HLA-Face: Joint High-Low Adaptation for Low Light Face Detection (Supplementary Material)

Wenjing Wang, Wenhan Yang, Jiaying Liu*
Wangxuan Institute of Computer Technology, Peking University, Beijing, China

Contents

1. Implementation Details

1.1. Network Architecture

For brightening, the detailed architecture of our $E(\cdot)$ is shown in Fig. 1.

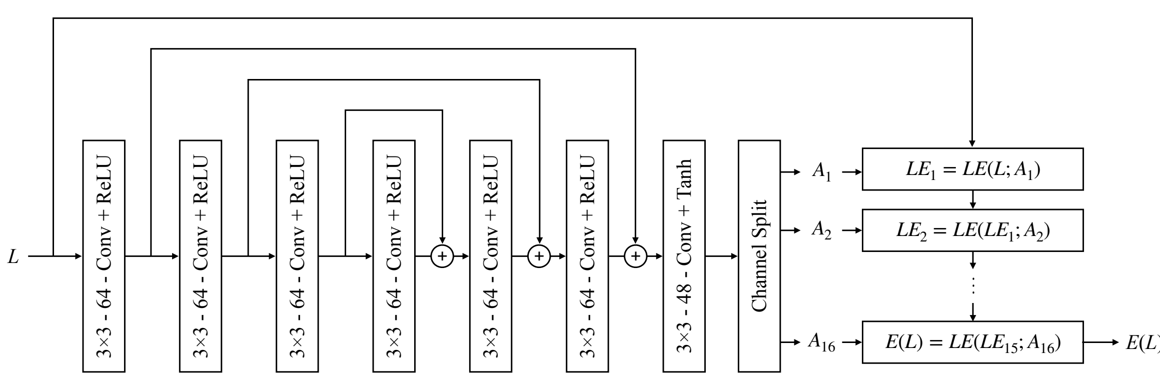


Figure 1: The architecture of our brightening network.

We use DSFD [15] as the face detector. DSFD uses an extended VGG16 [23] backbone, which is shown in Fig. 2. It extracts a 6-layer multi-scale feature: conv3_3, conv4_3, conv5_3, conv_fc7, conv6_2, and conv7_2.

*Corresponding author. Our project is publicly available at: https://daooshee.github.io/HLA-Face-Website/

For face detection, these six layers are first used as the first shot detection layers. Then, a Feature Enhance Module (FEM) proposed by DSFD is used to transfer these six feature maps into six enhanced feature maps. The enhanced features construct the second shot detection layers. Please refer to [15] for more details.

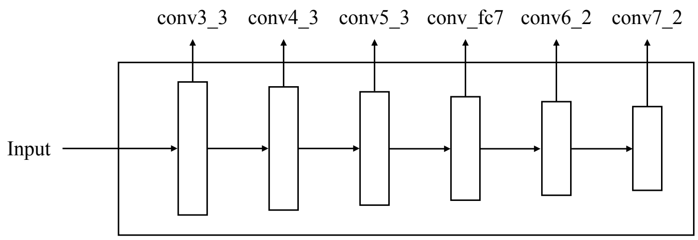


Figure 2: The architecture of our backbone.

Our self-supervised learning headers are added after all these six layers. The headers share the same architecture of Conv-Conv-FC, which is shown in Fig. 3. The output channels of the convolutional layers are all 64. For the rotation prediction pretext task, the output channel K of the last fully connected layer is 4. For the jigsaw puzzling pretext task, $K = 30$. For contrastive learning, $K = 128$. Losses are first computed for each layer, then added up to constitute the final objective.

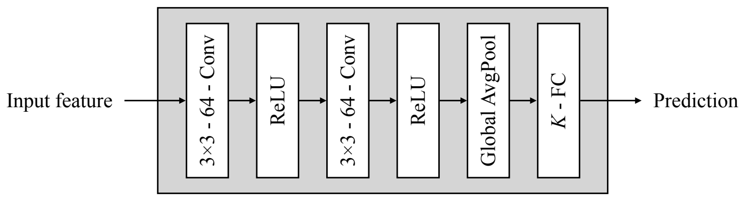


Figure 3: The architecture of our self-supervised learning header.

An example of the jigsaw pretext task on $E(L)$ is shown in Fig. 4. Given a 3×3 shuffled image $E(L)_{jig}$, the extended VGG16 backbone first extracts six feature maps, which are then processed by six headers. After that, the cross entropy loss is computed for each $F_{jig}^{E(L)}$. At last, the sum of the six losses is used as the final jigsaw loss on $E(L)$.

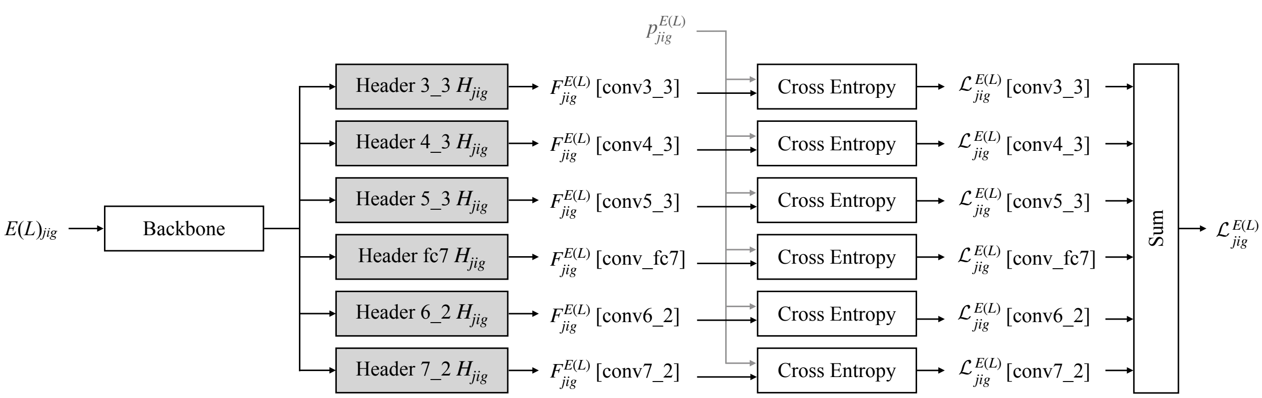


Figure 4: The architecture of our framework for the jigsaw pretext task on $E(L)$.

The implementation is based on `https://github.com/yxlijun/DSFD.pytorch`. We implement the whole framework using the PyTorch library.

1.2. Network Training

All experiments are based on WIDER FACE [26] and DARK FACE [27]. Our model is allowed to use the labels of WIDER FACE, but not allowed to use the labels of DARK FACE.

The model is first pre-trained on WIDER FACE for face detection only, then fine-tuned with both WIDER FACE and the images of DARK FACE. Pretraining follows the same process of original DSFD [15]. For fine-tuning, the batch size is set to 8. The self supervised learning layers are initialized by the "Kaiming" method [11]. We use SGD with 0.9 momentum and 5e-4 weight decay. The learning rate is set to 1e-4 for the first 20k iterations, and 1e-5 for another 40k iterations. Fine-tuning takes about 15 hours with two GeForce RTX 2080Ti.

In the objective, we set $\lambda_{det} = 1$, $\lambda_{E(L)\leftrightarrow H} = 0.05$, $\lambda_{H\leftrightarrow D(H)} = 0.05$, $\lambda_{E(L)\uparrow} = 0.05$. Our contrastive learning is based on MOCO [10]. Similar to MOCO, the temperature is set to $\tau = 0.07$. Different from MOCO, the momentum coefficient m is set to 0.99 and the number of negative samples N is set to 2048. In this way, the training process can be accelerated. For augmentation, we follow MOCO v2 [2] to use random resized crop, color jitter, random gray scale, random gaussian blur, and random horizontal flip.

With GeForce RTX 2080Ti and Intel Xeon E5-2650, the average inference time for 1080×720 images is 9.315s, only 0.038s (the time for $E(\cdot)$) slower than the original DSFD.

1.3. Benchmarking Settings

For WIDER FACE, we use the official train/val setting. WIDER FACE is not needed for testing. For DARK FACE, we use the official train/test setting, and further split 500 images from the training set for validation. Finally, there are 5500 images for training, 500 images for validation, and 4000 images for testing.

Following [27], performance is measured by mean Average Precision (mAP), and evaluated with the official tool of DARK FACE: https://github.com/Ir1d/DARKFACE_eval_tools.

The code sources of all compared methods are shown in Table 1. Three unsupervised domain adaptation methods, OSHOT [6], Progressive DA [12], and Pseudo Labeling [14], are reimplemented based on DSFD. WIDER FACE pre-training is used for all methods that need to be trained, including all ablation study versions, and compared methods of categories Darkening, Unsupervised DA, and Fully Supervised.

Table 1: The code sources of compared methods.

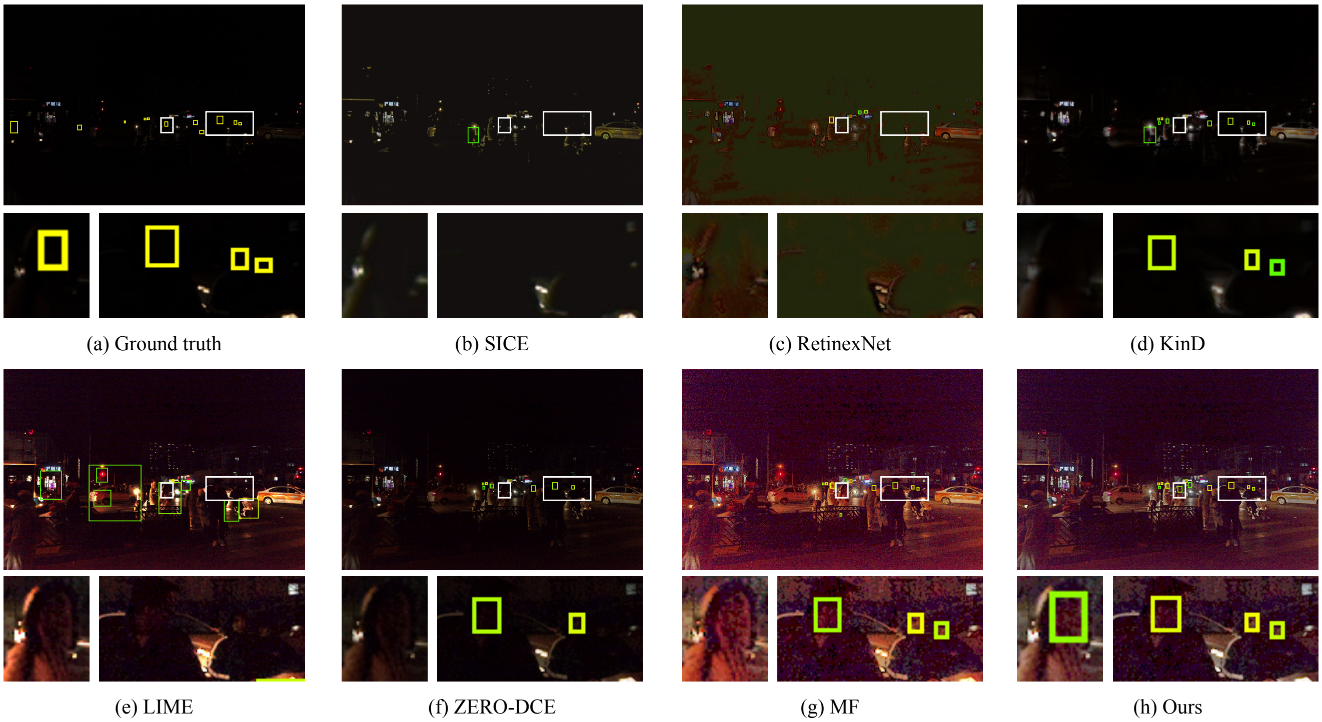
Method	Link
Faster-RCNN [21]	https://github.com/hdjsjyl/face-fasterrcnn.pytorch
SSH [19]	https://github.com/dechunwang/SSH-pytorch
RetinaFace [5]	https://github.com/biubug6/Pytorch_Retinaface
SRN [3]	https://github.com/ChiCheng123/SRN
SFA [18]	https://github.com/shiluo1990/SFA
PyramidBox [24]	https://github.com/yxlijun/Pyramidbox.pytorch
Small Hard Face [30]	https://github.com/bairdzhang/smallhardface
DSFD [15]	https://github.com/yxlijun/DSFD.pytorch
SICE [1]	https://github.com/csjcai/SICE
RetinexNet [25]	https://github.com/weichen582/RetinexNet
KinD [29]	https://github.com/zhangyhuaee/KinD
LIME [9]	https://sites.google.com/view/xjguo/lime
Zero-DCE [8]	https://github.com/Li-Chongyi/Zero-DCE
MF [7]	https://github.com/baidut/BIMEF
MUNIT [13]	https://github.com/NVlabs/MUNIT
CycleGAN [31]	https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix
CUT [20]	https://github.com/taesungp/contrastive-unpaired-translation

2. More Performance Analysis

2.1. More Comparison Results

In this section, we present more subjective comparison results. As shown in Fig. 5-10, our method can better locate the target faces and less recognize non-face objects as faces.

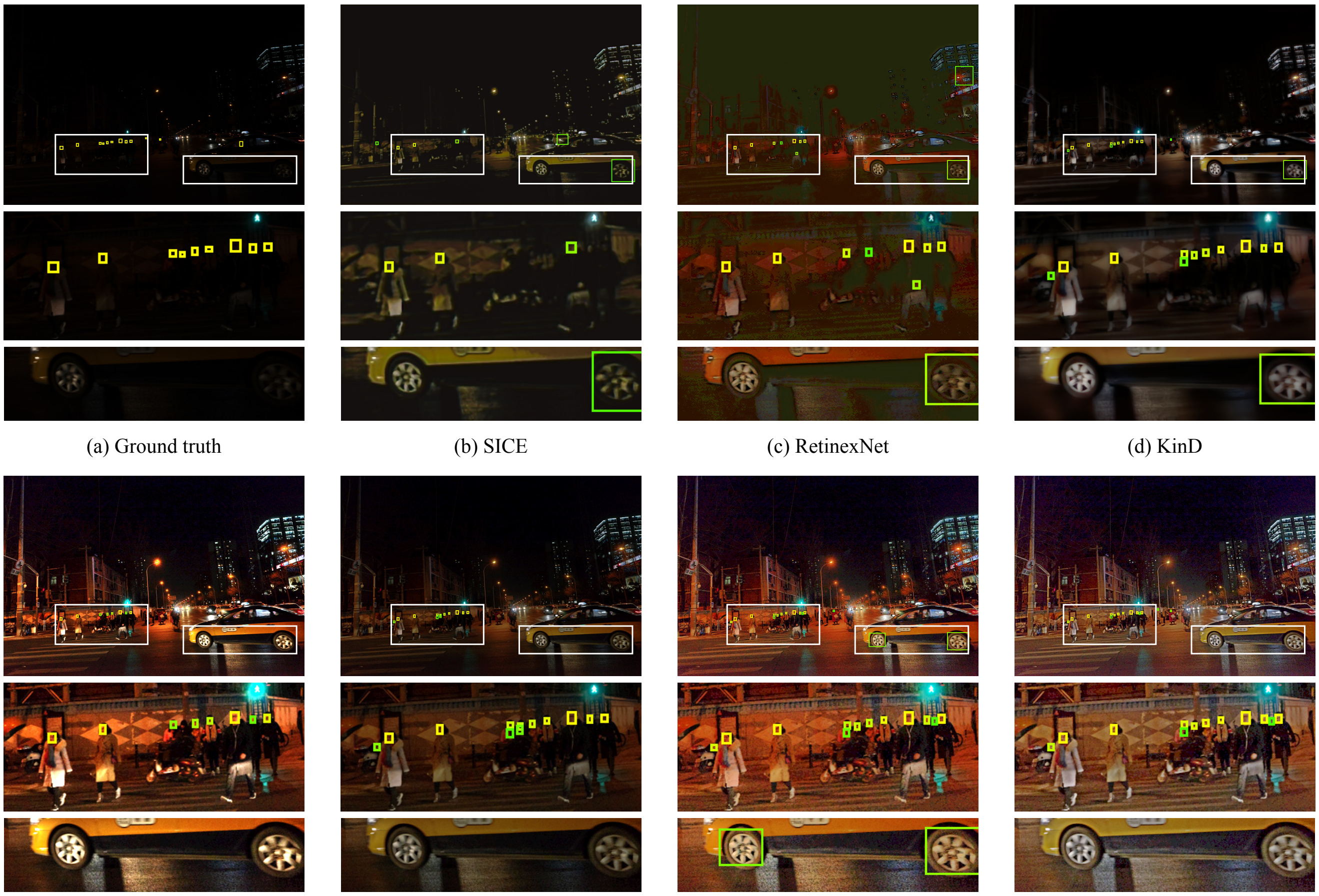
In Fig. 10, we show an interesting failure case where we "wrongly" detect a face on an advertisement board. This is because only the faces of real pedestrians are labeled in DARK FACE, while all kinds of faces are labeled in WIDER FACE.



(a) Ground truth (b) SICE (c) RetinexNet (d) KinD

(e) LIME (f) ZERO-DCE (g) MF (h) Ours

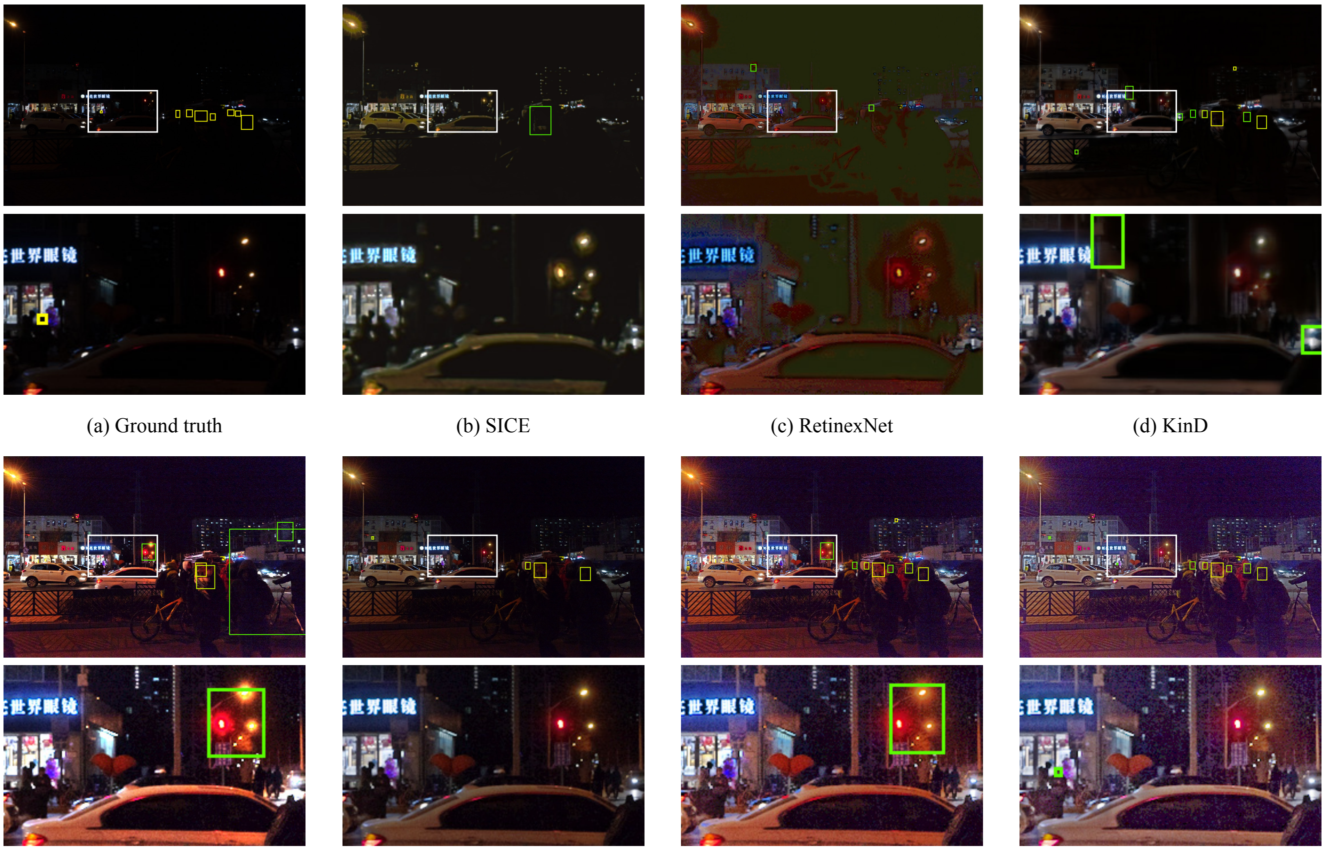
Figure 5: More qualitative comparison of different enhancement-based methods. (a) Input low light image and the ground truth boxes. (b)-(g) Results of low-light enhancement methods with DSFD. (h) Our result. The color of the bounding boxes indicates confidence. Yellow indicates high confidence, while green vice versa.



(a) Ground truth (b) SICE (c) RetinexNet (d) KinD

(e) LIME (f) ZERO-DCE (g) MF (h) Ours

Figure 6: More qualitative comparison of different enhancement-based methods. (a) Input low light image and the ground truth boxes. (b)-(g) Results of low-light enhancement methods with DSFD. (h) Our result. The color of the bounding boxes indicates confidence. Yellow indicates high confidence, while green vice versa.



(a) Ground truth (b) SICE (c) RetinexNet (d) KinD

(e) LIME (f) ZERO-DCE (g) MF (h) Ours

Figure 7: More qualitative comparison of different enhancement-based methods. (a) Input low light image and the ground truth boxes. (b)-(g) Results of low-light enhancement methods with DSFD. (h) Our result. The color of the bounding boxes indicates confidence. Yellow indicates high confidence, while green vice versa.

(a) Ground truth (b) SICE (c) RetinexNet (d) KinD

(e) LIME (f) ZERO-DCE (g) MF (h) Ours

Figure 8: More qualitative comparison of different enhancement-based methods. (a) Input low light image and the ground truth boxes. (b)-(g) Results of low-light enhancement methods with DSFD. (h) Our result. The color of the bounding boxes indicates confidence. Yellow indicates high confidence, while green vice versa.

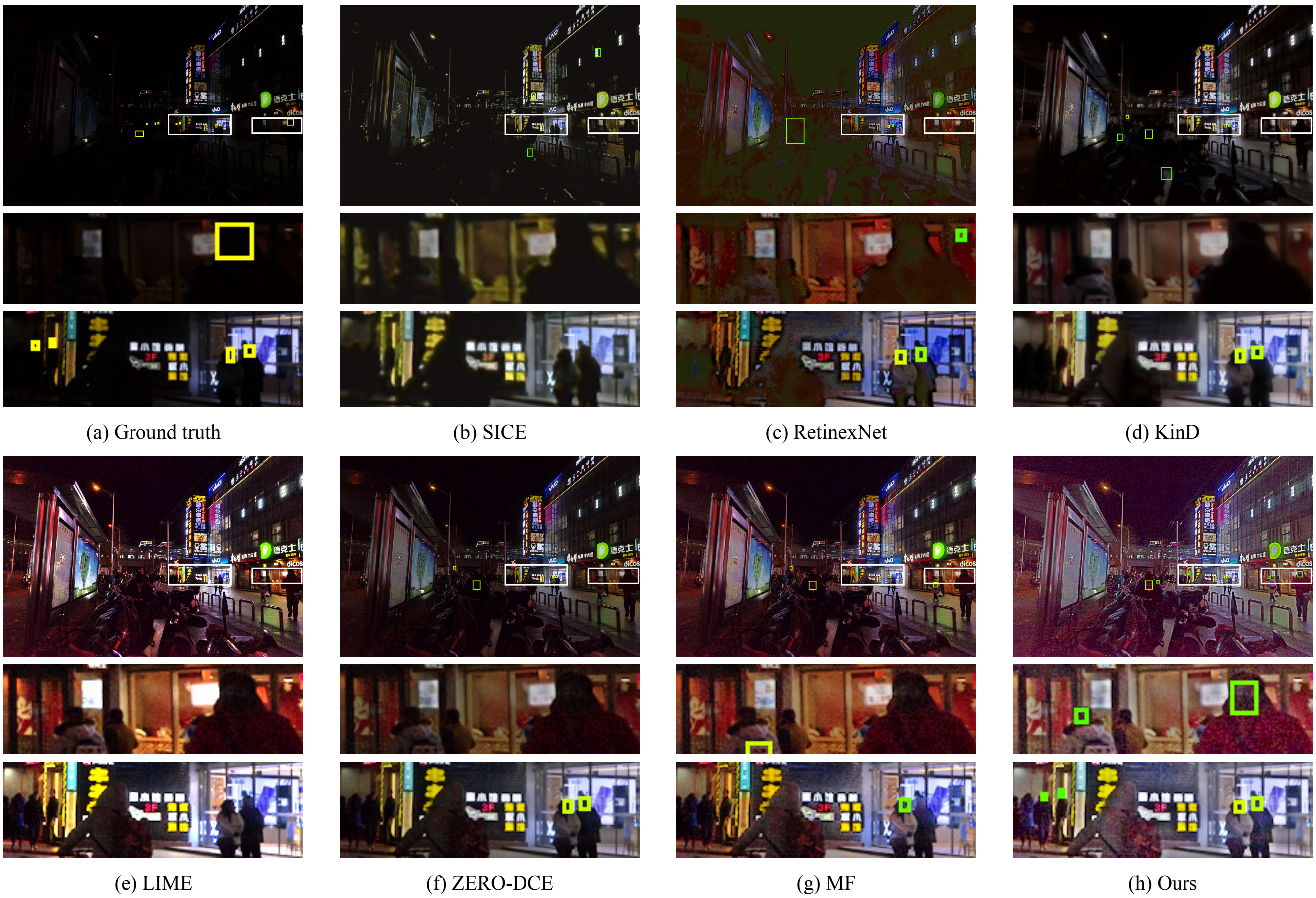
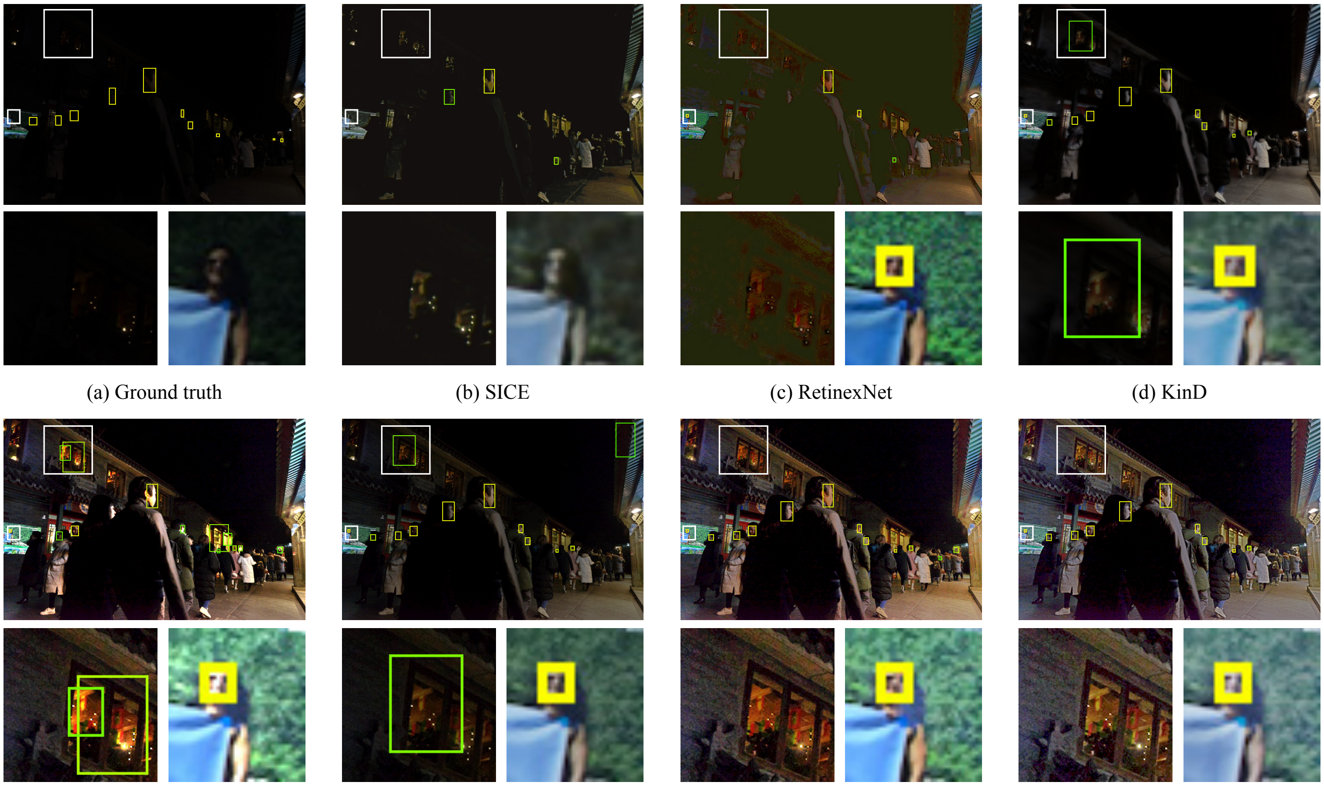


Figure 9: More qualitative comparison of different enhancement-based methods. (a) Input low light image and the ground truth boxes. (b)-(g) Results of low-light enhancement methods with DSFD. (h) Our result. The color of the bounding boxes indicates confidence. Yellow indicates high confidence, while green vice versa.



(a) Ground truth (b) SICE (c) RetinexNet (d) KinD

(e) LIME (f) ZERO-DCE (g) MF (h) Ours

Figure 10: More qualitative comparison of different enhancement-based methods. (a) Input low light image and the ground truth boxes. (b)-(g) Results of low-light enhancement methods with DSFD. (h) Our result. The color of the bounding boxes indicates confidence. Yellow indicates high confidence, while green vice versa.

2.2. Precision-Recall Curves for Ablation Study

The Precision-Recall (PR) curves for ablation study are shown in Fig. 11. The full version of our framework achieves the best performance, demonstrating the effectiveness of our technical designs.

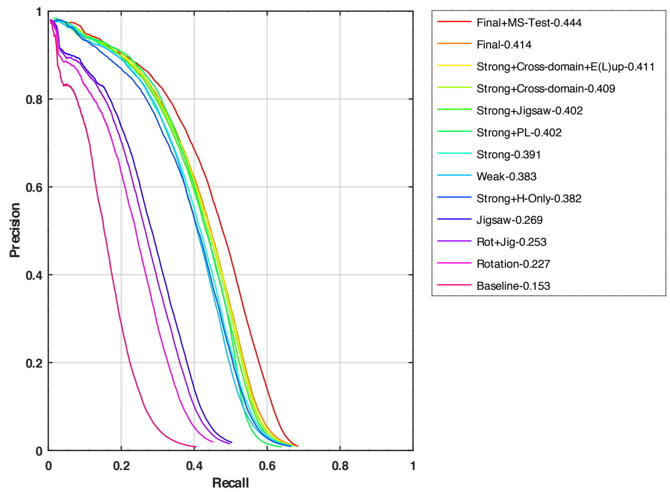


Figure 11: Precision-Recall (PR) curves for ablation study on DARK FACE.

2.3. Performance on WIDER FACE

Our model can detect normal light faces as well. To avoid over-exposure, we skip $E(\cdot)$ when the average pixel value of the image is higher than 45, which does not affect our original detection on DARK FACE. As shown in Table 2, the performance on WIDER FACE is even improved, which may be because that our adaptation makes the model more robust to noise.

Table 2: AP (%) on WIDER FACE

	Easy	Medium	Hard
DSFD	94.6	93.7	88.0
Ours	**95.0**	**93.9**	**88.3**

2.4. Real World Cases

Given a random unseen image I_m, our model can quickly adapt to it by fast one-shot fine-tuning. We first crop patches from I_m, treat these patches as the new L domain, then fine-tune the model with $\mathcal{L}_{E(L)\leftrightarrow H}$, $\mathcal{L}_{E(L)\uparrow}$, and the losses of $E(\cdot)$ for 100 steps (about 2 minutes). As shown in Fig. 12, the clothes button wrongly detected by DSFD and our model can be corrected after one-shot fine-tuning.

Test image | Enhanced for better visualization | DSFD | Ours | w/ One-shot

Figure 12: Effectiveness of fast one-shot adaptation.

2.5. Compared with UG2 Solutions

Details of UG2 solutions are in [27]. Compared with UG2 top 3-10 teams, our fine-tuned baseline (46% mAP) performs better probably because: (1) Baseline: U-Net [22] (Team PHI-AI) or LIME [9] + DPSR [28] + BM3D [4] (Team tjfirst) may distort details and hurt performance. (2) Domain gap: the testing set is more difficult than the training set. Many faces in the testing set are much smaller. Some solutions may over-fit the training set, while our method has better generality.

Compared with UG2 top 1-2 teams, the top 2 teams use very strong backbones, state-of-the-art enhancement methods, and advanced settings, *e.g*. different maximum testing scales. We instead use the original settings of DSFD.

2.6. Improving Supervised Model

For fine-tuning DSFD with labels, by combining with our adaptation, the mAP improves from 0.460 to 0.486.

2.7. Generalization

Our joint high-low adaptation can also be extended to other tasks. For example, the AP of COCO-pretrained [16] Faster-RCNN [21] on ExDark [17] is 29.261. By applying our joint high-low adaptation, the performance can be improved to 30.056 as shown in Table 3.

Table 3: Adapting Faster-RCNN from COCO to ExDark.

	AP	AP_{50}	AP_{75}
original	29.261	59.784	24.538
w/ Ours	**30.056**	**60.746**	**25.835**

References

[1] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE TIP*, 27(4):2049–2062, 2018. 3

[2] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020. 3

[3] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Lei, Stan Z. Li, and Xudong Zou. Selective refinement network for high performance face detection. In *AAAI*, 2019. 3

[4] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen O. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE TIP*, 16(8):2080–2095, 2007. 10

[5] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *CoRR*, abs/1905.00641, 2019. 3

[6] Antonio D'Innocente, Francesco Cappio Borlino, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. One-shot unsupervised cross-domain detection. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, 2020. 3

[7] Xueyang Fu, Delu Zeng, Yue Huang, Yinghao Liao, Xinghao Ding, and John W. Paisley. A fusion-based enhancing method for weakly illuminated images. *Signal Process.*, 129:82–96, 2016. 3

[8] 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 *CVPR*, 2020. 3

[9] Xiaojie Guo, Yu Li, and Haibin Ling. LIME: low-light image enhancement via illumination map estimation. *IEEE TIP*, 26(2):982–993, 2017. 3, 10

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1026–1034. IEEE Computer Society, 2015. 3

[12] Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Kumar Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *WACV*, 2020. 3

[13] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV*, 2018. 3

[14] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, 2018. 3

[15] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. DSFD: dual shot face detector. In *CVPR*, 2019. 1, 2, 3

[16] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014. 11

[17] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Comput. Vis. Image Underst.*, 178:30–42, 2019. 11

[18] Shi Luo, Xiongfei Li, Rui Zhu, and Xiaoli Zhang. SFA: small faces attention face detector. *IEEE Access*, 7:171609–171620, 2019. 3

[19] Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry S. Davis. SSH: single stage headless face detector. In *ICCV*, 2017. 3
[20] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, 2020. 3
[21] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *NeurIPS*, pages 91–99, 2015. 3, 11
[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 10
[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015. 1
[24] Xu Tang, Daniel K. Du, Zeqiang He, and Jingtuo Liu. Pyramidbox: A context-assisted single shot face detector. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV*, 2018. 3
[25] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *BMVC*, 2018. 3
[26] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. WIDER FACE: A face detection benchmark. In *CVPR*, 2016. 2
[27] Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J. Scheirer, Zhangyang Wang, Taiheng Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, Yuqiang Zheng, Yanyun Qu, Yuhong Xie, Liang Chen, Zhonghao Li, Chen Hong, Hao Jiang, Siyuan Yang, Yan Liu, Xiaochao Qu, Pengfei Wan, Shuai Zheng, Minhui Zhong, Taiyi Su, Lingzhi He, Yandong Guo, Yao Zhao, Zhenfeng Zhu, Jinxiu Liang, Jingwen Wang, Tianyi Chen, Yuhui Quan, Yong Xu, Bo Liu, Xin Liu, Qi Sun, Tingyu Lin, Xiaochuan Li, Feng Lu, Lin Gu, Shengdi Zhou, Cong Cao, Shifeng Zhang, Cheng Chi, Chubin Zhuang, Zhen Lei, Stan Z. Li, Shizheng Wang, Ruizhe Liu, Dong Yi, Zheming Zuo, Jianning Chi, Huan Wang, Kai Wang, Yixiu Liu, Xingyu Gao, Zhenyu Chen, Chang Guo, Yongzhou Li, Huicai Zhong, Jing Huang, Heng Guo, Jianfei Yang, Wenjuan Liao, Jiangang Yang, Liguo Zhou, Mingyue Feng, and Likun Qin. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE TIP*, 29:5737–5752, 2020. 2, 3, 10
[28] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Deep plug-and-play super-resolution for arbitrary blur kernels. In *CVPR*, 2019. 10
[29] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *ACM MM*, 2019. 3
[30] Zhishuai Zhang, Wei Shen, Siyuan Qiao, Yan Wang, Bo Wang, and Alan L. Yuille. Robust face detection via learning small faces on hard images. In *WACV*, 2020. 3
[31] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 3