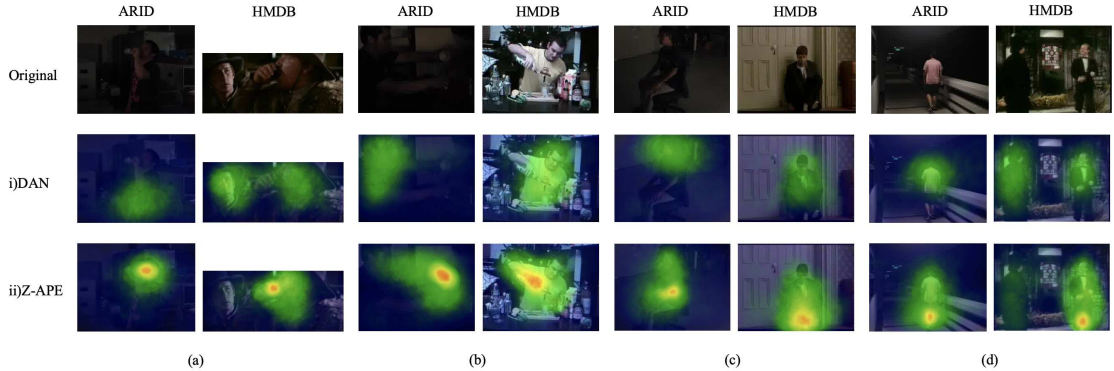


Fig. 2 Class activation maps (CAMs) on ARID and HMDB, employing i)DAN and ii)Z-APE. The CAMs are obtained from four actions: (a) “Drink”; (b) “Pour”; (c) “Sit”; (d) “Walk”. We also present the original frames at the top row from which the CAMs are computed.

pseudo label component can flexibly adjust the pseudo-label threshold to adapt to the alignment effect of the model. The application of our DPL can also effectively reclassify mis-classified categories and adjust the alignment of source and target domains.

Table 3 Ablation experiment of DPL, on HMDB $\rightarrow$ ARID

Method	DPL	Top-1 acc(%)
Z-APE(ours)	✓	56.9%
Z-APE w/o DPL	×	52.1%

4.5 Qualitative Analysis

Zero-DCA.

In order to better understand the effect of the proposed Zero-DCA module, we show the target video class activation map CAM [39] in Figure 2. We show the CAMs with and without the Zero-DCA (baseline/ours) model. The baseline model without Zero-DCA often does not perform well on noticing human body. However, the method with Zero-DCA enhances the human part more flexibly, allowing the model to focus more on human actions. The Zero-DCA model can be better extended to the dark environment, while the model without Zero-DCA is greatly affected by the dark environment. Thus Zero-DCA shows a much powerful outcome on the HMDB $\rightarrow$ ARID setting.

DPL.

The result shows how the DPL component can better distinguish similar actions. For example, the “sit” and “stand up” videos, while the person’s body was in mid-air, the baseline without DPL has a nearly 50-50 chance of confusing the two categories. When using the DPL, the model does not categorize ambiguous unlabeled clips, but rather categorizes them dynamically as the video progresses forward. Then we can observe that at the end of the video, when the characteristics of the action are more obvious, the model dynamically places the label of the action and turns it into a correct classification. Therefore, the DPL model has the function of dynamically generating correct pseudo labels for all unlabeled clips.

5 Conclusion

In conclusion, we propose a new method for Video Unsupervised Action Recognition, named Z-APE. The new Z-APE improves domain adaptive problem of high-light domain to low-light domain by employing a series of luminance curve augmentation network (cf., Subsection 3.1) to enhance the target domain data in dark environment. It also well takes advantage of ESE (cf., Subsection 3.2) and DPL (cf., Subsection 3.3) components to reduce the large domain gap between source domain and target domain. The method results in gaining a better accuracy on the HMDB-ARID dataset. Then we further optimize our approach through an ablation study. And the qualitative results indicate our approach’s effectiveness.

Declarations

- Ethics approval and consent to participate
Not
- Consent of publication
Yes
- Availability of data and materials
Not
- Competing interests
No, I declare that the authors have no competing interests as defined by Springer, or other interests that might be perceived to influence the results and/or discussion reported in this paper.
- Funding
No funding.
- Authors' contributions
Zhi Chen and Zijun Fan contributed to the conception of the study and contributed significantly to analysis and manuscript preparation; Yongjie Li performed the experiment; Shan Lin helped perform the analysis with constructive discussions; All authors reviewed the manuscript.
- Acknowledgments
I would like to acknowledge Yuecong Xu for inspiring my interest in the development of computer vision.

References

- [1] Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., et al.: End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316 (2016)
- [2] Lee, D.-H., Liu, J.-L.: End-to-end deep learning of lane detection and path prediction for real-time autonomous driving. *Signal, Image and Video Processing*, 1–7 (2022)
- [3] Mohammadian Fini, R., Mahlouji, M., Shahidinejad, A.: Real-time face detection using circular sliding of the gabor energy and neural networks. *Signal, Image and Video Processing* **16**(4), 1081–1089 (2022)
- [4] Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3339–3348 (2018)
- [5] Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-weak distribution alignment for adaptive object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6956–6965 (2019)
- [6] Pan, W., Zhou, X., Zhou, T., Chen, Y.: Fake license plate recognition in surveillance videos. *Signal, Image and Video Processing*, 1–9 (2022)
- [7] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: *2011 International Conference on Computer Vision*, pp. 2556–2563 (2011). IEEE
- [8] Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
- [9] Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., Zisserman, A.: A short note about kinetics-600. arXiv preprint

arXiv:1808.01340 (2018)

- [10] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
- [11] Liang, Z., Yin, M., Gao, J., He, Y., Huang, W.: View knowledge transfer network for multi-view action recognition. *Image and Vision Computing* **118**, 104357 (2022)
- [12] Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going deeper with image transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 32–42 (2021)
- [13] Alhashim, I., Wonka, P.: High quality monocular depth estimation via transfer learning. arXiv preprint arXiv:1812.11941 (2018)
- [14] Xu, Y., Yang, J., Cao, H., Mao, K., Yin, J., See, S.: Aligning correlation information for domain adaptation in action recognition. arXiv preprint arXiv:2107.04932 (2021)
- [15] Xu, Y., Yang, J., Cao, H., Chen, Z., Li, Q., Mao, K.: Partial video domain adaptation with partial adversarial temporal attentive network. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9332–9341 (2021)
- [16] Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: *International Conference on Machine Learning*, pp. 1180–1189 (2015). PMLR
- [17] Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: *International Conference on Machine Learning*, pp. 97–105 (2015). PMLR
- [18] Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. *Advances in neural information processing systems* **31** (2018)
- [19] Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Adversarial dropout regularization. arXiv preprint arXiv:1711.01575 (2017)
- [20] Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014)
- [21] Dubey, A., Gupta, O., Raskar, R., Naik, N.: Maximum-entropy fine grained classification. *Advances in neural information processing systems* **31** (2018)
- [22] Erkan, A., Altun, Y.: Semi-supervised learning via generalized maximum entropy. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 209–216 (2010). JMLR Workshop and Conference Proceedings
- [23] Saito, K., Kim, D., Sclaroff, S., Darrell, T., Saenko, K.: Semi-supervised domain adaptation via minimax entropy. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8050–8058 (2019)
- [24] Xu, Y., Yang, J., Cao, H., Mao, K., Yin, J., See, S.: Arid: A new dataset for recognizing action in the dark. In: *International Workshop on Deep Learning for Human Activity Recognition*, pp. 70–84 (2021). Springer
- [25] Sankaranarayanan, S., Balaji, Y., Jain, A., Lim, S.N., Chellappa, R.: Learning from synthetic data: Addressing domain shift for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3752–3761 (2018)
- [26] Munro, J., Damen, D.: Multi-modal domain adaptation for fine-grained action recognition. *IEEE* (2020)
- [27] Choi, J., Sharma, G., Schuler, S., Huang, J.B.: Shuffle and Attend: Video Domain Adaptation, (2020)
- [28] Kalfaoglu, M., Kalkan, S., Alatan, A.A.: Late temporal modeling in 3d cnn architectures with bert for action recognition. In: *European Conference on Computer Vision*, pp. 731–747 (2020). Springer

- [29] Singh, A., Chakraborty, O., Varshney, A., Panda, R., Feris, R., Saenko, K., Das, A.: Semi-supervised action recognition with temporal contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10389–10399 (2021)
- [30] Li, J., Shlizerman, E.: Sparse semi-supervised action recognition with active learning. arXiv preprint arXiv:2012.01740 (2020)
- [31] Xiao, J., Jing, L., Zhang, L., He, J., She, Q., Zhou, Z., Yuille, A., Li, Y.: Learning from temporal gradient for semi-supervised action recognition. arXiv preprint arXiv:2111.13241 (2021)
- [32] Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: European Conference on Computer Vision, pp. 213–226 (2010). Springer
- [33] Huang, Y., Xu, H.: Fully convolutional network with attention modules for semantic segmentation. *Signal, Image and Video Processing* **15**(5), 1031–1039 (2021)
- [34] Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., Shinozaki, T.: Flex-match: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems* **34** (2021)
- [35] Guo, C., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S., Cong, R.: Zero-reference deep curve estimation for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1780–1789 (2020)
- [36] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6450–6459 (2018)
- [37] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2017)
- [38] Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3723–3732 (2018)
- [39] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)