Supplementary Material for Watching it in Dark: A Target-aware Representation Learning Framework for High-Level Vision Tasks in Low Illumination

Yunan Li[1,2,3], Yihao Zhang[1,2], Shoude Li[1,2], Long Tian[1], Dou Quan[4], Chaoneng Li[5], and Qiguang Miao*[1,2,3]

[1] School of Computer Science and Technology, Xidian University, Xi'an, China
[2] Xi'an Key Laboratory of Big Data and Intelligent Vision, Xi'an, China
[3] Key Laboratory of Smart Human-Computer Interaction and Wearable Technology of Shaanxi Province, Xi'an, China
[4] School of Artificial Intelligence, Xidian University, Xi'an, China
[5] School of Electronics and Information Engineering, Lanzhou Jiaotong University, Lanzhou, China
{yunanli, tianlong, quandou, qgmiao}@xidian.edu.cn, {zhangyihao,shoulderlee}@stu.xidian.edu.cn, chaonengli@lzjtu.edu.cn

1 Experimental Details

In this section, we will introduce the experimental details for the training and evaluation on various datasets for three tasks of object recognition, object detection, and action recognition.

1.1 Object Recognition

Settings for CODaN. Following [6,8], we train our model on the well-lit test set and evaluate it on the low-lit test set. We utilize ResNet-18 [3] as the backbone architecture, with images set to 224×224 pixels, and optimize parameters with the Adam optimizer. The training process stops after 180 epochs. The initial learning rate is set to 0.002. We apply multi-step learning rate decay at predetermined intervals, reducing the learning rate by a factor of 0.1 at the 60th and 120th epochs.

Settings for ExDark. According to [7], we extract 400 normal light images per class from the COCO dataset as training data, totaling 4,800 images, and conduct evaluations on the complete ExDark dataset. Consistent with [7], we employ the deeper ResNet-50 [3] as the backbone, with images set to 224×224 pixels, and optimize parameters using SGD with momentum of 0.9 and weight decay of 1e-5. The training is continued over 50 epochs with an initial learning rate of 0.0005, and we adopt the scheme of linear decay.

* Corresponding author

1.2 Object Detection

For our study on low-light object detection, we utilize the ExDark dataset. To ensure compatibility with the 12 categories of the ExDark dataset, we randomly select 1,000 images for each category from the COCO and VOC datasets, obtaining a total of 12,000 images for our normal light training dataset. The ExDark validation set, partitioned in a manner consistent with MAET [2], serves as our low-light test dataset. We employ YOLOv3 [10] as the backbone. Input images are first resized to 608×608, maintaining the original aspect ratio. We employ SGD with a momentum of 0.937 and a weight decay of 0.0005, for network optimization. The initial learning rate is set to 0.01 and spans over 30 epochs. We adopt a linear decay strategy.

1.3 Action Recognition

For action recognition, we use the normal light dataset provided in the CVPR UG2+ 2022 challenge as our training data, which consists of 2.6k normal light video clips from HMDB51 [5], UCF101 [11], Kinetics-600 [4], and Moments in Time [9]. After training, we evaluate our method on the official test split of the ARID dataset released by CVPR UG2+, which contains videos of 11 classes under low-light conditions. In our experiments, we employ I3D-RGB [1] based on 3D-ResNet as the backbone. We randomly extract 16 frames from each video, with at least 2 frames' intervals between sampled frames, and perform random cropping of 224×224 after scaling the videos from 0.8 to 1.25. We employ SGD optimizer with a momentum of 0.9 and a weight decay of 0.0001, and stop the training process after 60 epochs. The initial learning rate is set to 0.01. We apply a fixed-interval learning rate adjustment strategy, executing learning rate decay by a factor of 0.1 after training with the 20,000th, 40,000th, and 80,000th samples.

2 Additional Experiments

In this section, we present more results of evaluating the impact of different components of our framework for the other two tasks: object detection and action recognition.

Impacts of framework components for detection. As shown in Table 1, for object detection tasks, only using image-level alignment hampers the performance. It is primarily due to the introduction of noise during the alignment process. This noise can interfere with the target detector's ability to extract features from images; however, a synergistic effect is observed when combined with feature-level domain alignment, enhancing the detector's performance. This is because the performance degradation caused by the noise introduced at the image level is mitigated by the complementary nature of feature-level domain alignment. Meanwhile, the combination of contrastive learning with masking can

Table 1: Impacts of framework components for object detection.

two-level bi-directional alignment		target-aware mask representation learning		mAP@0.5
image-level	feature-level	target-aware mask	contrastive learning	
×	×	×	×	58.7
√	×	×	×	56.8
√	√	×	×	60.3
√	√	√	×	60.8
√	√	×	√	60.5
√	√	√	√	64.2

lead to another significant improvement, which means finding the target region and improving feature representation are equally important for detection tasks.

Impacts of framework components for action recognition. As Table 2 illustrates, both image-level and feature-level domain alignments contribute to performance improvement for action recognition tasks. However, compared with the other two tasks, the target-aware mask should be associated with contrastive learning for boosting performance. This is attributed to the complex illumination changes between frames for video action recognition tasks, which pose a significant challenge to feature extraction. With TGMM, the movement-relevant target region can be extracted, but the robust feature representation is not so easy to obtain when compared with object recognition tasks. However, it is still effective to provide a more specific task-oriented view and consequently improve the performance of feature representation learning.

Table 2: Impacts of components for action recognition.

two-level bi-directional alignment		target-aware mask representation learning		Accuracy(%)
image-level	feature-level	target-aware mask	contrastive learning	
×	×	×	×	47.02
√	×	×	×	50.09
√	√	×	×	51.63
√	√	√	×	51.28
√	√	×	√	54.43
√	√	√	√	58.08

Effect of different masking strategies. We compare our target-aware masking with random masking in Table 3. This comparison takes the bi-directional

domain alignment results as the baseline. As observed, random region masking yields an accuracy of 61.12%, which even decreases at about 0.52% compared to the baseline. This is because although masking can randomly remove redundant information, it comes at the cost of losing some useful features, especially after domain alignment. In contrast, our target-aware masking strategy achieves sufficient improvement since it strikes a balance between preserving useful information and fostering sample diversity.

Table 3: Results of different masking strategies

Strategies	Accuracy(%)
random	61.12
target-aware	64.40

3 Technical Supplement

Our TGMM is based on the Gaussian Mixture Model (GMM) framework described in [12], but it extends it to incorporate temporal dynamics for video-based tasks. It can be formulated as:

$$B(x) = \sum_{i=1}^{K} \omega_{i,t}(x) \cdot \mathcal{N}(I_t(x); \mu_{i,t}(x), \sigma_{i,t}(x)), \tag{1}$$

where $B(x)$ represents the distribution map of TGMM, $\mathcal{N}$ denotes a Gaussian distribution, and $I_t(x)$ is the observed value of pixel x at time t. The "distribution map of TGMM" refers to the combination of these Gaussian components that collectively describe the statistical behavior of each pixel over time. The weight $\omega_{i,t}(x)$ indicates the relevance of the i-th Gaussian component for modeling the pixel x at time t, $\mu_{i,t}(x)$ is the expected pixel value for component i at frame t, and $\sigma_{i,t}(x)$ is the covariance matrix, reflecting the variability or stability of the pixel intensity. Each Gaussian component in the model represents a possible background state (e.g., lighting conditions). A higher $\sigma_{i,t}(x)$ suggests that the pixel x is more likely influenced by motion and belongs to a dynamic, salient object. The movement of the target entity will lead to a mismatch between the pixel values within the corresponding region and the Gaussian distribution of the background model. When the mismatching becomes apparent enough, we can infer that it is caused by the target. Subsequently, the dynamic attention map can be obtained via binary processing, where pixels significantly deviating from the background model are marked as part of the target, while others are marked as background. Binary processing means image binarization, where pixels are classified as non-salient (0) or salient (1).

4 Visualization

We also present the visualization of target recognition tasks using different strategies on the night test set of CODaN, encoded by t-SNE. Figure 1 illustrates the feature distribution when employing different strategies, including 1) directly applying the daytime trained model to low-lit data, 2) achieving low light enhancement with Zero-DCE before classification, 3) employing our bidirectional domain alignment, and 4) our entire framework. It is observed that low-light enhancement has a limited effect on feature extraction enhancement. In contrast, the bi-directional domain alignment at both the image and feature levels significantly improves the extraction effectiveness. Notably, with feature representation learning, our framework reduces the distance between features of samples within the same class and correspondingly leads to a better feature clustering performance, thereby extracting more robust high-level semantic features in low-light conditions to enhance performance.

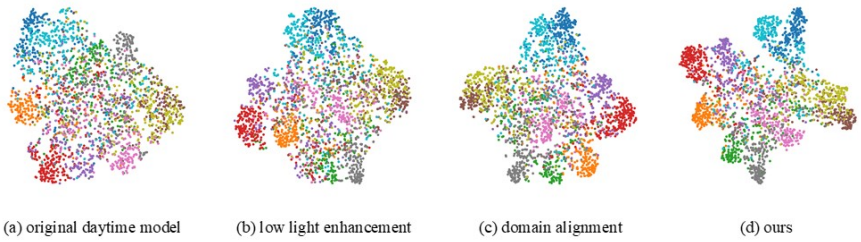

Fig. 1: Visualizing the projection of feature distribution of low-illumination conditions of different components. (a) Original model trained with daytime data. (b) With low light enhancement. (c) With our domain alignment strategy. (d) With our entire framework.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017) 2
2. Cui, Z., Qi, G.J., Gu, L., You, S., Zhang, Z., Harada, T.: Multitask aet with orthogonal tangent regularity for dark object detection. In: ICCV. pp. 2553–2562 (2021) 2
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 1
4. 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) 2
5. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: ICCV. pp. 2556–2563 (2011) 2
6. Lengyel, A., Garg, S., Milford, M., van Gemert, J.C.: Zero-shot day-night domain adaptation with a physics prior. In: ICCV. pp. 4399–4409 (2021) 1
7. Loh, Y.P., Chan, C.S.: Getting to know low-light images with the exclusively dark dataset. CVIU **178**, 30–42 (2019) 1

8. Luo, R., Wang, W., Yang, W., Liu, J.: Similarity min-max: Zero-shot day-night domain adaptation. In: ICCV. pp. 8104–8114 (2023) 1
9. Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S.A., Yan, T., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C., et al.: Moments in time dataset: one million videos for event understanding. IEEE TPAMI **42**(2), 502–508 (2019) 2
10. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018) 2
11. 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) 2
12. Zivkovic, Z., van der Heijden, F.: Efficient adaptive density estimation per image pixel for the task of background subtraction. PR Letters **27**(7), 773–780 (2006) 4