arXiv:2104.01984v1 [cs.CV] 5 Apr 2021

HLA-Face: Joint High-Low Adaptation for Low Light Face Detection

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

Abstract

Face detection in low light scenarios is challenging but vital to many practical applications, e.g., surveillance video, autonomous driving at night. Most existing face detectors heavily rely on extensive annotations, while collecting data is time-consuming and laborious. To reduce the burden of building new datasets for low light conditions, we make full use of existing normal light data and explore how to adapt face detectors from normal light to low light. The challenge of this task is that the gap between normal and low light is too huge and complex for both pixel-level and object-level. Therefore, most existing low-light enhancement and adaptation methods do not achieve desirable performance. To address the issue, we propose a joint High-Low Adaptation (HLA) framework. Through a bidirectional low-level adaptation and multi-task high-level adaptation scheme, our HLA-Face outperforms state-of-the-art methods even without using dark face labels for training. Our project is publicly available at: https://daooshee.github.io/HLA-Face-Website/

1. Introduction

Face detection is fundamental for many vision tasks, and has been widely used in a variety of practical applications, such as intelligent surveillance for smart city, face unlock, and beauty filters in mobile phones. Over the past decades, extensive researches have made great progress in face detection. However, face detection under adverse illumination conditions is still challenging. Images captured without insufficient illumination suffer from a series of degradations, *e.g.*, low visibility, intensive noise, and color cast. These degradations can not only affect the human visual quality, but also worsen the performance of machine vision tasks, which may cause potential risks in surveillance video analysis and nighttime autonomous driving. In Fig. 1 (a), the state-of-the-art face detector DSFD [1] can hardly detect faces under insufficient illumination, in direct contrast to its over 90% precision on WIDER FACE [2].

(a) DSFD (b) LIME + DSFD (c) Ours

(d) Our learning paradigm for dark face detection

Figure 1. Dark face detection visual results and our learning paradigm. Compared with the result of DSFD [1] on original low light images and the enhanced version by LIME [5], our method can better recognize the faces in dark scenarios.

To promote the research of low light face detection, a large scale benchmark DARK FACE [3] is constructed. The emergence of dark face data gives birth to a number of dark face detection researches [4]. However, existing methods are dependent on extensive annotations, therefore have poor robustness and scalability.

In this paper, based on the benchmarking platform provided by DARK FACE, we explore how to adapt normal light face detection models to low light scenarios without the requirement of dark face annotations. We find that there are two levels of gaps between normal light and low light. One is the gap in pixel-level appearance, such as the insufficient illumination, camera noise, and color bias. The other is the object-level semantic differences between normal and low light scenes, including but not limited to the existence of street lights, vehicle headlights, and advertisement

*Corresponding author. This work was supported by the National Key Research and Development Program of China under Grant No. 2018AAA0102702 and the Fundamental Research Funds for the Central Universities and the National Natural Science Foundation of China under Contract No.61772043 and a research achievement of Key Laboratory of Science, Techonology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

boards. Traditional low light enhancement methods [5, 6] are designed for improving visual quality, therefore cannot fill the semantic gap, as shown in Fig. 1 (b). Typical adaptation methods [7, 8] are mainly designed for the scenario where the two domains share the same scene, such as adapting from Cityscapes [9] to Foggy Cityscapes [10]. But for our task, the domain gap is more huge, raising a more difficult challenge for adaptation.

To adapt from normal light to low light, we propose a High-Low Adaptation Face detection framework (HLA-Face). We consider joint low-level and high-level adaptation. Specifically, for low-level adaptation, typical methods either brighten the dark image or darken the bright image. However, due to the huge domain gap, they do not achieve desirable performance. Instead of unidirectional low-to-normal or normal-to-low translation, we bidirectionally make two domains each take a step towards each other. By brightening the low light images and distorting the normal light images, we build intermediate states that lie between the normal and low light. For high-level adaptation, we use multi-task self-supervised learning to close the feature distance between the intermediate states built by low-level adaptation. By combining low-level and high-level adaptation, we outperform state-of-the-art face detection methods even though we do not use the labels of dark faces. Our contributions are summarized as follows:

- We propose a framework for dark face detection without annotated dark data. Through a joint low-level and high-level adaptation, our model achieves superior performance compared with state-of-the-art face detection and adaptation methods.
- For low-level adaptation, we design a bidirectional scheme. Through brightening low light data and distorting normal light data with noise and color bias, we set up intermediate states and make two domains each take a step towards each other.
- For high-level adaptation, we introduce cross-domain self-supervised learning for feature adaptation. With context-based and contrastive learning, we comprehensively close the feature distance among multiple domains and further strengthen the representation.

2. Related Works

Low Light Enhancement. Low illumination is a common kind of visual distortion, which might be caused by undesirable shooting conditions, wrong camera operations, and equipment malfunctions, *etc*. There have been many literatures for low light enhancement. Histogram equalization and its variants [11] stretch the dynamic range of the images. Dehazing-based methods [12] regard dark images as inverted hazy images. Retinex theory assumes that images can be decomposed into illumination and reflectance. Based on the Retinex theory, a large portion of works [5, 13] estimate illumination and reflectance, then process each component separately or simultaneously. Recent methods are mainly based on deep learning. Some design end-to-end processing models [14], while some inject traditional ideas such as the Retinex theory [15, 16, 6]. Besides processing 8-bit RGB images, there are also models for RAW images [17].

The problem is that these methods are mainly designed for human vision rather than machine vision. How pixel-level adjustment can benefit and guide high-level tasks has not been well explored. In this paper, we provide corresponding solutions for dark face detection.

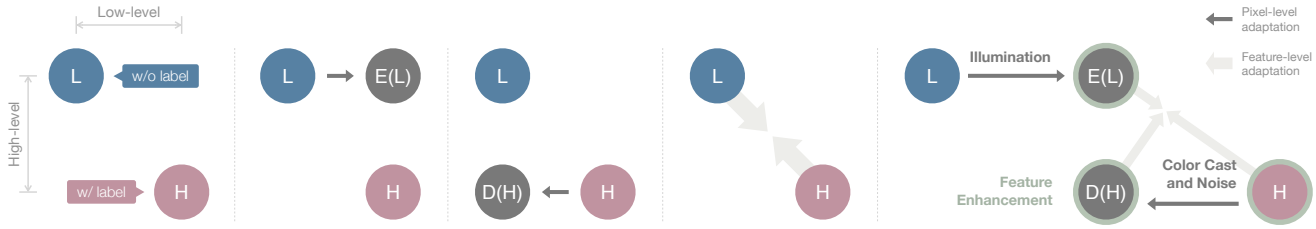
Face Detection. Early face detectors rely on hand-crafted features [18], which are now replaced by deep features learned from data-driven convolutional neural networks. Inherit from generic object detection, typical face detectors can be classified into two categories: two-stage and one-stage. Two-stage models [19, 20] first generate region proposals, then refine them for the final detection. One-stage models [21] instead directly predict the bounding boxes and confidence. The difference between generic object and face detection is that, in face detection, the scale variation is often much larger. Existing methods solve this problem by multi-scale image and feature pyramids [22, 23], or various anchor sampling and matching strategies [24, 25, 26].

Despite the prosperity of face detection researches, existing models seldom consider the scenario of insufficient illumination. In this paper, we propose a dark face detector that outperforms state-of-the-art methods even without using dark annotations.

Dark Object Detection. With the rapid development of deep learning, object detection has attracted more and more attention. However, few efforts have been made for dark objects. For RAW images, YOLO-in-the-Dark [27] merges models pre-trained in different domains using glue layers and a generative model. For RGB images, Loh *et al*. build the ExDark [28] dataset and analyze the low light images using both hand-crafted and learned features. DARK FACE [3] is a large-scale low light face dataset, giving birth to a series of dark face detectors in the UG^2 Prize Challenge[1]. However, most of these models highly rely on annotations, thus are of limited flexibility and robustness.

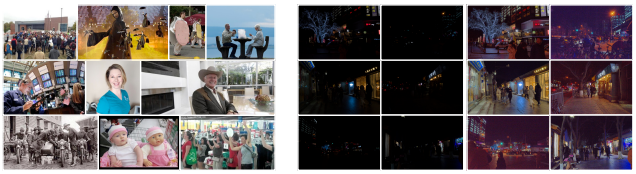
To get rid of the dependency on labels, Unsupervised Domain Adaptation (UDA) may be a plain solution [8, 29]. Although UDA has been demonstrated to be effective in many applications, due to the huge gap between normal and low light, these methods have limited performance in dark face detection. In this paper, we propose a superior method by combining low- and high-level adaptation.

[1] http://cvpr2020.ug2challenge.org/



(a) Problem description (b) Enhancement (c) Darkening (d) Feature adaptation (e) The proposed method

Figure 2. Comparison of different adaptive low light detection techniques. L: low light data. H: normal light data. Existing enhancement-based, darkening-based, and feature adaptation methods either ignore the high-level gap, or have limited effects due to the huge and complex gap between L and H. Our method instead considers both low-level and high-level adaptation, therefore achieves better performance.



(a) WIDER FACE (b) DARK FACE

Figure 3. Comparison of WIDER FACE and DARK FACE. On the right, DARK FACE is enhanced for better visibility.

3. Joint Adaptation for Dark Face Detection

In this section, we firstly introduce the motivation of our learning paradigm, then describe the detailed designs.

3.1. Motivation

The task is to adapt face detectors trained on normal light data H to unlabeled low light data L. As shown in Fig. 2, existing methods can be roughly divided into three categories: enhancement, darkening, and feature adaptation. **Enhancement**-based methods [30] brighten the low light images and directly test on them. They usually require no model fine-tuning, therefore are highly flexible. **Darkening**-based methods [31, 32, 33] first darken the normal light data into a dark version, then re-train the model on the transferred annotated data. Enhancement and darkening are all pixel-level. For **feature adaptation**, typical methods use alignment [34], adversarial learning [35], or pseudo labeling [29] to directly adapt the features of the model.

The problem for dark face detection is that the gap between H and L is too huge and complex for existing methods to handle. As shown in Fig. 3, the images in WIDER FACE [2] and DARK FACE [3] not only have different pixel-level appearance (bright v.s. dark, clean v.s. noisy), but also contain different objects and scenes (photos, paintings v.s. street views). However, enhancement- and darkening-based methods only consider the pixel-level gap. Feature adaptation methods try to fill the whole gap in one step. But as shown Sec. 4.2, the effect is limited.

To jointly fill both pixel-level and feature-level gaps for dark face detection, we propose a High-Low Adaptation (HLA) scheme. As shown in Fig. 2 (e), we set low-level intermediate states between L and H, and based on these states adapt the corresponding high-level representations. Specifically, the low-level distance is reduced by both *enhancing* and *darkening*. Compared with L-to-H or H-to-L unidirectional translation, our bidirectional translation: L-to-$E(L)$ and H-to-$D(H)$, can not only ease the difficulty of adaptation, but also provide more tools for feature-level adaptation. The high-level distance is reduced by pushing the feature spaces of multiple states towards each other. Moreover, the feature representation is further enhanced by contrastive learning. While testing, we first process the image by $E(\cdot)$, then apply the adapted face detector.

The framework detail is shown in Fig. 4. In the following, we will respectively introduce the proposed low-level and high-level adaptation schemes.

3.2. Bidirectional Low-Level Adaptation

The challenge of low-level adaptation lies in two aspects. One is the co-existence of the high-level gap, which can confuse pixel-level transfer models. For example, we show the effect of some methods for transferring H to L in Fig. 5. Different from WIDER FACE, DARK FACE contains many street lights, vehicle highlights, and signboards. Accordingly, CUT [36] generates weird lights on human bodies, and CycleGAN [37] generates street lights on faces. MUNIT [38] can distinguish content and style, therefore has no street light artifact. However, MUNIT cannot completely darken the image, and the result is visually far from L.

The other challenge is the difficulty of low light enhancement itself. Existing low light enhancement methods are mainly designed for human vision rather than machine vision. Some methods draw black edges, keep noisy parts dark, or enhance the contrast to improve the comprehensive visual quality, which can damage the high-level detection performance. Moreover, images in DARK FACE suffer from intensive noise and color bias. However, existing denoising and color reconstruction methods are not robust enough to handle this extreme case.

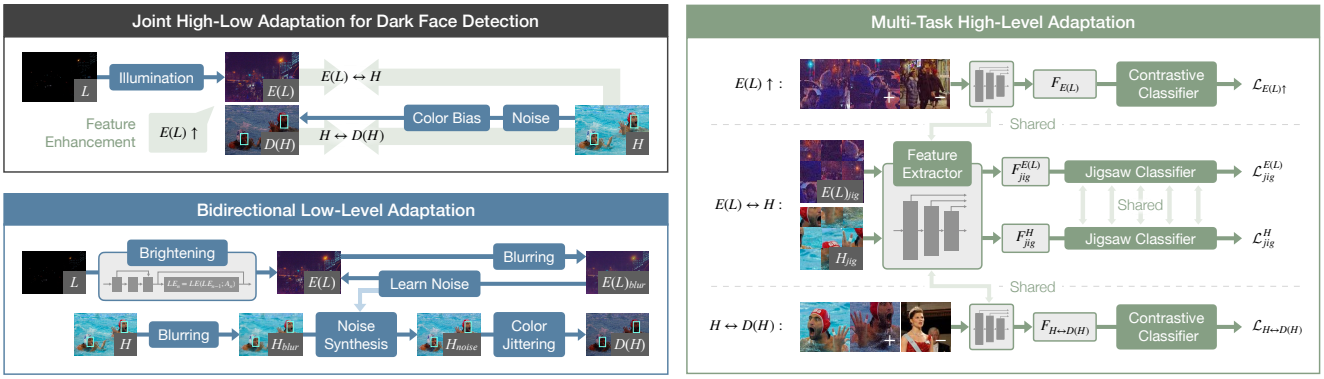


Figure 4. The overview of our joint High-Low Adaptation (HLA) framework for dark face detection. Low-level adaptation fills the gap by creating intermediate states. We bidirectionally brighten the low light data as well as distort the normal light data with noise and color bias. Based on the built intermediate states, we use multi-task cross-domain self-supervised learning to fill the high-level gap.

(a) WIDER FACE (b) DARK FACE (c) Enhanced (b)

(d) CycleGAN (e) Enhanced (d) (f) MUNIT

(g) CUT (h) Our H_{noise} (i) Our $E(L)$

Figure 5. Results of transferring between WIDER FACE and DARK FACE. (b) and (d) are enhanced for better visualization.

To solve the above challenges, we propose the bidirectional low-level adaptation scheme. Low light degradation is a complex process. We roughly decompose the related factors into three aspects: illumination, noise, and color bias. Although denoising and color correction is difficult, adding noise and applying color bias inversely is relatively easy. On this basis, we bright L into $E(L)$, and distort H with noises and color bias to form $D(H)$. Compared with L and H, $E(L)$ and $D(H)$ are more similar. In this way, we ease the difficulty of adaptation by making L and H take a step towards each other. Also, by formulating the specific components of low light degradation, the transfer model will not be disturbed by the semantic gap between the domains. In the following, we will introduce the detailed designs of each procedure.

Brightening. Different from common low light enhancement tasks, here we want to adjust the illumination without denoising or color reconstruction. Moreover, low light images suffer from nonuniform illumination. Some faces may be brightened by street lights, while some may be covered in severe darkness. Therefore, we also need to prevent over-exposure as well as under-exposure.

Our module is based on nonlinear curve mapping [14], which is made up of iterative quadratic curves $LE(\cdot)$:

$$LE(x, A) = x + Ax(1 - x), \quad (1)$$

$$LE_n = LE(LE_{n-1}; A_n), \quad (2)$$

where LE_0 is the input image, LE_n is the result at iteration n, and A_n is a pixel-wise three-channel adjustment map estimated by neural networks. Compared with common end-to-end or Retinex-based deep enhancement methods, curve mapping does not introduce extra noise or artifacts. We follow [14] to use a 7-layer CNN with symmetrical skip-connections and the corresponding training objectives.

The issue of [14] is that, the enhancement is conservative (Weak). As shown in Fig. 6 (b), many faces are still covered in darkness. This is because further enhancing the image can bring more noise, and [14] choose to hide these noises in darkness, so that the visual quality of the whole image is better. We instead propose strong illumination enhancement (Strong). By doubling the iteration number in Eq. (1) and widening the curve estimation network, the model can enhance the image with higher brightness. The drawback may be that noise and color bias come along, but we can leave it to the following $H \rightarrow D(H)$ process. This is also the difference between our enhancement module and common low light enhancement methods.

Noise Synthesis. Although the pixel-level distance can be reduced by brightening, the gap between $E(L)$ and H is still challenging. Therefore, we further decompose the gap

(a) Input (b) Weak (c) Strong

Figure 6. Effects of weak and strong brightening. Compared with (c), many faces are still covered in darkness in (b).

remained into color and noise. Also, by separating out the color, we can use color to guide the noise synthesis process.

As shown in Fig. 4, we first blur $E(L)$ by a strong Bilateral filter of $d = 25$ and $\sigma = 75$. The blurring result $E(L)_{blur}$ works as the color guidance. Then, a Pix2Pix [39] is trained for transferring from $E(L)_{blur}$ to $E(L)$. Finally, we blur H in the same way and use the trained Pix2Pix to add noise. As shown in Fig. 5, H_{noise} successfully imitates the noise pattern of $E(L)$. Their difference in color distribution will be handled in the next step.

Color Jittering. We want the color distribution of $D(H)$ to match that of $E(L)$. Based on statistical analysis, we set the jittering range to brightness: (0.4, 1.2), contrast: (0.6, 1.4), saturation: (0.6, 1.4) and hue: (0.8, 1.2).

3.3. Multi-Task High-Level Adaptation

Most feature adaptation methods are based on alignment, pseudo labeling, and adversarial learning. However, alignment and pseudo labeling cannot well handle the huge gaps, while adversarial learning is not stable. We instead fully use the natural information of the images themselves, *i.e.*, self-supervised learning. By forcing the self-supervised learning classifiers to be shared across domains, the features are forced to be mapped into the same high dimensional subspace, therefore closing the high-level gap.

To push $E(L)$, H and $D(H)$ towards each other, we first close $E(L)$-H by cross-domain context-based self-supervised learning, then close H-$D(H)$ by cross-domain contrastive learning. We further enhance the representation of $E(L)$ by single-domain contrastive learning. The whole adaptation works in a multi-task way. In the following, we will introduce the details of each learning scheme.

Closing E(L) and H. Context-based self-supervised learning designs pretext tasks, through which the model can learn to understand the spatial context of objects. Here, we use the jigsaw puzzling game [40]. We have also tried rotation [41] and combining jigsaw with rotation, but find that using jigsaw alone works the best. One possible explanation for this may be that many images in WIDER FACE are paintings or advertisements, where the faces may have strange angles. Therefore, the rotation prediction pretext task can be ambiguous.

Similar to [42], we assemble 3×3 patches into a whole image and set the patch permutation number to 30, *i.e.*, 30 classification problem. Denote p_{jig} as the permutation label, and $\mathcal{L}_c$ as the cross-entropy loss, we have:

$$\mathcal{L}_{jig}^{E(L)} = \mathcal{L}_c(F_{jig}^{E(L)}, p_{jig}^{E(L)}), \tag{3}$$

$$\mathcal{L}_{jig}^{H} = \mathcal{L}_c(F_{jig}^{H}, p_{jig}^{H}), \tag{4}$$

where F_{jig} stands for the feature extracted from the corresponding domain. $E(L)$ and H share classification heads, which can force the semantic features to be mapped into the same space, therefore closing high-level gaps. The final loss for closing $E(L)$ and H is:

$$\mathcal{L}_{E(L)\leftrightarrow H} = \mathcal{L}_{jig}^{E(L)} + \mathcal{L}_{jig}^{H}. \tag{5}$$

Closing H and D(H). The idea of contrastive learning is that, given a query v, identifying its "positive" pair v^+ and "negatives" pairs $v^- = \{v_1^-, v_2^- ..., v_N^-\}$. With similarity measured by dot product, the objective $\mathcal{L}_q(v, v^+, v^-)$ is:

$$\mathcal{L}_q = -\log\left[\frac{\sigma(v, v^+)}{\sigma(v, v^+) + \sum_{n=1}^{N}\sigma(v, v_n^-)}\right], \tag{6}$$

$$\sigma(x, y) = \exp(x \cdot y/\tau), \tag{7}$$

where τ is a temperature hyper-parameter. Intuitively, this is an $(N + 1)$ classification problem.

To reduce the distance between H and $D(H)$, we take advantage of the behavior that contrastive learning brings positive samples closer. In specific, we make the positive pair of H to be the patch from $D(H)$, and vice versa:

$$\begin{aligned}\tilde{\mathcal{L}}_{H\leftrightarrow D(H)} = &\mathcal{L}_q(H, D(H)^+, H^-) \\ &+ \mathcal{L}_q(D(H), H^+, D(H)^-).\end{aligned} \tag{8}$$

In this way, the feature similarity between H and $D(H)$ can be improved, and the high-level gap can be closed.

We also introduce single-domain contrastive learning on H and $D(H)$ themselves to make the features better. In the implementation, the above four losses are simplified by regarding $D(\cdot)$ as a part of the augmentation:

$$\mathcal{L}_{H\leftrightarrow D(H)} = \mathcal{L}_q(D_i^*(H), D_j^*(H)^+, D_k^*(H)^-), \tag{9}$$

where $D^*(H)$ has a 50% probability of being H, and 50% of being $D(H)$. While training, we use the Momentum Contrast (MoCo) [43] and follow [44] for other settings.

Enhancing E(L). We also find that it is beneficial to enhance the feature on $E(L)$ by contrastive learning:

$$\mathcal{L}_{E(L)\uparrow} = \mathcal{L}_q(E(L), E(L)^+, E(L)^-). \tag{10}$$

Final objective. Our model learns in a multi-task way. Denote $\mathcal{L}_{det}$ as the detection loss, the final objective is:

$$\mathcal{L} = \lambda_{det}\mathcal{L}_{det} + \lambda_{E(L)\leftrightarrow H}\mathcal{L}_{E(L)\leftrightarrow H} + \lambda_{H\leftrightarrow D(H)}\mathcal{L}_{H\leftrightarrow D(H)} + \lambda_{E(L)\uparrow}\mathcal{L}_{E(L)\uparrow}, \quad (11)$$

where λs are hyper-parameters to balance different losses.

4. Experimental Results

4.1. Implementation Details

Network Architecture. DSFD [1] is used as the face detection baseline. Our headers for self-supervised learning are added on the conv3_3, conv4_3, conv5_3, conv_fc7, conv6_2, and conv7_2 layers of the backbone. For more details, please refer to the supplementary material.

Training and Evaluation Settings. All experiments are based on WIDER FACE [2] and DARK FACE [3]. Our model is allowed to use the labels of WIDER FACE, but not allowed to use the labels of DARK FACE. The framework is first pre-trained on WIDER FACE, then fine-tuned with both WIDER FACE and the images of DARK FACE. Pre-training follows the same process of [1]. For fine-tuning, the batch size is set to 8. 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. The testing process is the same as the original DSFD implementation.

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. Performance is measured by mean Average Precision (mAP), and evaluated with the official tool[2] of DARK FACE.

4.2. Comparisons with State-of-the-Art Methods

The proposed model is compared with 22 state-of-the-art methods, covering the categories of face detection, low light enhancement, image-to-image translation, and unsupervised domain adaptation. The benchmarking results are shown in Table 1 and Fig. 8.

Face Detection. Our model is compared with seven face detectors and one generic object detector. Due to the poor visibility caused by low light conditions, existing detectors all achieve undesirable performance. As shown in Table 1, Faster-RCNN[3] [45] (re-trained on WIDER FACE) performs worse than detection models designed especially for faces. State-of-the-art face detection methods, SSH [46], RetinaFace [47], SRN [48], SFA [49], and PyramidBox [50] all have mAP scores less then 15%, showing that insufficient illumination can greatly hurt the performance of high-level tasks. The best results here are achieved by DSFD [1] and Small Hard Faces [51], but their mAP scores are still unsatisfactory. By adapting to the dark environment, our model outperforms these detectors by a significant margin.

Table 1. Comparison results on DARK FACE.

Category	Method	mAP (%)
Face Detection	Faster-RCNN [45]	1.7
	SSH [46]	6.9
	RetinaFace [47]	8.6
	SRN [48]	9.0
	SFA [49]	9.3
	PyramidBox [50]	12.5
	Small Hard Face [51]	16.1
	DSFD [1]	16.1
Enhancement (with Small Hard Face)	Zero-DCE [14]	37.7
	MF [13]	38.3
Enhancement (with DSFD)	SICE [15]	4.7
	RetinexNet [16]	12.0
	KinD [6]	15.8
	EnlightenGAN † [52]	20.8
	EnlightenGAN [52]	31.3
	Zero-DCE † [14]	37.3
	LIME [5]	40.7
	Zero-DCE [14]	41.3
	MF [13]	41.4
Darkening (with DSFD)	MUNIT [38]	29.7
	CycleGAN [37]	31.9
	CUT [36]	32.7
Unsupervised DA (with DSFD)	OSHOT [53]	25.4
	Progressive DA [54]	28.5
	Pseudo Labeling [29]	35.1
Fully Supervised	Fine-tuned DSFD [1]	46.0
	Ours	**44.4**

† denotes retrained with DARKFACE.

Enhancement. We also explore the effect of illumination adjustment, *i.e.*, the scheme in Fig. 2 (b). We first use low light enhancement methods to enhance the DARK FACE images, then apply the face detectors. Although DSFD and Small Hard Face are comparable on original dark images, when the images are brightened, DSFD outperforms Small Hard Face by 3.35% in mAP on average. This indicates that DSFD is of better robustness and generalization. Therefore, in the rest of the experiments, we use DSFD as the baseline.

Although some low light enhancement methods can improve the performance to a large extent, some may even damage the detection performance. This is because these methods introduce more artifacts to the images. As shown in Fig. 7, SICE [15] distorts the details. KinD [6] over-

[2]https://github.com/Ir1d/DARKFACE_eval_tools
[3]https://github.com/playerkk/face-py-faster-rcnn

(a) Ground truth (b) SICE (c) RetinexNet (d) KinD (e) LIME (f) Zero-DCE (g) MF (h) Ours

Figure 7. 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.

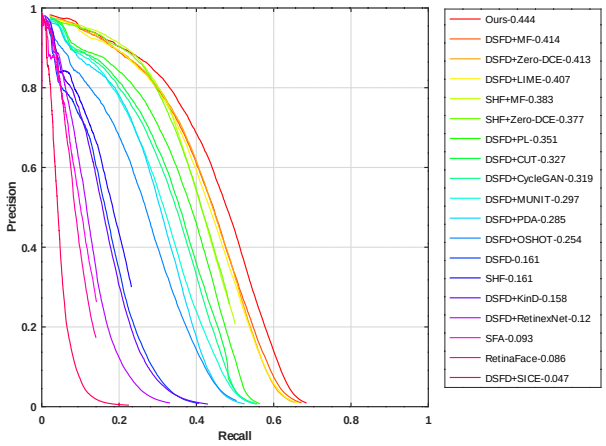


Figure 8. Precision-Recall (PR) curves on DARK FACE.

Table 2. Comparison with Top 10 teams (with labels) in the UG^2 Prize Challenge. Scores are copied from the official website.

Rank	Team Name	mAP (%)
1	CAS-Newcastle-TUM	62.45
2	CAS-NEU	61.84
-	**Ours**	**44.44**
3	MSFace	42.71
4	iie	40.49
5	NTU-MiRA	37.50
6	DUTMedia	35.65
7	SCUT-CVC	35.18
8	IIAI VOS	34.73
9	USTC-NELSLIP	32.81
10	PHI-AI	29.95

denoises the images, leading to blurry edges and dull color. RetinexNet [5] instead generates weird green colors on dark regions. These three methods widen the gap between the testing images and the daytime natural photographies, therefore hurt the performance of the face detector. MF [13] and Zero-DCE [14], LIME [5] can help DSFD better recognize faces. The visual quality of their subjective results is also better. However, compared with our model, their performance is still relatively poor. This is because when simply combining low light enhancement and face detection, the semantic gap between WIDER FACE and DARK FACE still remains.

Darkening. Darkening-based adaptation, *i.e.*, the scheme in Fig. 2 (c), proposes to re-train models on synthetic dark data. Specifically, we first transfer WIDER FACE to DARK FACE, then use the transferred WIDER FACE to re-train DSFD. Most darkening-based methods [32, 55] are based on the classic unsupervised image-to-image translation model CycleGAN [37]. We also test more powerful MUNIT [38], and the newest CUT [36].

Quantitative and qualitative results can be found in Fig. 5 and Table 1, respectively. Although MUNIT is more powerful than CycleGAN, the effect of benefitting dark face detection is worse. This is because MUNIT cannot fully darken the image as shown in Fig. 5 (f). In Fig. 5 (d) and (e), although the result of CycleGAN looks like night street views at the first glance, after enhancing the image, we can see that CycleGAN actually distorts the details and puts street light on faces. CUT generates less artifact than CycleGAN, therefore the mAP score is higher. However, compared with our model, the performances of darkening-based methods are all unsatisfactory. This demonstrates our assumption that the gap between normal and low light is too huge and complex for pixel-level transfer models to handle.

Unsupervised Domain Adaptation. Most UDA methods are based on Faster-RCNN, which performs too poor on face detection as shown in Table 1. For a fair comparison, we re-implement all compared UDA methods with DSFD.

OSHOT [53] directly closes the gap by self-supervised learning of rotation angle prediction. It is originally designed for one-shot adaptation. We change it into fine-tuning on the whole DARK FACE. The performance of OSHOT is poor. This is because the gap between normal light and low light faces is too huge to be handled by feature adaptation. Pseudo Labeling [29] is a two-step progressive UDA method. It first uses CycleGAN to artificially generate training data, then uses pseudo labels to fine-tune the detector. Compared with directly training on images synthesized by CycleGAN, the performance improves from 31.9% to 35.1% in mAP, demonstrating the effectiveness of pseudo labels. However, the mAP is still less than 40%. Progressive

Table 3. Ablation study results on DARK FACE. † denotes using the pyramid multi-scale testing scheme in DSFD.

$E(\cdot)$	$E(L) \leftrightarrow H$	$H \leftrightarrow D(H)$	$E(L) \uparrow$	mAP (%)
-	-	-	-	15.3
Weak	-	-	-	38.3
Strong	-	-	-	39.1
-	Rotation	-	-	22.7
-	Jigsaw	-	-	26.9
-	Rot + Jig	-	-	25.3
Strong	-	Pseudo labels	-	40.2
Strong	-	H only	-	38.2
Strong	-	Cross-domain	-	40.9
Strong	Jigsaw	-	-	40.2
Strong	-	Cross-domain	✓	41.1
Strong	Jigsaw	Cross-domain	✓	41.4
Strong	Jigsaw	Cross-domain	✓	44.4 †

DA [54] combines pixel-level transferring and feature-level adversarial learning. But adversarial learning still cannot close the huge gap between normal and low light domains.

With Dark Annotations. Our model is also compared with face detection methods that have access to the labels of DARK FACE. The result of fine-tuning DSFD with labels is shown in Table 1. We can see that our model is much closer to the supervised learning upper bound 46.0% in mAP, demonstrating the effectiveness of our adaptation framework. We also show the leader board of the UG^2 Prize Challenge[4] in Table 2, where our model outperforms most of the teams. Notice that the teams in UG^2 are allowed to use labels for training, while our model uses no DARK FACE annotations.

4.3. Ablation Studies

To support our motivation and the joint high-low adaptation framework, in this section, we analyze the effect of each technical design. The results are shown in Table 3.

Effectiveness of E(L). Enhancing the testing images can improve the performance from 15.3% to 39.1% in mAP. Compared with the baseline (Weak), the performance of our $E(L)$ is higher by 0.8%, supporting our proposed strong enhancement.

Effectiveness of E(L) $\leftrightarrow$ H. We show the effect of different choices of context-based self-supervised learning for closing $E(L)$ and H. Using jigsaw alone works the best. Adding rotation can damage the performance. As we mentioned in Sec. 3.3, since many images in WIDER FACE are paintings or advertisements, the rotation angle prediction pretext task can be ambiguous. As shown in Table 4, the rotation top-1 classification accuracy on WIDER FACE is only slightly over random guess. In comparison, although the jigsaw pretext task is a 30-class problem, the top-1 accuracies are higher than 85%. We also notice that for both jigsaw and rotation, the performance on DARK FACE is higher than that on WIDER FACE. This is because the images in WIDER FACE are more diverse.

[4] http://cvpr2020.ug2challenge.org/program19/leaderboard19_t2.html

Table 4. Top-1 classification accuracy of jigsaw and rotation self-supervised learning pretext tasks in different domains.

Layer	Jig, $E(L)$	Jig, H	Rot, $E(L)$	Rot, H
conv3_3	97.5%	80.6%	19.2%	15.4%
conv4_3	98.9%	87.8%	33.6%	26.2%
conv5_3	99.0%	89.6%	51.2%	36.7%
conv_fc7	99.3%	89.7%	54.9%	33.5%
conv6_2	99.3%	89.7%	51.2%	28.3%
conv7_2	99.3%	90.0%	41.8%	19.3%
average	98.9%	87.9%	42.0%	26.6%

Effectiveness of H $\leftrightarrow$ D(H). We further explore the strategy for closing H and $D(H)$. The proposed cross-domain contrastive learning scheme $\mathcal{L}_{H \leftrightarrow D(H)}$ can improve the mAP score by 1.8%. If we only use contrastive learning on H, the performance even drops from 39.1% to 38.2%. This is because if we enhance features only in H, the detection model will concentrate more on H, therefore increasing the distance between H and $E(L)$. The result also supports our design of cross-domain contrastive learning and the necessity of setting the intermediate domain $D(H)$.

Contrastive learning can be regarded as a kind of "soft" label. Naturally, we wonder about the effect of "hard" label. We test the result of directly training with transferred labels on $D(H)$, *i.e.*, pseudo labeling. The mAP can be improved from 39.1% to 40.2%, but the improvement is smaller than using our contrastive learning. This is because representation learning based on "soft" labels can avoid the inaccuracy of manual annotations and better refine the features.

Combination Effect. Finally, we demonstrate the combination effect of our design. Enhancing the feature on $E(L)$ ($\mathcal{L}_{E(L)\uparrow}$) can further improve the mAP score. The full version of our model achieves the best performance, demonstrating the effectiveness of our joint high-level and low-level adaptation framework.

DSFD uses a pyramid multi-scale testing scheme. Although it can improve the performance, the running time increases from 1.25 hours to 10 hours. Even without this multi-scale scheme, *i.e.*, using a more weak face detection baseline, the performance of our model (41.4% in mAP) can still outperform most of the state-to-the-arts in Table 1.

5. Conclusion

We design a joint high-level and low-level adaptation framework for dark face detection. We propose a bidirectional pixel translation pipeline for the low level, and a multi-task adaptation strategy based on self-supervised learning for the high level. Our framework demonstrates the potential of joint high-low adaptation and can inspire other related low light high-level vision tasks.

References

[1] 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, 6

[2] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. WIDER FACE: A face detection benchmark. In *CVPR*, 2016. 1, 3, 6

[3] 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. 1, 2, 3, 6

[4] Jinxiu Liang, Jingwen Wang, Yuhui Quan, Tianyi Chen, Jiaying Liu, Haibin Ling, and Yong Xu. Recurrent exposure generation for low-light face detection. *CoRR*, abs/2007.10963, 2020. 1

[5] Xiaojie Guo, Yu Li, and Haibin Ling. LIME: low-light image enhancement via illumination map estimation. *IEEE TIP*, 26(2):982–993, 2017. 1, 2, 6, 7

[6] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *ACM MM*, 2019. 2, 6

[7] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, 2019. 2

[8] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *CVPR*, 2019. 2

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2

[10] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 126(9):973–992, 2018. 2

[11] S. M. Pizer, R. E. Johnston, J. P. Ericksen, B. C. Yankaskas, and K. E. Muller. Contrast-limited adaptive histogram equalization: speed and effectiveness. In *Conference on Visualization in Biomedical Computing*, pages 337–345, 1990. 2

[12] Xuan Dong, Guan Wang, Yi Pang, Weixin Li, Jiangtao Wen, Wei Meng, and Yao Lu. Fast efficient algorithm for enhancement of low lighting video. In *ICME*, 2011. 2

[13] 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. 2, 6, 7

[14] 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. 2, 4, 6, 7

[15] 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. 2, 6

[16] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *BMVC*, 2018. 2, 6

[17] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, 2018. 2

[18] Paul A. Viola and Michael J. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004. 2

[19] 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*, 2015. 2

[20] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE TPAMI*, 39(6):1137–1149, 2017. 2

[21] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 2

[22] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *CVPR*, 2017. 2

[23] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 2

[24] Chenchen Zhu, Ran Tao, Khoa Luu, and Marios Savvides. Seeing small faces from robust anchor's perspective. In *CVPR*, 2018. 2

[25] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. S^3fd: Single shot scale-invariant face detector. In *ICCV*, 2017. 2

[26] Xiang Ming, Fangyun Wei, Ting Zhang, Dong Chen, and Fang Wen. Group sampling for scale invariant face detection. In *CVPR*, 2019. 2

[27] Yukihiro Sasagawa and Hajime Nagahara. Yolo in the dark - domain adaptation method for merging multiple models -. *ECCV*, 2020. 2

[28] 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. 2

[29] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, 2018. 2, 3, 6, 7

[30] Se Woon Cho, Na Rae Baek, Ja Hyung Koo, Muhammad Arsalan, and Kang Ryoung Park. Semantic segmentation with low light images by modified cyclegan-based image enhancement. *IEEE Access*, 8:93561–93585, 2020. 3

[31] Hongjun Lee, Moonsoo Ra, and Whoi-Yul Kim. Nighttime data augmentation using GAN for improving blind-spot detection. *IEEE Access*, 8:48049–48059, 2020. 3

[32] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *ICCV*, pages 7373–7382. IEEE, 2019. 3, 7

[33] Vinicius F. Arruda, Thiago M. Paixão, Rodrigo Ferreira Berriel, Alberto F. De Souza, Claudine Badue, Nicu Sebe, and Thiago Oliveira-Santos. Cross-domain car detection using unsupervised image-to-image translation: From day to night. In *IJCNN*, 2019. 3

[34] Baochen Sun and Kate Saenko. Deep CORAL: correlation alignment for deep domain adaptation. In Gang Hua and Hervé Jégou, editors, *ECCV*, 2016. 3

[35] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17:59:1–59:35, 2016. 3

[36] 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, 6, 7

[37] 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, 6, 7

[38] 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, 6, 7

[39] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 5

[40] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *ECCV*, 2016. 5

[41] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 5

[42] Fabio Maria Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, 2019. 5

[43] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 5

[44] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020. 5

[45] 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. 6

[46] Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry S. Davis. SSH: single stage headless face detector. In *ICCV*, 2017. 6

[47] 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. 6

[48] 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. 6

[49] Shi Luo, Xiongfei Li, Rui Zhu, and Xiaoli Zhang. SFA: small faces attention face detector. *IEEE Access*, 7:171609–171620, 2019. 6

[50] 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. 6

[51] 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. 6

[52] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE TIP*, 30:2340–2349, 2021. 6

[53] 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. 6, 7

[54] 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. 6, 8

[55] Tong Liu, Zhaowei Chen, Yi Yang, Zehao Wu, and Haowei Li. Lane detection in low-light conditions using an efficient data enhancement : Light conditions style transfer. *CoRR*, abs/2002.01177, 2020. 7