# Unsupervised Domain-Specific Deblurring via Disentangled Representations

Boyu Lu          Jun-Cheng Chen          Rama Chellappa

UMIACS, University of Maryland, College Park

bylu@umiacs.umd.edu          pullpull@cs.umd.edu          rama@umiacs.umd.edu

## Abstract

*Image deblurring aims to restore the latent sharp images from the corresponding blurred ones. In this paper, we present an unsupervised method for domain-specific, single-image deblurring based on disentangled representations. The disentanglement is achieved by splitting the content and blur features in a blurred image using content encoders and blur encoders. We enforce a KL divergence loss to regularize the distribution range of extracted blur attributes such that little content information is contained. Meanwhile, to handle the unpaired training data, a blurring branch and the cycle-consistency loss are added to guarantee that the content structures of the deblurred results match the original images. We also add an adversarial loss on deblurred results to generate visually realistic images and a perceptual loss to further mitigate the artifacts. We perform extensive experiments on the tasks of face and text deblurring using both synthetic datasets and real images, and achieve improved results compared to recent state-of-the-art deblurring methods.*

## 1. Introduction

Image blur is an important factor that adversely affects the quality of images and thus significantly degrades the performances of many computer vision applications, such as object detection [15] and face recognition [22, 21]. To address this problem, blind image deblurring methods aim to restore the latent sharp image from a blurred image. Conventional approaches often formulate the image deblurring task as a blur kernel estimation problem. Since this problem is highly ill-posed, many priors have been proposed to model images and kernels [29, 41, 13]. However, most of these priors only perform well on generic natural images, but cannot generalize to specific image domains, like face [36], text [9] and low-illumination images [10]. Therefore, some priors (*e.g.* $L_0$-regularized intensity and gradient prior [28], face exemplars [27]) have been developed to handle these domain-specific image deblurring problems. But still these methods can only handle certain types of blur and



(a) Blurred          (b) Madam *et al*. [25]          (c) Ours

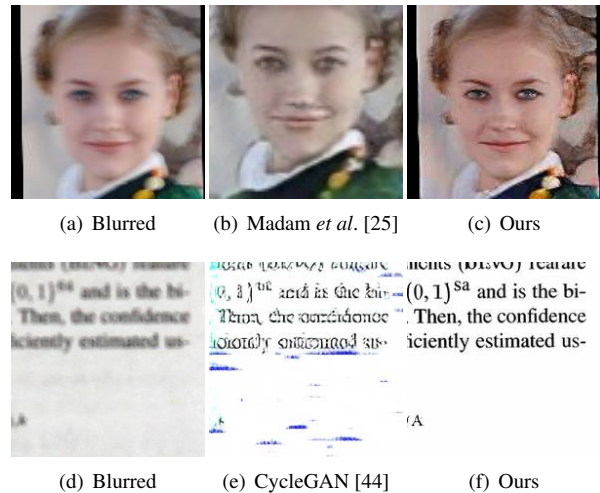(d) Blurred          (e) CycleGAN [44]          (f) Ours

Figure 1. Real-world blurred face and text results of the proposed method compared with other state-of-the-art unpaired deblurring methods. The deblurred image of (b) is from [25]. For (e), we apply our trained model using the publicly available code of [44].

often require longer inference time.

Recently, some learning-based approaches have been proposed for blind image deblurring [15, 26, 36]. CNN-based models can tackle more complex blur types and the inference is fast due to GPU acceleration. Meanwhile, the Generative Adversarial Networks (GAN) have been found to be effective in generating more realistic images. Nonetheless, most of these methods need paired training data, which is expensive to collect in practice. Although numerous blur generation methods have been developed [15, 38, 4], they are not capable of learning all types of blur variants in the wild. Moreover, strong supervision may cause algorithms to overfit training data and thus cannot generalize well to real images.

More recently, Nimisha *et al*. [25] proposed an unsupervised image deblurring method based on GANs where they add reblur loss and multi-scale gradient loss on the model. Although they achieved good performance on the synthetic datasets, their results on some real blurred images are not satisfactory (Fig. 1(b)). Another solution might be
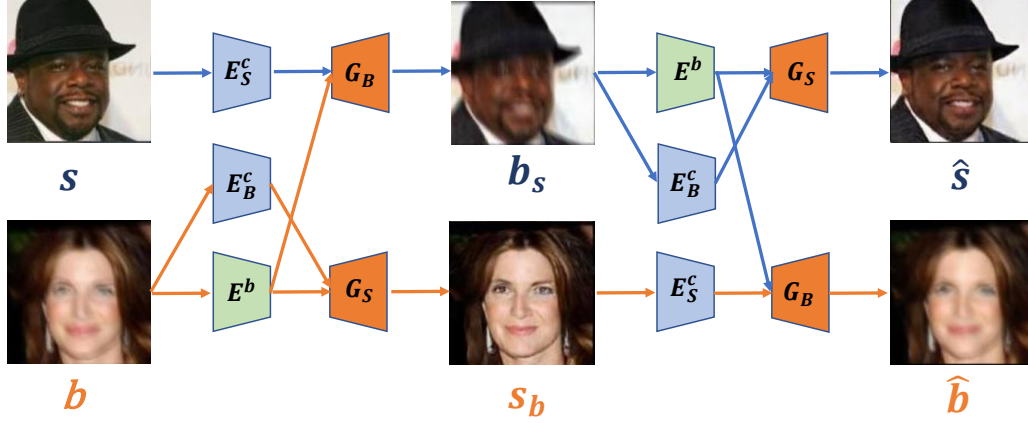
IEEE
computer
society

Figure 2. Overview of the deblurring framework. The data flow of the top *blurring* branch (bottom *deblurring* branch) is represented by blue (orange) arrows. $E_B^c$ and $E_S^c$ are content encoders for blurred and sharp images; $E^b$ is the blur encoder; $G_B$ and $G_S$ are blurred image and sharp image generators. Two GAN losses are added to distinguish $b_s$ from blur images, and to distinguish $s_b$ from sharp images. The KL divergence loss is added to the output of $E^b$. Cycle-consistency loss is added to $s$ and $\hat{s}$, $b$ and $\hat{b}$. Perceptual loss is added to $b$ and $s_b$.

directly using some existing unsupervised methods (Cycle-GAN [44], DualGAN [42]) to learn the mappings between sharp and blurred image domains. However, these generic approaches often encode other factors (*e.g.*, color, texture) rather than blur information into the generators, and thus do not produce good deblurred images (Fig. 1(e)).

In this paper, we propose an unsupervised domain-specific image deblurring method based on disentangled representations. More specifically, we disentangle the content and blur features from blurred images to accurately encode blur information into the deblurring framework. As shown in Fig. 2, the content encoders extract content features from unpaired sharp and blurred images, and the blur encoder captures blur information. We share the weights of the last layer of these two content encoders so that the content encoders can project the content features of both domains onto a common space. However, this structure by itself does not guarantee that the blur encoder captures blur features — it may encode content or other features as well. Inspired by [2], we add a KL divergence loss to regularize the distribution of blur features to suppress the contained content information. Then, the deblurring generator $G_S$ and the blurring generator $G_B$ take corresponding content features conditioned on blur attributes to generate deblurred and blurred images. Similar to CycleGAN [44], we also use the adversarial loss and the cycle-consistency loss as regularizers to assist the generator networks to yield more realistic images, and also preserve the content of the original image. To further remove the unpleasant artifacts introduced by the deblurring generator $G_S$, we add the perceptual loss to the proposed approach. Some sample deblurred images are shown in Fig. 1.

We conduct extensive experiments on face and text de-blurring and achieve competitive performance compared with other state-of-the-art deblurring methods. We also evaluate the proposed method on face verification and optical character recognition (OCR) tasks to demonstrate the effectiveness of our algorithm on recovering semantic information.

## 2. Related Works

### 2.1. Single Image Blind Deblurring

**Generic methods** Single image blind deblurring is a highly ill-posed problem. Over the past decade, various natural image and kernel priors have been developed to regularize the solution space of the latent sharp images, including heavy-tailed gradient prior [35], sparse kernel prior [7], $l_0$ gradient prior [41], normalized sparsity prior [14], and dark channels [29]. However, these priors are estimated from limited observations, and are not accurate enough. As a result, the deblurred images are often under-deblurred (images are still blurred) or over-deblurred (images contain many artifacts).

On the other hand, due to the recent immense success of deep networks and GANs, several CNN-based methods have been proposed for image deblurring. Sun *et al*. [38] and Schuler *et al*. [33] use CNN to predict the motion blur kernels. Chakrabarti *et al*. [4] predict the Fourier coefficients of the deconvolution filters by a neural network and perform deblurring in frequency domains. These methods combine the advantage of CNN and conventional maximum a posteriori probability (MAP)-based algorithms. Differently, Nah *et al*. [26] train a multi-scale CNN in an end-to-end manner to directly deblur images without explicitly estimating the blur kernel. Similarly, Kupyn *et al*. [15] use

WGAN and perceptual loss and achieve state-of-the-art performance on natural image deblurring.

**Domain specific methods** Although the above mentioned methods perform well for natural image deblurring, it is difficult to generalize them to some specific image domains, such as face and text images. Pan *et al.* [28] propose the $L_0$-regularized prior on image intensity and gradients for text image deblurring. Hradis *et al.* [9] train an end-to-end CNN specific for text image deblurring. Pan *et al.* [27] utilize exemplar faces in a reference set to guide the blur kernel estimation. Shen *et al.* [36] use the face parsing labels as global semantic priors and local structure regularization to improve face deblurring performance.

## 2.2. Disentangled Representation

There has been many recent efforts on learning disentangled representations. Tran *et al.* [40] propose DR-GAN to disentangle the pose and identity components for pose-invariant face recognition. Bao *et al.* [2] explicitly disentangle identity features and attributes to learn an open-set face synthesizing model. Liu *et al.* [19] construct an identity distill and dispelling auto encoder to disentangle identity with other attributes. BicycleGAN [45] combines cVAE-GAN and cLR-GAN to model the distribution of possible outputs in image-to-image translation. Recently, some unsupervised methods decouple images into domain-invariant content features and domain-specific attribute vectors, which produce diverse image-to-image translation outputs [17, 1, 11].

## 3. Proposed Method

The proposed approach consists of four parts: 1) content encoders $E_B^c$ and $E_S^c$ for blurred and sharp image domains; 2) blur encoder $E^b$; 3) blurred and sharp image generators $G_B$ and $G_S$; 4) blurred and sharp image discriminators $D_B$ and $D_S$. Given a training sample $b \in B$ in the blurred image domain and $s \in S$ in the sharp image domain, the content encoders $E_B^c$ and $E_S^c$ extract content information from corresponding samples and $E^b$ estimates the blur information from $b$. $G_S$ then takes $E_B^c(b)$ and $E^b(b)$ to generate a sharp image $s_b$ while $G_B$ takes $E_S^c(s)$ and $E^b(b)$ to generate a blurred image $b_s$. The discriminators $D_B$ and $D_S$ distinguish between the real and generated examples. The end-to-end architecture is illustrated in Fig. 2.

In the following subsections, we first introduce the method to disentangle content and blur components in Section 3.1. Then, we discuss the loss functions used in our approach. In Section 3.5, we describe the testing procedure of the proposed framework. Finally, the implementation details are discussed in Section 3.6.

## 3.1. Disentanglement of Content and Blur

Since the ground truth sharp images are not available in the unpaired setting, it is not trivial to disentangle the content information from a blurred image. However, since sharp images only contain content components without any blur information, the content encoder $E_S^c$ should be a good content extractor. We enforce the last layer of $E_B^c$ and $E_S^c$ to share weights so as to guide $E_B^c$ to learn how to effectively extract content information from blurred images.

On the other hand, the blur encoder $E^b$ should only encode blur information. To achieve this goal, we propose two methods to help $E^b$ suppress as much content information as possible. First, we feed $E^b(b)$ together with $E_S^c(s)$ into $G_B$ to generate $b_s$. Since $b_s$ is a blurred version of $s$ and it will not contain content information of $b$, this structure discourages $E^b(b)$ to encode content information of $b$. Second, we add a KL divergence loss to regularize the distribution of the blur features $z_b = E^b(b)$ to be close to the normal distribution $p(z) \sim N(0,1)$. As shown in [2], this will further suppress the content information contained in $z_b$. The KL divergence loss is defined as follows:

$$KL(q(z_b)||p(z)) = -\int q(z_b) \log \frac{p(z)}{q(z_b)} dz \qquad (1)$$

As proved in [13], minimizing the KL divergence is equivalent to minimizing the following loss:

$$\mathcal{L}_{KL} = \frac{1}{2} \sum_{i=1}^{N} (\mu_i^2 + \sigma_i^2 - \log(\sigma_i^2) - 1) \qquad (2)$$

where $\mu$ and $\sigma$ are the mean and standard deviation of $z_b$ and $N$ is the dimension of $z_b$. Similar to [13], $z_b$ is sampled as $z_b = \mu + z \circ \sigma$, where $p(z) \sim N(0,1)$ and $\circ$ represents element-wise multiplication.

## 3.2. Adversarial Loss

In order to make the generated images look more realistic, we apply the adversarial loss on both domains. For the sharp image domain, we define the adversarial loss as:

$$\mathcal{L}_{D_S} = \mathbb{E}_{s \sim p(s)}[\log D_S(s)] \\ + \mathbb{E}_{b \sim p(b)}[\log(1 - D_S(G_S(E_B^c(b), z_b)))] \qquad (3)$$

where $D_S$ tries to maximize the objective function to distinguish between deblurred images and real sharp images. In contrast, $G_S$ aims to minimize the loss to make deblurred images look similar to real samples in domain $S$. Similarly, we define the adversarial loss in blurred image domain as $\mathcal{L}_{D_B}$:

$$\mathcal{L}_{D_B} = \mathbb{E}_{b \sim p(b)}[\log D_B(b)] \\ + \mathbb{E}_{s \sim p(s)}[\log(1 - D_B(G_B(E_S^c(s), z_b)))] \qquad (4)$$

10219

### 3.3. Cycle-Consistency Loss

After competing with discriminator $D_S$ in the minimax game, $G_S$ should be able to generate visually realistic sharp images. However, since no pairwise supervision is provided, the deblurred image may not retain the content information in the original blurred image. Inspired by CycleGAN [44], we introduce the cycle-consistency loss to guarantee that the deblurred image $s_b$ can be reblurred to reconstruct the original blurred sample, and $b_s$ can be translated back to the original sharp image domain. The cycle-consistency loss further limits the space of the generated samples and preserve the content of original images. More specifically, we perform the forward translation as:

$$s_b = G_S(E_B^c(b), E^b(b)), b_s = G_B(E_S^c(s), E^b(b)) \quad (5)$$

and the backward translation as:

$$\hat{b} = G_B(E_S^c(s_b), E^b(b_s)), \hat{s} = G_S(E_B^c(b_s), E^b(b_s)) \quad (6)$$

We define the cycle-consistency loss on both domains as:

$$\mathcal{L}_{cc} = \mathbb{E}_{s \sim p(s)}[||s - \hat{s}||_1] + \mathbb{E}_{b \sim p(b)}[||b - \hat{b}||_1] \quad (7)$$

### 3.4. Perceptual Loss

From our preliminary experiments, we find that the generated deblurred samples often contain many unpleasant artifacts. Motivated by the observations from [39, 5] that features extracted from pre-trained deep networks contain rich semantic information, and their distances can act as perceptual similarity judgments, we add a perceptual loss between the deblurred images and the corresponding original blurred images:

$$\mathcal{L}_p = ||\phi_l(s_b) - \phi_l(b)||_2^2 \quad (8)$$

where $\phi_l(x)$ is the features of the $l$-th layer of the pre-trained CNN. In our experiments, we use the $\texttt{conv}_{3,3}$ layer of VGG-19 network [37] pre-trained on ImageNet [6].

There are two main reasons why we use the blurred image $b$ instead of the sharp one $s$ as the reference image in the perceptual loss. First, we make an assumption that the content information of $b$ can be extracted by the pre-trained CNN. As shown in Section 4.2, the experimental results confirm this point. Second, since $s$ and $b$ are unpaired, applying the perceptual loss between $s$ and $s_b$ will force $s_b$ to encode irrelevant content information from $s$. However, we also notice that the perceptual loss is sensitive to blur as shown in [43]. We thus carefully balance the weight of the perceptual loss and other losses to prevent $s_b$ from staying too close to $b$. The sensitivity evaluation of varying this weight is shown in the supplementary materials.

It is worth mentioning that the perceptual loss is not added to $b_s$ and $s$. This is because we do not find obvious artifacts in $b_s$ during training. Moreover, for text image deblurring, since we observe that the perceptual loss does not

help but sometimes hurt the performance, we do not include it for this task. One possible reason is that the pixel intensity distribution of text images is very different from natural images, which leads to the model pre-trained on ImageNet to be ineffective for text images.

The full objective function is a weighted sum of all the losses from (2) to (8):

$$\mathcal{L} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{KL}\mathcal{L}_{KL} + \lambda_{cc}\mathcal{L}_{cc} + \lambda_p\mathcal{L}_p \quad (9)$$

where $\mathcal{L}_{adv} = \mathcal{L}_{D_S} + \mathcal{L}_{D_B}$. We empirically set the weights of each loss to balance their importance.

### 3.5. Testing

At test time, the blurring branch is removed. Given a test blurred image $b_t$, $E_B^c$ and $E^b$ extract the content and blur features. Then $G_S$ takes the outputs and generates the deblurred image $s_{b_t}$:

$$s_{b_t} = G_S(E_B^c(b_t), E^b(b_t)) \quad (10)$$

### 3.6. Implementation Details

**Architecture and training details.** For the network architectures, we follow the similar structures as the one used in [17]. The content encoder is composed of three strided convolution layers and four residual blocks. The blur encoder contains four strided convolution layers and a fully connected layer. For the generator, the architecture is symmetric to the content encoder with four residual blocks followed by three transposed convolution layers. The discriminator applies a multi-scale structure where feature maps at each scale go through five convolution layers and then are fed into sigmoid outputs. The end-to-end design is implemented in PyTorch [31]. During training, we use Adam solver [12] to perform two steps of update on discriminators, and then one step on encoders and generators. The learning rate is initially set to 0.0002 for the first 40 epochs, then we use exponential decay over the next 40 epochs. In all the experiments, we randomly crop $128 \times 128$ patches with batch size of 16 for training. For hyper-parameters, we experimentally set: $\lambda_{adv} = 1$, $\lambda_{KL} = 0.01$, $\lambda_{cc} = 10$ and $\lambda_p = 0.1$.

**Motion blur generation.** We follow the procedure in DeblurGAN [15] to generate motion blur kernels to blur face images. A random trajectory is generated as described in [3]. Then the kernels are generated by applying sub-pixel interpolation to the trajectory vector. For parameters, we use the same values as in [15] except that we set the probability of impulsive shake as 0.005, the probability of Gaussian shake uniformly distributed in $(0.5, 1.0)$, and the max length of the movement as 10.

(a)       (b)       (c)       (d)       (e)       (f)       (g)

Figure 3. Ablation study. (a) shows the blurred image and (g) is the sharp image. (b) only contains deblurring branch (bottom branch of Fig. 2), (c) adds blurring branch (bottom branch of Fig. 2), (d) adds disentanglement ($E^b$), (e) adds the KL divergence loss, and (f) adds perceptual loss.
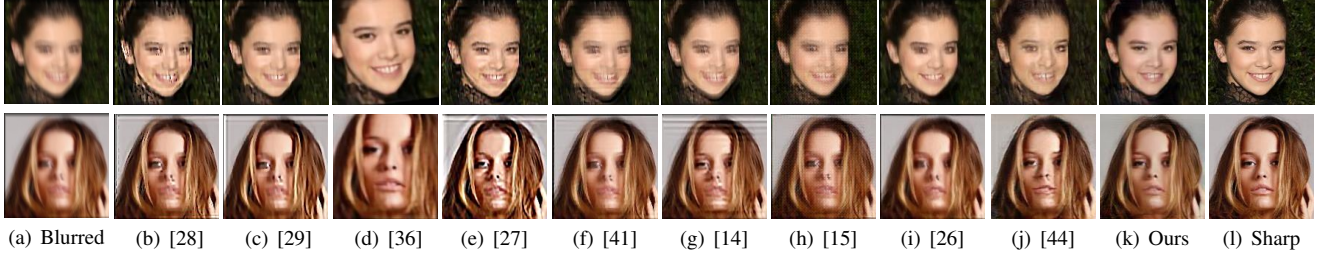


(a) Blurred   (b) [28]   (c) [29]   (d) [36]   (e) [27]   (f) [41]   (g) [14]   (h) [15]   (i) [26]   (j) [44]   (k) Ours   (l) Sharp

Figure 4. Visual performance comparison with state-of-the-art methods on CelebA dataset. Best viewed in color and by zooming in.

| Method | PSNR | SSIM | $d_{VGG}$ |
|---|---|---|---|
| Only deblurring branch | 18.83 | 0.56 | 82.9 |
| Add blurring branch | 19.84 | 0.59 | 65.5 |
| Add disentanglement | 19.58 | 0.57 | 69.8 |
| Add KL divergence loss | 20.29 | 0.61 | 60.6 |
| Add perceptual loss | **20.81** | **0.65** | **57.6** |

Table 1. Ablation study on the effectiveness of different components. $d_{VGG}$ represents the distance of feature from VGG-Face, lower is better.

## 4. Experimental Results

We evaluate the proposed approach on three datasets: CelebA dataset [20], BMVC_Text dataset [9], and CFP dataset [34].

### 4.1. Datasets and Metrics

**CelebA dataset:** This dataset consists of more than 202,000 face images. Most of the faces are of good quality and at near-frontal poses. We randomly split the whole dataset into three mutually exclusive subsets: sharp training set (100K images), blurred training set (100K images) and test set (2137 images). For the blurred training set, we use the method in Section 3.6 to blur the images. The faces are detected and aligned using the method proposed in [32].

**BMVC_text dataset:** This dataset is composed of 66,000 text images with size $300 \times 300$ for training and 94 images with size $512 \times 512$ for OCR testing. Similar to CelebA, we evenly split the training sets as sharp and blurred set. Since the dataset already contains the blurred

text images, we directly use them instead of generating new ones.

**CFP dataset:** This dataset consists of 7,000 still images from 500 subjects and for each subject, it has ten images in frontal pose and four images in profile pose. The datasets are divided into ten splits and two protocols: frontal-to-frontal (FF) and frontal-to-profile (FP). We used the same method as described above to blur the images. The faces are detected and aligned similarly as the CelebA dataset.

For CelebA and BMVC_Text datasets, we use standard debluring metrics (PSNR, SSIM) for evaluation. We also use feature distance (*i.e.*, the $L_2$ distance of the outputs from some deep networks) between the deblurred image and the ground truth image as a measure of semantic similarity because we find this to be a better perceptual metric than PSNR and SSIM [43]. For the CelebA dataset, we use the outputs of `pool5` layer from VGG-Face [30] and for the text dataset, we use the outputs of `pool5` layer from a VGG-19 network. For text deblurring, another meaningful metric is the OCR recognition rate for the deblurred text. We follow the same protocol as in [9] to report the character error rate (CER) for OCR evaluation.

To study the influence of motion blur on face recognition and test the performance of different deblurring algorithms, we perform face verification on the CFP dataset. Both frontal-to-frontal and frontal-to-profile protocol are evaluated. The frontal-to-profile protocol can further be used to examine the robustness of the deblurring methods on pose.

In order to test the generalization capability of the proposed method, we also try our approach on natural images.
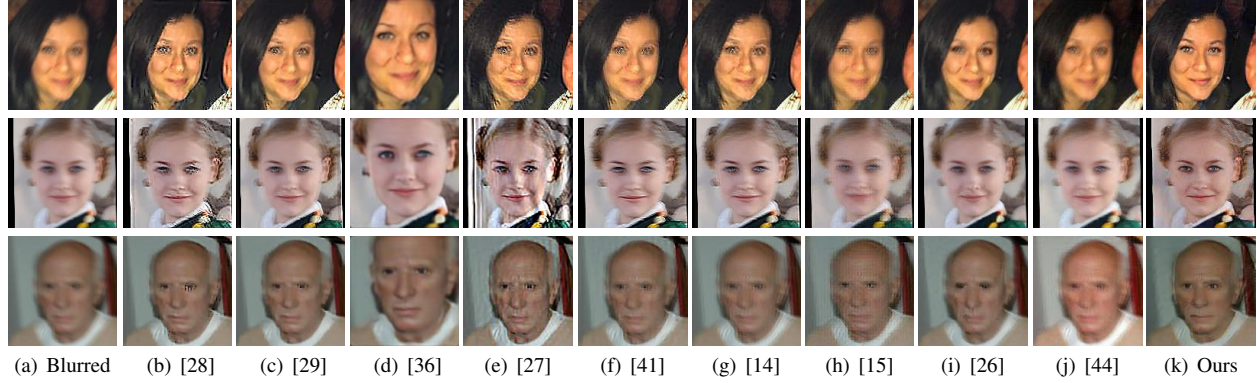
(a) Blurred  (b) [28]  (c) [29]  (d) [36]  (e) [27]  (f) [41]  (g) [14]  (h) [15]  (i) [26]  (j) [44]  (k) Ours

Figure 5. Visual comparisons with state-of-the-art methods on real blurred face images. Best viewed in color and by zooming in.

| Method | PSNR | SSIM | $d_{VGG}$ |
|---|---|---|---|
| Pan *et al.* [28] | 17.34 | 0.52 | 96.6 |
| Pan *et al.* [29] | 17.59 | 0.54 | 85.6 |
| Shen *et al.* [36] | **21.50** | **0.69** | **57.9** |
| Pan *et al.* [27] | 15.16 | 0.38 | 166.6 |
| Xu *et al.* [41] | 16.84 | 0.47 | 102.0 |
| Krishnan *et al.* [14] | 18.51 | 0.56 | 89.4 |
| Kupyn *et al.* [15] | 18.86 | 0.54 | 116.5 |
| Nah *et al.* [26] | 18.26 | 0.57 | 75.6 |
| Zhu *et al.* [44] | 19.40 | 0.56 | 103.2 |
| Ours | **20.81** | **0.65** | **57.6** |

Table 2. Quantitative performance comparison with state-of-the-art methods on CelebA dataset. $d_{VGG}$ represents the distance of feature from VGG-Face, lower is better.

| Methods | F2F Accuracy | F2P Accuracy |
|---|---|---|
| Blurred | 0.920±0.014 | 0.848±0.013 |
| Sharp | 0.988±0.005 | 0.949±0.014 |
| Pan *et al.* [28] | 0.930±0.013 | 0.853±0.010 |
| Pan *et al.* [29] | 0.935±0.015 | 0.872±0.015 |
| Shen *et al.* [36] | 0.959±0.008 | 0.821±0.022 |
| Pan *et al.* [27] | 0.916±0.011 | 0.825±0.016 |
| Xu *et al.* [41] | 0.944±0.012 | 0.865±0.013 |
| Krishnan *et al.* [14] | 0.941±0.012 | 0.857±0.014 |
| Kupyn *et al.* [15] | 0.948±0.012 | 0.872±0.007 |
| Nah *et al.* [26] | **0.960±0.007** | **0.885±0.016** |
| Zhu *et al.* [44] | 0.941±0.012 | 0.864±0.015 |
| Ours | **0.948±0.006** | **0.872±0.015** |

Table 3. Face verification results on the CFP dataset. F2F, F2P represent frontal-to-frontal and frontal-to-profile protocols.

More details are presented in the supplementary materials.

## 4.2. Ablation Study

In this section, we perform an ablation study to analyze the effectiveness of each component or loss in the proposed framework. Both quantitative and qualitative results on CelebA dataset are reported for the following five variants of our methods where each component is gradually added: 1) only including deblurring branch (*i.e.*, removing the top cycle in Fig. 2 and the blur encoder $E^b$); 2) adding blurring branch (adding the top cycle of Fig. 2); 3) adding content and blur disentanglement; 4) adding the KL divergence loss; 5) adding the perceptual loss.

We present the PSNR, SSIM and VGG-Face distance ($d_{VGG}$) for each variant in Table 1 and the visual comparisons are shown in Fig. 3. From Table 1, we can see that adding the blurring branch significantly improves the deblurring performance, especially for the perceptual distance. As shown in Fig. 3 (c) many artifacts are removed from face and colors are preserved well compared to (b). This confirms the findings in CycleGAN [44] that only one

direction cycle-consistency loss is not enough to recover good images. However, we find that adding a disentanglement component does not help but rather hurt the performance ( Fig. 3 (d)). This demonstrates that the blurring encoder $E^b$ will induce some noise and confuse the generator $G_S$ if the KL divergence loss is not enforced. In contrast, when the KL divergence loss is added to $E^b$ (Fig. 3 (e)), content and blur information can be better disentangled and we observe some improvements on both PSNR and perceptual similarities. Finally, the perceptual loss can improve the perceptual reality of the face notably. By comparing Fig. 3 (e) and (f), we find that the artifacts on cheek and forehead are further removed. Furthermore, the mouth region of (f) is more realistic than (e).

## 4.3. Face Results

**Compared methods:** We compare the proposed method with some state-of-the-art deblurring methods [28, 29, 36, 27, 41, 14, 26, 44, 15]. We directly use the pre-trained models provided by authors except for CycleGAN [44], where

10222

Figure 6. Visual results compared with state-of-the-art methods on BMVC_Text dataset. Best viewed by zooming in.

|  (a) Blurred | (b) [28] | (c) [29] | (d) [26] | (e) [44] | (f) [9] | (g) Ours | (h) Sharp |



Figure 7. Visual results compared with state-of-the-art methods on real blurred text images. Best viewed by zooming in.

|  (a) Blurred | (b) [28] | (c) [29] | (d) [26] | (e) [44] | (f) [9] | (g) Ours |

we retrain the model by using the same training set as our method. Both CNN-based models [36, 26, 44, 15] and conventional MAP-based methods are included [28, 29, 27, 41, 14]. Among these approaches, two are specific for face deblurring [27, 36] while others are generic deblurring algorithm. The kernel size for [28, 29] is set to 9. We found that the face deblurring method [36] is very sensitive to face alignment, we follow the sample image provided by the author to align the faces before running their algorithm. Meanwhile, CycleGAN is the only unsupervised CNN-based method we compare with.

**CelebA dataset results.** The quantitative results for CelebA dataset are shown in Table 2 and the visual comparisons are illustrated in Fig. 4. Our approach shows superior performance to other unsupervised algorithms on both conventional metrics and VGG-Face distance. Furthermore, we achieve comparable results with state-of-the-art supervised face deblurring method [36]. From Fig. 4 we can see that conventional methods often over-deblur or under-deblur the blurred images. Among them, Krishnan *et al*. [14] perform the best in PSNR and SSIM and Pan *et al*. [29] perform the best in perceptual distance. For CNN-based methods, Shen *et al*. [36] include a face parsing branch and achieve the best

performance among the compared methods. The results for DeblurGAN [15] contain some ringing artifacts and Cycle-GAN [44] cannot recover the mouth part of both images that well. Nah *et al*. [26] shows better visual results than other CNN-based generic methods but still contains some blur in local structures.

**Face verification results.** The face verification results for the CFP dataset are reported in Table 3. We train a 27-layer ResNet [22] on the curated MS-Celeb1M dataset [8, 18] with 3.7 millions face images and extract features of the deblurred faces for each method. Cosine similarities of test pairs are used as similarity scores for face verification. We follow the protocols used in [23, 24] and the verification accuracy for both frontal-to-frontal and frontal-to-profile protocols are reported. As shown in Table 3, the proposed method improves the baseline results of blurred images and outperforms CycleGAN [44] on both protocols. Moreover, we achieve comparable performance compared to other state-of-the-art supervised deblurring methods. Shen *et al*. [36] perform very well for frontal-to-frontal protocol, yet provide the worst performance on frontal-to-profile protocol, which shows that the face parsing network in their method is sensitive to poses. In contrast, the pro-

10223

| Method | PSNR | SSIM | $d_{VGG}$ | CER |
|---|---|---|---|---|
| Pan *et al*. [28] | 21.18 | 0.92 | 19.7 | 42.3 |
| Pan *et al*. [29] | 21.84 | 0.93 | 15.7 | 35.3 |
| Nah *et al*. [26] | 22.27 | 0.92 | 31.9 | 50.6 |
| Hradis *et al*. [9] | **30.6** | **0.98** | **1.6** | **7.2** |
| Zhu *et al*. [44] | 19.57 | 0.89 | 18.8 | 53.0 |
| Ours | **22.56** | **0.95** | **2.2** | **10.1** |

Table 4. Quantitative performance comparison with state-of-the-art methods on BMVC_Text dataset. $d_{VGG}$ represents the distance of feature from VGG net, lower is better. CER is the OCR character error rate, lower is better.

posed method works for both frontal and profile face images even though we do not explicitly train on faces with extreme poses.

**Real blurred images results** We also evaluate the proposed method on some real-world images from the datasets of Lai *et al*. [16], and the results are shown in Fig. 5. Similar to what we have observed for CelebA, our method shows competitive performance compared to other state-of-the-art approaches. Conventional methods [28, 29, 27, 41, 14] still tend to under-deblur or over-deblur images, especially on local regions such as eyes and mouths. On the other hand, the generic CNN-based method [15] does not perform very well on face deblurring. CycleGAN [44] fails to recover sharp faces but only changes the background color of images (*e.g*., third row of Fig. 5(j)). Nah *et al*. [26] produce good results on the first two faces, but generate some artifacts in the third image. Deep semantic face deblurring [36] generate better results than other compared methods. Nonetheless, due to the existence of face parsing, they tend to sharpen some facial parts (eye, nose and mouth) but over-smooth the ears and the background. In contrast, our method can not only recover sharp faces, but also restore sharp textures in the background (*e.g*., third row of Fig. 5(k)).

### 4.4. Text results

**BMVC_Text dataset results.** Similar to face experiments, we train a CycleGAN model using the same training set as our method. The kernel size for [28, 29] is set to 12. The quantitative results for BMVC_Text dataset are shown in Table 4 and some sample images are presented in Fig. 6. We can see that conventional methods [28, 29] and generic deblurring approaches [26] do not perform well on text deblurring. The visual quality is poor and the OCR error rate is very high. The results for CycleGAN [44] contain some unexplainable blue background. Although it removes the blur in images, it fails to recover recognizable text. In contrast, our method achieves good visual quality and its performance is comparable to the state-of-the-art supervised text deblurring method [9] on semantic metrics (*i.e*., per-

ceptual distance and OCR error rate). Interestingly, we find the PNSR performance for our approach is worse than the method [9] by large margins. We carefully examine our visual results and find that the proposed method sometimes changes the font of the text while deblurring. For example, as shown in the first row of Fig. 6(g), the font of our deblurred text becomes lighter and thinner compared to the original sharp text image (Fig. 6(h)). The main reason for this phenomenon is that our method does not utilize paired training data so that the deblurring generator cannot preserve some local details of text images.

**Real blurred text images results** We also evaluate our deblurring method on real blurred text images provided by Hradis *et al*. [9]. Due to space limitation, $200 \times 200$ patches are randomly cropped, and some visual results are illustrated in Fig. 7. Similar to the results of BMVC_Text dataset, we find that conventional methods [28, 29] fail to deblur the given text images. Nah *et al*. [26], in contrast, generate a reasonable deblurred result for the first image but cannot handle the second one. CycleGAN [44] again produces blue artifacts and cannot recover meaningful text information. Hradis*et al*. [9] and our approach both generate satisfactory results. Although we mis-recognize some characters (*e.g*., in the second images, "*i.e*., BING" is recovered as "*Le*.,BING"), we still correctly recover most of the blurred images.

## 5. Conclusions

In this paper, we propose an unsupervised method for domain-specific single image deblurring. We disentangle the content and blur features in a blurred image and add the KL divergence loss to discourage the blur features to encode content information. In order to preserve the content structure of the original images, we add a blurring branch and cycle-consistency loss to the framework. The perceptual loss helps the blurred image remove unrealistic artifacts. Ablation study on each component shows the effectiveness of different modules. We conduct extensive experiments on face and text deblurring. Both quantitative and visual results show promising performance compared to other state-of-the-art approaches.

# References

[1] Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *arXiv preprint arXiv:1802.10151*, 2018.

[2] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6713–6722, 2018.

[3] G Boracchi and A Foi. Modeling the performance of image restoration from motion blur. *IEEE Transactions on Image Processing*, 21(8):3502–3517, 2012.

[4] Ayan Chakrabarti. A neural approach to blind motion deblurring. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 221–235. Springer, 2016.

[5] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 1520–1529. IEEE, 2017.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[7] Rob Fergus, Barun Singh, Aaron Hertzmann, Sam T Roweis, and William T Freeman. Removing camera shake from a single photograph. In *ACM transactions on graphics (TOG)*, volume 25, pages 787–794. ACM, 2006.

[8] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large scale face recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016.

[9] Michal Hradi, Jan Kotera, Pavel Zemk, and Filip roubek. Convolutional neural networks for direct text deblurring. In Mark W. Jones Xianghua Xie and Gary K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 6.1–6.13. BMVA Press, September 2015.

[10] Zhe Hu, Sunghyun Cho, Jue Wang, and Ming-Hsuan Yang. Deblurring low-light images with light streaks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3382–3389, 2014.

[11] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[14] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 233–240. IEEE, 2011.

[15] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[16] Wei-Sheng Lai, Jia-Bin Huang, Zhe Hu, Narendra Ahuja, and Ming-Hsuan Yang. A comparative study for single image blind deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 1701–1709, 2016.

[17] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 36–52. Springer, 2018.

[18] Wei-An Lin, Jun-Cheng Chen, and Rama Chellappa. A proximity-aware hierarchical clustering of faces. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 294–301. IEEE, 2017.

[19] Yu Liu, Fangyin Wei, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Exploring disentangled feature representation beyond face identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[20] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[21] Boyu Lu, Rama Chellappa, and Nasser M. Nasrabadi. Incremental dictionary learning for unsupervised domain adaptation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.

[22] Boyu Lu, Jun-Cheng Chen, Carlos D Castillo, and Rama Chellappa. An experimental evaluation of covariates effects on unconstrained face verification. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2019.

[23] Boyu Lu, Jun-Cheng Chen, and Rama Chellappa. Regularized metric adaptation for unconstrained face verification. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 4112–4117. IEEE, 2016.

[24] Boyu Lu, Jingxiao Zheng, Jun-Cheng Chen, and Rama Chellappa. Pose-robust face verification by exploiting competing tasks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1124–1132. IEEE, 2017.

[25] Thekke Madam Nimisha, Kumar Sunil, and AN Rajagopalan. Unsupervised class-specific deblurring. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 353–369, 2018.

[26] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3883–3891, 2017.

[27] Jinshan Pan, Zhe Hu, Zhixun Su, and Ming-Hsuan Yang. Deblurring face images with exemplars. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 47–62. Springer, 2014.

[28] Jinshan Pan, Zhe Hu, Zhixun Su, and Ming-Hsuan Yang. Deblurring text images via l0-regularized intensity and gradient prior. In *Proceedings of the IEEE Conference on Computer*

*Vision and Pattern Recognition (CVPR)*, pages 2901–2908, 2014.

[29] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Blind image deblurring using dark channel prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1628–1636, 2016.

[30] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.

[31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[32] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D Castillo, and Rama Chellappa. An all-in-one convolutional neural network for face analysis. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 17–24. IEEE, 2017.

[33] Christian Schuler, Michael Hirsch, Stefan Harmeling, and Bernhard Scholkopf. Learning to deblur. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (1):1–1.

[34] S Sengupta, JunCheng Cheng, C.D Castillo, V.M Patel, R Chellappa, and D.W Jacobs. Frontal to profile face verification in the wild. In *IEEE Winter Conference on Applications of Computer Vision*, February 2016.

[35] Qi Shan, Jiaya Jia, and Aseem Agarwala. High-quality motion deblurring from a single image. In *Acm transactions on graphics (tog)*, volume 27, page 73. ACM, 2008.

[36] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Deep semantic face deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[38] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 769–777, 2015.

[39] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.

[40] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, July 2017.

[41] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural l0 sparse representation for natural image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1107–1114, 2013.

[42] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 2868–2876. IEEE, 2017.

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[44] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2017.

[45] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2017.