Human-Aware Motion Deblurring

Ziyi Shen[1,2*], Wenguan Wang[1,2*], Xiankai Lu[1], Jianbing Shen[1,2†], Haibin Ling[3], Tingfa Xu[2], Ling Shao[1]

[1] Inception Institute of Artificial Intelligence, UAE [2] Beijing Institute of Technology, China [3] Stony Brook University, USA

https://github.com/joanshen0508/HA_deblur

Abstract

This paper proposes a human-aware deblurring model that disentangles the motion blur between foreground (FG) humans and background (BG). The proposed model is based on a triple-branch encoder-decoder architecture. The first two branches are learned for sharpening FG humans and BG details, respectively; while the third one produces global, harmonious results by comprehensively fusing multi-scale deblurring information from the two domains. The proposed model is further endowed with a supervised, human-aware attention mechanism in an end-to-end fashion. It learns a soft mask that encodes FG human information and explicitly drives the FG/BG decoder-branches to focus on their specific domains. To further benefit the research towards Human-aware Image Deblurring, we introduce a large-scale dataset, named HIDE, which consists of 8,422 blurry and sharp image pairs with 37,985 densely annotated FG human bounding boxes. HIDE is specifically built to span a broad range of scenes, human object sizes, motion patterns, and background complexities. Extensive experiments on public benchmarks and our dataset demonstrate that our model performs favorably against the state-of-the-art motion deblurring methods, especially in capturing semantic details.

1. Introduction

Image deblurring, *i.e.* recovering a sharp latent image with significant details from a single degraded image, has long been an active research area in computer vision. With the increasing use of handheld devices, such as cell-phones and onboard cameras, motion blur has become a ubiquitous problem to confront with. In this work, we focus on the dynamic scene deblurring problem and propose a human-aware deblurring model by *explicitly discriminating between the blurred FG humans and BG*. Basically, due to the relative motion between a camera and objects, FG and BG often undergo different types of image degradation [31, 29].

(a) Blurred image (b) Sun *et al*. [40] (c) Nah *et al*. [29] (d) Kupyn *et al*. [22] (e) Tao *et al*. [42] (f) Ours

Figure 1: A challenging blurred image undergoes heterogeneous blur caused by camera motion and human movement.

In addition, according to the imaging mechanism, each independent object experiences a varied motion blur as well, due to the specific distance between the object and image plane. Among different objects, human beings are the most common and essential in our daily lives. Humans are often accompanied by unpredictable perturbations, thus providing a representative example for the in-depth study of the dynamic deblurring problem. In addition, restoring humans in a scene has broad application prospects in tasks such as pedestrian detection. Furthermore, with the dramatic increase in popularity of live shows and hand-held devices, massive amounts of human-focused photos and short videos have been created (since humans in these setting often draw viewer attention [3]). A specially designed and human-aware deblurring approach would be highly beneficial for processing and editing human-focused visual data.

Most existing non-uniform deblurring models [29] attempt to deblur the FG and BG simultaneously. However, this leads to inferior performance and potential artifacts due to the neglection of the multiple motion patterns. Only a

*The first two authors contributed equally to this work.

†Corresponding author: *Jianbing Shen*.

few pioneering heuristic methods [31, 18] estimate object motion blur kernels. However they do not emphasize the importance and particularity of human-focused deblurring and instead rely solely on pre-computed FG masks (*e.g.*, Fig. 1).

Though their promising results do address the value of handling FG/BG blurs separately, the last generation of deblurring models put FG/BG information aside in favor of directly learning a uniform blur kernel using neural networks. We believe that the main reasons for this radical choice are the lack of **(1)** an effective method for incorporating FG/BG information into neural networks in an end-to-end manner, and **(2)** a large-scale, carefully designed deblurring dataset with FB/BG annotations.

To tackle the first issue, we propose a novel human-aware attention guided deblurring network that learns and leverages FG human and BG masks to explicitly capture the heterogeneous FG/BG blurs in a fully differentiable and end-to-end trainable manner. More specifically, our model is built upon a fully convolutional encoder-decoder scheme. It is equipped with a *differentiable and supervised attention mechanism*, which is specially designed for learning a soft human-mask and can be trained end-to-end. Based on this attention network design, we further extend our model with a *multi-head decorder* structure containing three branches. The first two decoder branches are used to explicitly model the FG human and BG blurs, respectively, and the last one is designed for collecting and fusing both FG and BG multi-scale deblurring information and producing a final harmonious deblurring result for the whole scene. Here, the human-aware attention acts as a gate mechanism that filters out unrelated encoder features and allows the FG/BG decoder-branches to focus on their specific domains. By comprehensively fusing the deblurring features from different domains, it is able to reconstruct the image with explicit structure and semantic details. Such a design leads to a unified, human-aware, and attentive deblurring network. By explicitly and separately modeling the human-related and BG blurs, our method can better capture the diverse motion patterns and rich semantics of humans, leading to better deblurring results for both FG humans and BGs.

To address the second issue, we introduce a large-scale dataset, HIDE, which is specifically designed for *Human-aware Image DEblurring*. The dataset contains 8,422 pairs of realistic blurry images and the corresponding ground truth sharp images, which are obtained using a high-speed camera. Each pair of images is further combined with densely and professionally annotated FG human bounding boxes. Additionally, these image pairs are intentionally collected to cover a wide range of daily scenes, diverse FG human motions, sizes, and complex BG. The components described above represent a complete image deblurring dataset, which is expected to advance this field.

In summary, our **contributions** are four-fold:

- A human-aware attentive deblurring network is proposed to explore the task of motion deblurring by explicitly disentangling the blurs of FG humans and BG.
- A differentiable, supervised attention mechanism is integrated for the first time to enable the network to exclusively concentrate on FG human and BG areas.
- A novel multi-head decoder architecture is explored to explicitly model FG/BG motion blur and comprehensively fuse different-domain information for global and harmonious deblurring.
- A large-scale dataset, HIDE, is carefully constructed for human-aware image deblurring, covering a wide range of scenes, motions, *etc*., with densely annotated FG human bounding boxes.

2. Related Work

This section first provides a review of previous representative image deblurring datasets as shown in Table 1, followed by a survey of recent image deblurring models and a brief overview of differentiable neural attention.

Image Deblurring Datasets: Image deblurring has experienced remarkable progress in recent years. One of the critical factors bootstrapping this progress is the availability of large-scale datasets. Several early works [20, 40] directly convolved sharp images with a set of pre-defined motion kernels to synthesize blurry images. For instance, the BM4CS dataset [20] contains 48 blurred images generated by convolving four natural images with twelve 6D trajectories (representing real camera motions) in a patch-wise manner. Similarly, Sun *et al*. [40] built a larger dataset, which has 1,000 images sourced from the PASCAL VOC 2010 dataset [6]. Though widely used, such patch-wise generated datasets yield discrete approximations of real blurry images with pixel-wise heterogeneous blurs. Later, Gong *et al*. [9] used 200 sharp images and 10,000 pixel-wise motion maps to develop a new dataset, by associating each pixel with the corresponding motion vector. Recently, to construct a more real blurry image dataset, several researchers [29, 30] have generated dynamic blurred images by averaging multiple successive frames captured by high frame-rate video cameras. More specifically, the GoPro dataset [29] contains 2,103 pairs of blurred and sharp images in 720p quality, taken from 240 fps videos with a GoPro camera. A similar strategy was adopted in building the MSCNN (WILD) dataset [30].

Despite having greatly promoted the advancement of this field, current datasets seldom target the task of human-aware deblurring with ground truth FG annotations. This severely restrains the research progress towards a more comprehensive understanding of the underlying mechanisms of motion blur. This work proposes a new dataset,

Dataset	Pub.	Year	# Images	Resolution	Systhesis	Motion Description	Content	Pub. Ava.	Fore. Anno.
BM4CS [20]	ECCV	2012	4×12	800×800	Convolution	Camera Motion: 6D Trajectories	Natural Images	✓	
VOC-Sampled [40]	CVPR	2015	1,000	∼500×300	Convolution	Camera Motion: Rotation & Translation	Static Object & Scenes	✓	
BSD-Sampled [9]	CVPR	2016	200×10k	300×460	Convolution	Camera Motion: Rotation & Translation	Static Object & Scenes	✓	
GoPro [29]	CVPR	2017	3,214	1280×720	Integration	Dynamic Scenes	Outdoor Scenes	✓	
MSCNN(WILD) [30]	GCPR	2017	-	1280×720	Integration	Dynamic Scenes	Outdoor Scenes		
HIDE (ours)	-	2019	8,422	1280×720	Integration	Human Motion & Dynamic Scenes	Pedestrians in Outdoor	✓	✓

Table 1: Summary of existing popular non-uniform deblurring datasets and our proposed HIDE dataset (see §2).

HIDE, which is carefully constructed for human-aware deblurring and expected to inspire further explorations of non-uniform motion blur caused by object motion.

Blind-Image Deblurring: For the uniform blur problem, conventional methods typically resort to natural image priors to estimate latent images [8, 21, 36, 23, 32, 28, 51, 5, 14, 41, 50, 53, 35]. Furthermore, rather than simply assuming the whole image is associated with a uniform blur kernel, some methods estimate a global blur kernel descriptor [47, 10], or predict a set of blur kernels [13, 11]. However, they are limited by assuming that the sources of blurs are camera motions. For more general dynamic scenes with object motion blending, some other methods [17, 38, 31] estimate patch-wise blur kernels, assuming different positions with the corresponding uniform blurs. They deblur the background and object regions separately by relying on pre-generated segments [31, 17], or estimating motion flow to facilitate blur kernel prediction in a segmentation-free framework [18]. More recently, with the renaissance of neural networks, several researchers [40, 9] have turned to using deep learning to predict patch-wise blur kernels. Such methodology hinges on an intermediate for the final reconstruction. Numerous CNN-based methods have also been applied in an end-to-end fashion for image processing and generation problems, such as segmentation [26, 44], super-resolution [52, 24, 15, 16, 46], denoising [27, 59], dehazing and deraining [34, 55], enhancement [33, 56] *etc*. In a similar spirit, more advanced deep learning based deblurring models [37, 54, 29, 22, 58, 42] have been designed, for instance, the coarse-to-fine framework [29], recurrent neural networks [42, 58], and adversarial learning [22].

The promising results of these CNN-based models demonstrate well the benefits of exploring neural networks in this problem. However, in general, they do not take into account different FG human motion patterns or BGs, nor address human-aware deblurring. A few heuristic studies [17, 38, 31] have addressed the use of FG information. These methods are effective on the images with a wide-range scene or undergoing a straightforward defocus blur. However, the diacritical mechanism relies heavily on the method of segmentation and fails to learn a robust solution for multi-motion superposition in real dynamic scenes.

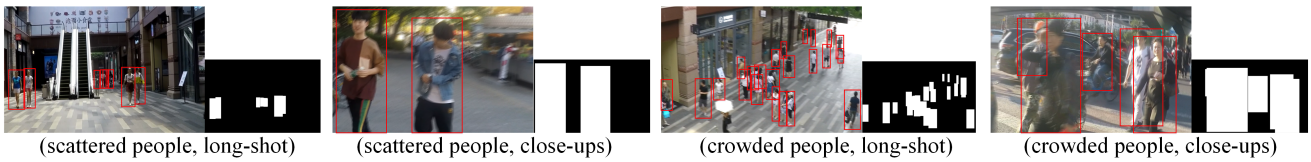
Furthermore, their significant feature engineering, high computational cost, complicated pipeline, and dependency on segmentation pre-processing, limit the performance and applicability od these methods. In this work, in addition to contributing a human-aware motion deblurring dataset, we explore FG/BG information by integrating a trainable attention mechanism into a fully convolutional encoder-decoder architecture. By associating the encoder with soft attention-based FG and BG masks, a multi-head decoder is developed to enable explicit FG and BG blur modeling, and improve the performance, simultaneously.

Differentiable Attention in Neural Networks: Recently, differentiable neural attentions have gained significant research interest. They mimic the human cognitive attention mechanism, which selectively focuses on the most visually informative parts of a scene. They were first explored in neural machine translation [2], and later proved effective in various natural language processing and computer vision tasks, such as image captioning [49], question answering [57], scene recognition [4, 43], fashion analysis [45], *etc*. In the above studies, an attention mechanism is learned in a goal-driven, end-to-end manner, allowing the network to concentrate on the most task-relevant parts of the inputs.

We propose an essential attention mechanism, called human-aware attention, which explicitly encodes FG human information by learning a soft human-mask. It drives our FG/BG decoders to focus on their specific domains and suppresses irrelevant information. Instead of learning attention implicitly, as done in the above mentioned methods, our attention module is learned from human annotations in a *supervised* manner. Additionally, the attention mechanism keeps our model fully differentiable, enabling it to be trained end-to-end. As far as we know, this is the first time an attention mechanism is leveraged for image deblurring.

3. Proposed HIDE Dataset

Dynamic blurs are caused by the relative movement between an imaging device and a scene, mainly including camera shaking and object movements. Most representative datasets [25, 41] were constructed on the basis of a simplified camera-driven assumption that camera disturbance dominates the cause of blur [47]. To model a more realistic situation, the GoPro dataset [29] further proposed to display the dynamic scenes with extra active actions. However, it is mainly concerned with wide-range scenes, ignoring significant FG moving objects, especially in close-up shots. To fully capture the dynamic blurs caused by the passive device interference and initiative actions, our HIDE dataset is



(scattered people, long-shot) (scattered people, close-ups) (crowded people, long-shot) (crowded people, close-ups)

Figure 2: Example images from our HIDE dataset with human bounding box annotations, human masks and attributes (see §3).

HIDE	Quantity of People		Depth of Object		Dataset Splits	
	Scattered	Crowded	Long-Shot (HIDE I)	Close-Ups (HIDE II)	Train	Test
# Images	4,202	4,220	1,304	7,118	6,397	2,025
# FG Human	8,866	29,119	9,122	28,863	27,714	10,271

Table 2: Statistics of the proposed HIDE dataset (see §3).

elaborately collected to cover both wide-range and close-range scenes and address human-aware motion deblurring.

Data Collection: Following the non-uniform blur generation methodology proposed in [39, 48], we capture videos at 240fps with a GoPro Hero camera. Frames from these high-fps videos are then integrated to produce plausible motion blurs. More specifically, as our dataset is designed for the multi-blur attention problem, we focus on collecting videos with humans in a close-up view to help facilitate moving-object annotation. To guarantee the diversity of the dataset, we select various real-world scenarios with different quantities of humans. Blurred images are then synthesized by averaging 11 sequential frames from a video to simulate the degradation process, and the central frame is kept as the sharp image.

We clean the initial collection by taking into account two aspects. First, to account for hardware limitations, overly quick movements are equivalent to skip frames, resulting in streak artifacts in the blurred images. Second, not all images contain an explicit structure or human in the close-up, especially if massive flat areas or pure scenes are present. Thus, we remove candidates with these drawbacks, finally reaching 8,422 sharp and blurry image pairs in total. Fig. 2 shows some sample images from our dataset.

Data Annotation: Unlike conventional pixel-wise tasks (*e.g.*, segmentation and parsing), which preserve clear and sharp object edges for labeling, for our motion deblurring dataset, the FG humans are typically subject to motion displacements, and thus cannot be annotated with precise boundaries. Therefore, we annotate the FG humans in our dataset using bounding boxes. To improve annotation efficiency, we first apply a state-of-the-art human detection model [7] to each sharp image in our dataset, which can provide roughly accurate human bounding boxes for most human objects. Then, we manually refine the inferior results and add annotations for undetected humans. To adequately apply the multi-motion blending model, we also remove the bounding boxes of BG humans in distant scenes to emphasize the close-up humans in the FG.

Dataset Statistics: Our HIDE dataset has 8,422 sharp and blurry image pairs, extensively annotated with 37,985 human bounding boxes. The images are carefully selected from 31 high-fps videos and cover realistic outdoor scenes containing humans with various numbers, poses, and appearances at various distances (see Fig. 2).

To describe this dataset in more depth, we present detailed attributes based on the quantity of humans in Table 2. The *Scattered* subset consists of 4,202 scenes with a small number of FG humans. Analogously, the *Crowded* set contains 4,220 images with large clusters of humans.

We subsequently organize the images into two categories, including long-shot (HIDE I) and regular pedestrians (close-ups, HIDE II), as shown in Table 2. Evaluating each group can capture different aspects of the multi-motion blurring problem. For the HIDE II dataset, as the FG human beings undergo more significant motions, it provides more emphasize on the challenges caused by FG actions.

Dataset Splits: For evaluation purpose, the images are split into separate training and test sets (no overlap in source videos). Following random selection, we arrive at a unique split containing 6,397 training and 2,025 test images.

4. Proposed Algorithm

4.1. Attentive Motion Deblurring Model

Vanilla Encoder-Decoder based Deblurring Model: Our human-aware motion deblurring model is built upon a convolutional encoder-decoder network architecture (see Fig. 3 (a)), which contains two parts, an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. The encoder and decoder are comprised of a stack of convolutional and transposed convolutional layers, respectively, interleaved with non-linear point-wise nonlinearity (*e.g.*, sigmoid). The encoder aims to extract a new representation $\mathbf{H} \in \mathbb{R}^{w \times h \times c}$ from an input blurry image $\mathbf{B} \in \mathbb{R}^{W \times H \times 3}$, which is used by the decoder to predict the corresponding sharp image $\hat{\mathbf{S}} \in \mathbb{R}^{W \times H \times 3}$:

$$\begin{aligned} \mathbf{H} &= \mathcal{E}(\mathbf{B}; \mathbf{W}_{\mathcal{E}}), \\ \hat{\mathbf{S}} &= \mathcal{D}(\mathbf{H}; \mathbf{W}_{\mathcal{D}}), \end{aligned} \tag{1}$$

where $\mathbf{W}_{\mathcal{E}}$ and $\mathbf{W}_{\mathcal{D}}$ are a stack of learnable convolutional kernels for the encoder and decoder, respectively. The non-linear activation layers and bias term are omitted for convenience.

To explicitly encode FG human information into our model, we further introduce a supervised, human-aware attention mechanism into the encoder-decoder architecture (see Fig. 3(b)). Before delving into the details of our model, we first elaborate the proposed attention module.

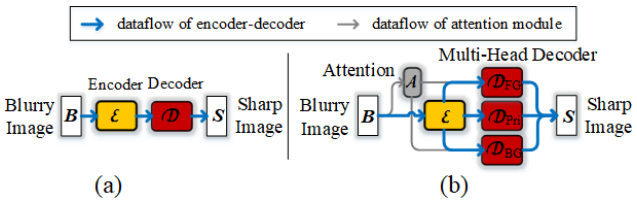


Figure 3: (a) A classical encoder-decoder based deblurring network. (b) Our proposed motion deblurring model, which is equipped with a human-aware attention module and a multi-head decoder. See §4.1 for details.

Human-Aware Attention Model: Here, we first provide a general formulation of differentiable neural attentions. Let $\mathbf{x} \in \mathbb{R}^{K\times C}$ be an input tensor, $\mathbf{z} \in \mathbb{R}^{k\times c}$ a feature obtained from $\mathbf{x}$, $\mathbf{a} \in [0,1]^k$ a soft attention vector, $\mathbf{g} \in \mathbb{R}^{k\times c}$ an attention-enhanced feature and $\mathcal{A} : \mathbb{R}^{K\times C} \to \mathbb{R}^k$ an attention network that learns to map $\mathbf{x}$ to a significance vector $\mathbf{y} \in \mathbb{R}^k$. The neural attention is implemented as:

$$\begin{aligned} \mathbf{a} &= \sigma(\mathbf{y}) = \sigma(\mathcal{A}(\mathbf{x})), \\ \mathbf{g} &= [\mathbf{a} \odot \mathbf{z}_1, \mathbf{a} \odot \mathbf{z}_2, \ldots, \mathbf{a} \odot \mathbf{z}_c], \end{aligned} \tag{2}$$

where σ indicates an activation function that maps the significance value into $[0,1]$, $\mathbf{z}_i \in \mathbb{R}^k$ indicates the feature in the i-th channel of $\mathbf{z}$, and '$\odot$' is an element-wise multiplication. The most popular strategy is to apply a softmax operation over $\mathbf{y}$ and learn (2) in an *implicit* manner.

In our approach, as we focus on image reconstruction, we extend the implicit attention above into a spatial domain. Similar to (2), our human-aware attention network $\mathcal{A}: \mathbb{R}^{W\times H\times 3} \to \mathbb{R}^{w\times h}$ learns to map the input blurry image $\mathbf{B} \in \mathbb{R}^{W\times H\times 3}$ to an importance map $\mathbf{Y} = \mathcal{A}(\mathbf{B}) \in \mathbb{R}^{w\times h}$. An attention map $\mathbf{A} \in [0,1]^{w\times h}$ can be computed:

$$\mathbf{A} = \sigma(\mathbf{Y}) = \text{sigmoid}(\mathcal{A}(\mathbf{B})). \tag{3}$$

Since we have annotations for humans, which provide the groundtruth for the attention, we relax the sum-to-one constraint of softmax and instead use a *sigmoid* activation function, *i.e.*, only constrain the attention response values ranging from 0 to 1: $\mathbf{A}_{i,j} = 1/(1+\exp(-\mathbf{Y}_{i,j}))$.

Then, we add supervision from human annotation over the attention map $\mathbf{A}$, *i.e.*, we *explicitly* train the attention network $\mathcal{A}$ by minimizing the following pixel-wise ℓ_2 loss:

$$\mathcal{L}_A = \|\mathbf{G} - \mathbf{A}\|_2^2, \tag{4}$$

where $\mathbf{G} \in \{0,1\}^{w\times h}$ is the binary FG human mask (see the small images in Fig. 2). This way, the attention $\mathbf{A}$ encodes FG human information in a fully differentiable and supervised manner, which can be viewed as a soft FG mask.

Attention-Enhanced Encoder-Features: Then, to obtain an FG human-aware attention-enhanced feature $\mathbf{H}_{\text{FG}} \in \mathbb{R}^{w\times h\times c}$, we have:

$$\mathbf{H}_{\text{FG}} = [\mathbf{A} \odot \mathbf{H}_1, \mathbf{A} \odot \mathbf{H}_2, \ldots, \mathbf{A} \odot \mathbf{H}_c]. \tag{5}$$

Similarly, we can obtain a soft BG mask through $(1-\mathbf{A}) \in [0,1]^{w\times h}$, and further obtain a BG-aware attention-enhanced feature $\mathbf{H}_{\text{BG}} \in \mathbb{R}^{w\times h\times c}$:

$$\mathbf{H}_{\text{BG}} = [(1-\mathbf{A}) \odot \mathbf{H}_1, (1-\mathbf{A}) \odot \mathbf{H}_2, \ldots, (1-\mathbf{A}) \odot \mathbf{H}_c]. \tag{6}$$

In this way, FG human and BG information are encoded into the attention-enhanced features, $\mathbf{H}_{\text{FG}}$ and $\mathbf{H}_{\text{BG}}$, while the overall image information is stored in $\mathbf{H}$.

Multi-Head Decoder: With the original image feature $\mathbf{H}$ and enhanced features $\mathbf{H}_{\text{FG}}$ and $\mathbf{H}_{\text{BG}}$, we propose a multi-head decoder. As shown in Fig. 3(b), it is comprised of three branches: a primary one, an FG one and a BG one. Each branch takes the corresponding encoder features as inputs, and performs deblurring over the corresponding regions:

$$\begin{aligned} \hat{\mathbf{S}}_{\text{FG}} &= \mathcal{D}_{\text{FG}}(\mathbf{H}_{\text{FG}}), \\ \hat{\mathbf{S}}_{\text{BG}} &= \mathcal{D}_{\text{BG}}(\mathbf{H}_{\text{BG}}), \\ \hat{\mathbf{S}} &= \mathcal{D}_{\text{Pri}}(\mathbf{H}). \end{aligned} \tag{7}$$

For brevity, the corresponding learnable weights are omitted. The three decoder branches have similar network architectures (but without weight sharing). The critical role of this multi-head decoder module is preserving domain-specific features via individual FG/BG decoder branches.

To further encourage the FG decoder $\mathcal{D}_{\text{FG}}$ and BG decoder $\mathcal{D}_{\text{BG}}$ to focus on their corresponding regions, their deblurring loss functions are designed as:

$$\begin{aligned} \mathcal{L}_{\text{D\_FG}} &= \mathbf{G} \odot \|\mathbf{S} - \hat{\mathbf{S}}_{\text{FG}}\|_2^2, \\ \mathcal{L}_{\text{D\_BG}} &= (1-\mathbf{G}) \odot \|\mathbf{S} - \hat{\mathbf{S}}_{\text{BG}}\|_2^2. \end{aligned} \tag{8}$$

Take $\mathcal{D}_{\text{FG}}$ as an example. By multiplying the squared error $\|\cdot\|$ with the binary FG human mask $\mathbf{G}$, the errors in the BG regions cannot be propagated back. This enables $\mathcal{D}_{\text{FG}}$ to handle FG blurs with more specific knowledge. Similarly, the employment of $(1-\mathbf{G})$ enables $\mathcal{D}_{\text{BG}}$ to concentrate more on deblurring of the background regions.

$\mathcal{D}_{\text{FG}}$ and $\mathcal{D}_{\text{BG}}$ capture domain-specific deblurring information, while the primary decoder $\mathcal{D}_{\text{Pri}}$ accounts for global information. To make use of different deblurring information from different decoders in an ensemble manner, our idea is to use the specific knowledge from the $\mathcal{D}_{\text{FG}}$ and $\mathcal{D}_{\text{BG}}$ branches to support $\mathcal{D}_{\text{Pri}}$. Instead of simply fusing their deblurring outputs (*i.e.*, $\hat{\mathbf{S}}_{\text{FG}}, \hat{\mathbf{S}}_{\text{BG}}$, and $\hat{\mathbf{S}}$) in a shallow manner, which might easily produce artifacts and inferior results, we use a deep knowledge-fusion strategy, *i.e.*, inject multiple intermediate features of $\mathcal{D}_{\text{FG}}$ and $\mathcal{D}_{\text{BG}}$ into $\mathcal{D}_{\text{Pri}}$. More specifically, each decoder has a total of L transposed convolutional blocks. Let us denote the features of the l-th block of $\mathcal{D}_{\text{FG}}$ ($\mathcal{D}_{\text{BG}}$) as $\mathbf{D}_{\text{FG}}^l$ ($\mathbf{D}_{\text{BG}}^l$) $\in \mathbb{R}^{w^l\times h^l\times c^l}$, where $l \in \{0, ..., L\}$, the corresponding l-th layer feature of $\mathcal{D}_{\text{Pri}}$, can be recursively defined as:

$$\mathbf{D}_{\text{Pri}}^l = \mathcal{D}_{\text{Pri}}^l(\langle \mathbf{D}_{\text{FG}}^{l-1}, \mathbf{D}_{\text{BG}}^{l-1}, \mathbf{D}_{\text{Pri}}^{l-1} \rangle), \tag{9}$$

where $\mathbf{D}_{\text{FG}}^0 = \mathbf{H}_{\text{FG}}, \mathbf{D}_{\text{BG}}^0 = \mathbf{H}_{\text{BG}}, \mathbf{D}_{\text{Pri}}^0 = \mathbf{H}$, and $\langle\cdot\rangle$ indicates concatenation. For the final L-th layer of $\mathcal{D}_{\text{Pri}}$, we have:

$$\hat{\mathbf{S}} = \mathbf{D}_{\text{Pri}}^L. \tag{10}$$

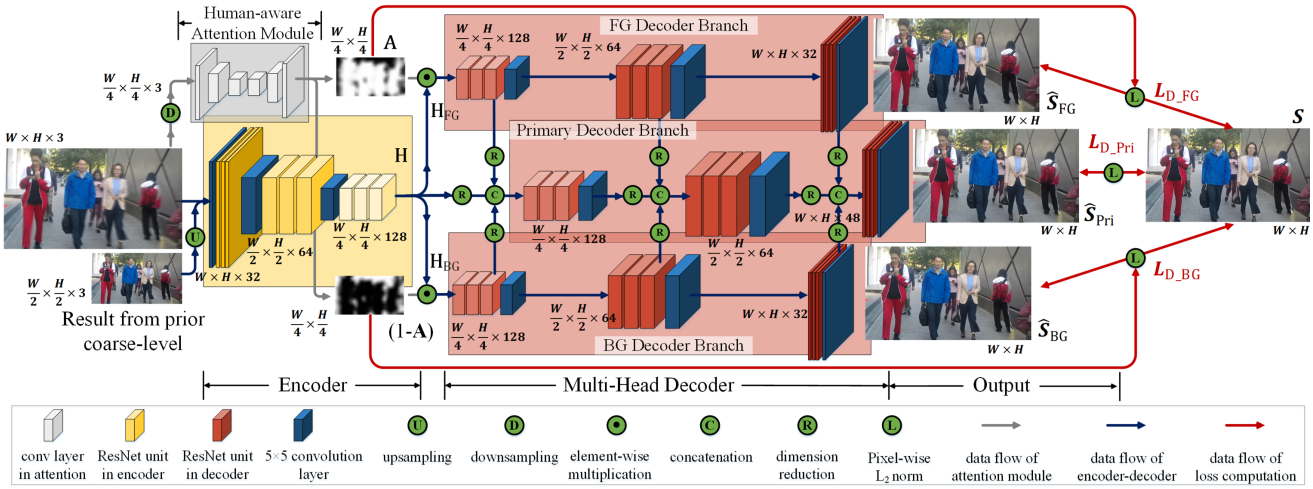


Figure 4: Overview of the proposed human-aware attentive deblurring network (single scale).

As the primary decoder $\mathcal{D}_{\text{Pri}}$ comprehensively embeds domain-specific as well as global deblurring information, its loss function is designed over the whole image domain:

$$\mathcal{L}_{\text{D\_Pri}} = \|\mathbf{S} - \hat{\mathbf{S}}_{\text{Pri}}\|_2^2. \tag{11}$$

4.2. Detailed Network Architecture

Fig. 4 illustrates the overall architecture of the proposed model. More details about the network architecture are provided in the supplementary material.

Human-Aware Attention Module: Our human-aware attention module is created as a small network (see the gray blocks in Fig. 4). First, there are three convolutional layers, interleaved with $\times 2$ max pooling and ReLU, which are stacked for effective image representation. Then, three transposed convolutional layers (each with a $\times 2$ dilation rate and ReLU) are further adapted to enhance the image representation and spatial resolution. Finally, a 1×1 convolutional layer with sigmoid nonlinearity is added to produce an FG human prediction map $\mathbf{A}$ with the same size as the input image $\mathbf{B}$, using (3).

Encoder: The encoder module $\mathcal{E}$ consists of 9 residual units [12] (the yellow blocks in Fig. 4). A 5×5 convolutional layer (the blue blocks in Fig. 4) is embedded between every three residual layers for dimensionality reduction. Then, we obtain a feature $\mathbf{H}$ and use the attention map $\mathbf{A}$ (a necessary downsampling operation is adopted) to obtain the enhanced features, $\mathbf{H}_{\text{FG}}$ and $\mathbf{H}_{\text{BG}}$, using (5) and (6), respectively.

Multi-Head Decoder: With $\mathbf{H}$, $\mathbf{H}_{\text{FG}}$ and $\mathbf{H}_{\text{BG}}$, we further apply a multi-head decoder, which has three decoder branches, $\mathcal{D}$, $\mathcal{D}_{\text{FG}}$, and $\mathcal{D}_{\text{BG}}$, to reconstruct the input blurred image in their corresponding regions (see the red blocks in Fig. 4). Briefly, each of the branches has a structure symmetrical to the encoder network, *i.e.*, comprising of nine transposed layers interleaved with dimensionality-reducing convolutional layers. In addition, a shortcut connection between the encoder and each decoder module is embedded to compensate for generalization error. Before fusing the enhanced features into the primary branch (see (10)), we first use a 1×1 convolutional layer as a feature-compression layer. We then concatenate the enhanced features to compensate the final deblurring task, using (9).

Multi-Scale Structure: A classical coarse-to-fine strategy is adopted, *i.e.*, the single-scale model above is aggregated over three scales. Here, weights are shared between scales to reduce the number of trainable parameters. The multi-scale network goes through a continuous mechanism by integrating the degraded inputs with the previous result. We extract features of each scale to enrich spatial information, and then extend the upper-scale representations by reusing the former collection. For the first scale, the input blurry image is repeated to guarantee a feed-forward formulation. We use the convolutional layer with a stride of 2 and 4×4 transposed convolutional layer to carry out downsampling and upsampling operations, respectively.

4.3. Implementation Details

Training Settings: We use the training sets of our HIDE and the GoPro dataset. There are, in total, 10,742 training images with a size of 1280×720. The GoPro dataset is only used to train the BG decoder as it contains very few pedestrians. We crop a 256×256 patch for each image and use a batch size of 10 for each iteration. In addition, since BG takes a significant fraction in the training images, randomly cropping will cause an imbalance of training data for the BG and FG decoders. To alleviate this issue, the fractions of BG and pedestrians patch in each mini-batch are set to be harmonious. We use the Adam [19] optimizer, with an initial learning rate of $1e^{-4}$. The attention network is first pre-trained with 70,000 iterations for convergence. Then, the whole deblurring network is trained over 500 epochs.

Reproducibility: Our model is implemented using Tensorflow [1]. All the experiments are done on a Titan X GPU.

Methods	GoPro [29]		HIDE			
			HIDE I		HIDE II	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Ours	**30.26**	**0.940**	**29.60**	**0.941**	**28.12**	**0.919**
w/o attention	29.30	0.929	28.40	0.927	27.58	0.912
FG branch	29.59	0.931	28.59	0.928	27.55	0.909
BG branch	29.78	0.934	28.89	0.931	27.68	0.911
Single-scale	29.85	0.934	28.47	0.930	27.81	0.916

Table 3: Ablation study of our proposed human-aware deblurring model, evaluated on the GoPro [29] and HIDE datasets using PSNR and SSIM. See §5.1 for details.

Our source code is released to provide full details of our training/testing processes and ensure reproducibility.

5. Experiments

In this section, we first perform an ablation study to evaluate the effect of each essential component of our model (§5.1). Then, we provide comparison results with several state-of-the-art deblurring methods on the GoPro [29] (§5.2) and HIDE (§5.3) datasets.

Evaluation Metrics: For quantitative evaluation, two standard metrics, Peak Signal-to-Noise-Ratio (PSNR) and Structural Similarity Index (SSIM), are adopted.

Compared Methods: As we focus on motion deblurring, we include four state-of-the-art dynamic motion deblurring models [40, 42, 29, 22] in our experiments. For a fair comparison, these models are also retrained using the training images of our HIDE and the GoPro datasets.

5.1. Ablation Study

Human-Aware Attention Module: We first assess the impact of our human-aware attention module. We derive a baseline *w/o attention* by retraining our model without the attention module. As demonstrated in Table 3, the baseline *w/o attention* clearly performs worse. From Fig. 5, we also observe that *w/o attention* cannot restore an accurate profile (see (d)), while our full model gains better results (see (e)). This shows that the attention module enables the deblurring network to reconstruct an image with more facial features and accurate shape.

Multi-Head Decoder: Next, to evaluate the effect of our multi-head decoder, we present a visual comparison between the deblurring results from different branches in Fig. 6. As shown in Fig. 6(b) and (c), the FG deblurring branch and BG branch can handle blurring in their respective regions. We further compare them with the final blending result in Fig. 6(d). We find that, by inheriting complementary features from the FG and BG branches, the main branch can successfully restore the content in the full picture.

Multi-Scale Framework: As described in §4.2, the proposed human-aware deblurring model works in a multi-scale fashion. To investigate the impact of such a design, we construct a *single-scale* baseline and present the results in Table 3. As the proposed multi-scale model comes with better convergence, a faithfully reconstructed feature with respect to the latent image can be extracted, and the feed-forward mechanism is simultaneously applied to guide the network to generate better restoration results. We provide more visual comparisons in the supplementary material.

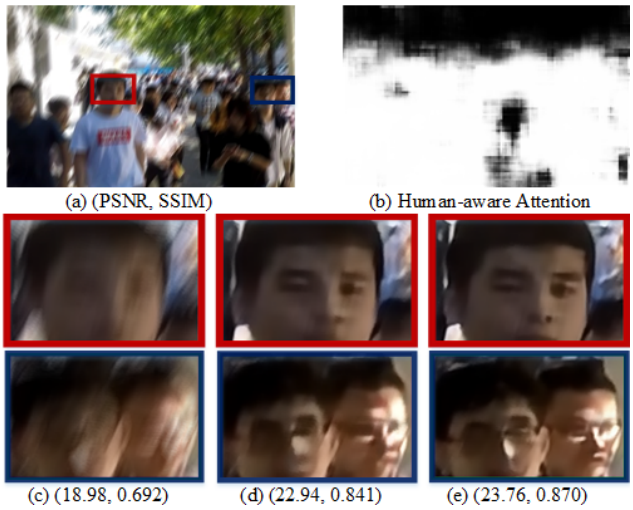


Figure 5: Demonstration of attention module for deblurring. (a) Blurred image. (b) Attention mask. (c) Blurred details. (d) Deblurred *w/o* attention. (e) Deblurred *w/* attention. See §5.1.

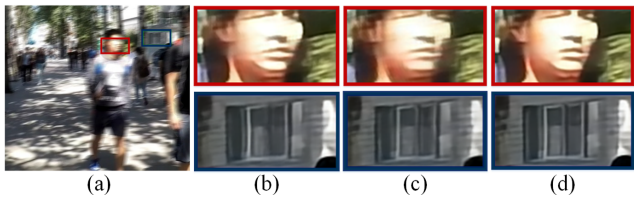


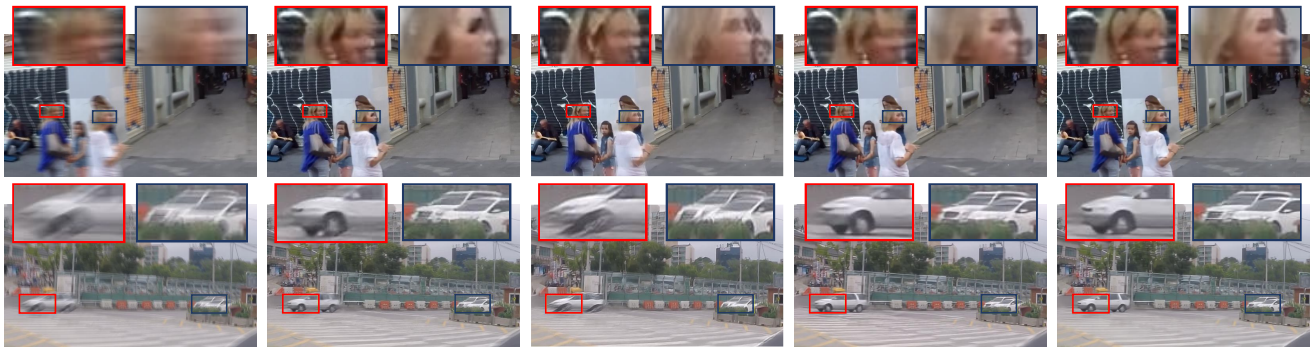
Figure 6: Effect of our multi-head structure. (a) Blurred image. (b) Deblurred result of the FG decoder branch. (c) Deblurred result of the BG decoder branch. (d) Blending deblurred result. See §5.1.

Methods	GoPro [29]		HIDE			
			HIDE I		HIDE II	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Sun *et al.* [40]	24.64	0.843	23.21	0.797	22.26	0.796
Nah *et al.* [29]	28.49	0.908	27.43	0.902	26.18	0.878
Tao *et al.* [42]	30.26	0.934	28.60	0.928	27.35	0.907
Kupyn *et al.* [22]	26.93	0.884	26.44	0.890	25.37	0.867
Ours	**30.26**	**0.940**	**29.60**	**0.941**	**28.12**	**0.919**

Table 4: Overall quantitative evaluation on the GoPro [29] and HIDE datasets using PSNR and SSIM. See §5.2 and §5.3.

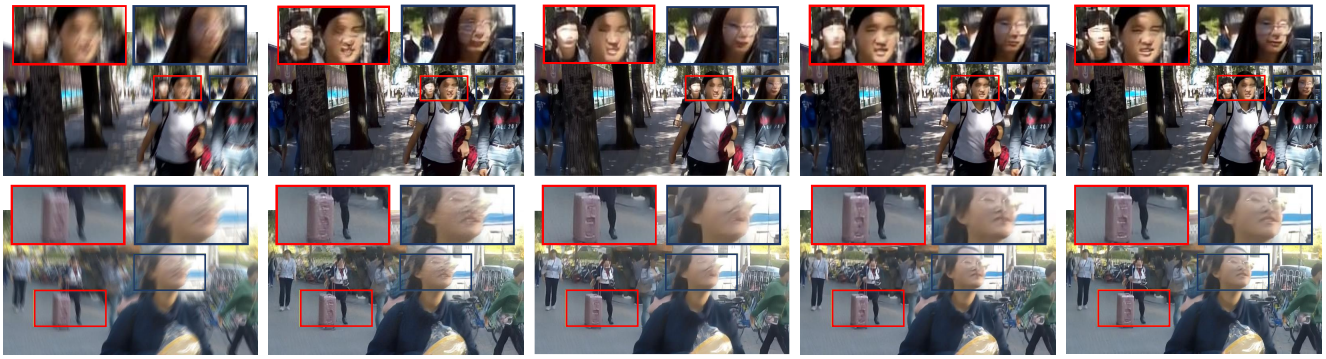
5.2. Performance on the GoPro Dataset

We first evaluate the proposed model on the GoPro dataset [29], which contains 1,111 blurred images for testing. Table 4 shows a quantitative evaluation in terms of PSNR and SSIM, where our method shows promising results. Furthermore, we provide a qualitative comparison in Fig. 7. To verify the effectiveness of the proposed attentive deblurring framework, we first provide an exam-



(a) Blurred image (b) Nah *et al*. [29] (c) Kupyn *et al*. [22] (d) Tao *et al*. [42] (e) Ours

Figure 7: Visual comparisons on the GoPro [29] dataset (see §5.2).



(a) Blurred image (b) Nah *et al*. [29] (c) Kupyn *et al*. [22] (d) Tao *et al*. [42] (e) Ours

Figure 8: Visual comparisons on our HIDE dataset (see §5.3).

ple of a blurred image consisting of a moving human with an independent motion. Our human-aware mechanism is able to perceive the specific movement and helps reconstruct a promising result with accurate face contours. Moreover, as shown in the second row of Fig. 7, our method can improve deblurring in the full frame and perform well on a scaled scene thanks to the multi-head reconstruction module, which introduces a reinforcing strategy with two branches. Overall, the proposed method applies a differentiable neural attention mechanism to a dynamic deblurring problem and achieves state-of-the-art performances.

5.3. Performance on the Proposed HIDE Dataset

We note that the proposed method focuses mainly on the moving human deblurring problem, with multiple blurs caused by camera motion and/or human movement. We further evaluate our approach on the HIDE testing set. In Fig. 8, we show visual comparisons, which relate to a specific human movement. Due to degradation by complicated motion factors, the FG human undergoes a serious blur, and thus might not be accurately restored, *e.g*., with precise facial features and unambiguous outlines. In contrast, the proposed human-aware attentive deblurring method exploits a multi-branch model to disentangle the FG human and BG. By fusing the reinforced features with the main branch, the network better restores the images with multiple blurs.

We also provide the associated qualitative comparison in Table 4. The GoPro and HIDE I datasets are mainly comprised of long-shot images, and hence involve only weak independent human movement. By contrast, HIDE II focuses on FG humans and provides a more comprehensive illustration for the moving human deblurring problem, for which our algorithm clearly outperforms previous state-of-the-arts. More deblurring results are available in the supplementary material.

6. Conclusion

In this paper, we study the problem of human-aware motion deblurring. We first create a novel large-scale dataset dedicated to this problem, which is used in our study and expected to facilitate future research on related topics as well. In addition, To handle multi-motion blur caused by camera motion and human movement, we propose a human-aware convolutional neural network for dynamic scene deblurring. We integrate a multi-branch deblurring model with a supervised attention mechanism to reinforce the foreground humans and background, selectively. By blending the different domain information, we restore the blurred images with more semantic details. Experimental results show that the proposed approach performs favorably in comparison with state-of-the-art deblurring algorithms.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, 2016. 6

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015. 3

[3] Zoya Bylinskii, Adrià Recasens, Ali Borji, Aude Oliva, Antonio Torralba, and Frédo Durand. Where should saliency models look next? In *ECCV*, 2016. 1

[4] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *ICCV*, 2015. 3

[5] Sunghyun Cho and Seungyong Lee. Fast motion deblurring. *ACM TOG (Proceedings of SIGGRAPH Asia)*, 28(5):145:1–145:8, 2009. 3

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 2

[7] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017. 4

[8] Robert Fergus, Barun Singh, Aaron Hertzmann, Sam T. Roweis, and William T. Freeman. Removing camera shake from a single photograph. *ACM TOG (Proceedings of SIGGRAPH)*, pages 787–794, 2006. 3

[9] Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian D Reid, Chunhua Shen, Anton Van Den Hengel, and Qinfeng Shi. From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur. In *CVPR*, 2017. 2, 3

[10] Ankit Gupta, Neel Joshi, C Lawrence Zitnick, Michael Cohen, and Brian Curless. Single image deblurring using motion density functions. In *ECCV*, 2010. 3

[11] Stefan Harmeling, Hirsch Michael, and Bernhard Schölkopf. Space-variant single-image blind deconvolution for removing camera shake. In *NIPS*, 2010. 3

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[13] Michael Hirsch, Christian J Schuler, Stefan Harmeling, and Bernhard Scholkopf. Fast removal of non-uniform camera shake. In *ICCV*, 2011. 3

[14] Neel Joshi, Richard Szeliski, and David J. Kriegman. PSF estimation using sharp edge prediction. In *CVPR*, 2008. 3

[15] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016. 3

[16] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, 2016. 3

[17] Tae Hyun Kim, Byeongjoo Ahn, and Kyoung Mu Lee. Dynamic scene deblurring. In *ICCV*, 2013. 3

[18] Tae Hyun Kim and Kyoung Mu Lee. Segmentation-free dynamic scene deblurring. In *CVPR*, 2014. 2, 3

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[20] Rolf Köhler, Michael Hirsch, Betty Mohler, Bernhard Schölkopf, and Stefan Harmeling. Recording and playback of camera shake: Benchmarking blind deconvolution with a real-world database. In *ECCV*, 2012. 2, 3

[21] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *CVPR*, 2011. 3

[22] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*, 2018. 1, 3, 7, 8

[23] Wei-Sheng Lai, Jian-Jiun Ding, Yen-Yu Lin, and Yung-Yu Chuang. Blur kernel estimation using normalized color-line prior. In *CVPR*, 2015. 3

[24] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *CVPR*, 2017. 3

[25] Wei-Sheng Lai, Jia-Bin Huang, Zhe Hu, Narendra Ahuja, and Ming-Hsuan Yang. A comparative study for single image blind deblurring. In *CVPR*, pages 1701–1709, 2016. 3

[26] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *CVPR*, pages 3623–3632, 2019. 3

[27] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *NIPS*, 2016. 3

[28] Tomer Michaeli and Michal Irani. Blind deblurring using internal patch recurrence. In *ECCV*, 2014. 3

[29] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 1, 2, 3, 7, 8

[30] Mehdi Noroozi, Paramanand Chandramouli, and Paolo Favaro. Motion deblurring in the wild. In *GCPR*, 2017. 2, 3

[31] Jinshan Pan, Zhe Hu, Zhixun Su, Hsin-Ying Lee, and Ming-Hsuan Yang. Soft-segmentation guided object motion deblurring. In *CVPR*, 2016. 1, 2, 3

[32] Jin-shan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Blind image deblurring using dark channel prior. In *CVPR*, 2016. 3

[33] Wenqi Ren, Sifei Liu, Lin Ma, Qianqian Xu, Xiangyu Xu, Xiaochun Cao, Junping Du, and Ming-Hsuan Yang. Low-light image enhancement via a deep hybrid network. *TIP*, 28(9):4364–4375, 2019. 3

[34] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. *CVPR*, 2018. 3

[35] Wenqi Ren, Jiawei Zhang, Lin Ma, Jinshan Pan, Xiaochun Cao, Wangmeng Zuo, Wei Liu, and Ming-Hsuan Yang. Deep non-blind deconvolution via generalized low-rank approximation. In *NeurIPS*, pages 297–307, 2018. 3

[36] Qi Shan, Jiaya Jia, and Aseem Agarwala. High-quality motion deblurring from a single image. *ACM TOG (Proceedings of SIGGRAPH)*, 27(3):73:1–73:10, 2008. 3

[37] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Deep semantic face deblurring. *CVPR*, 2018. 3

[38] Jianping Shi, Li Xu, and Jiaya Jia. Discriminative blur detection features. In *CVPR*, 2014. 3

[39] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *CVPR*, 2017. 4

[40] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *CVPR*, 2015. 1, 2, 3, 7

[41] Libin Sun, Sunghyun Cho, Jue Wang, and James Hays. Edge-based blur kernel estimation using patch priors. In *ICCP*, 2013. 3

[42] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, 2018. 1, 3, 7, 8

[43] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, 2017. 3

[44] Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven CH Hoi, and Haibin Ling. Learning unsupervised video object segmentation through visual attention. In *CVPR*, pages 3064–3074, 2019. 3

[45] Wenguan Wang, Yuanlu Xu, Jianbing Shen, and Song-Chun Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *CVPR*, 2018. 3

[46] Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, and Thomas Huang. Deep networks for image super-resolution with sparse prior. In *CVPR*, 2015. 3

[47] Oliver Whyte, Josef Sivic, Andrew Zisserman, and Jean Ponce. Non-uniform deblurring for shaken images. *IJCV*, 98(2):168–186, 2012. 3

[48] Patrick Wieschollek, Michael Hirsch, Bernhard Schölkopf, and Hendrik PA Lensch. Learning blind motion deblurring. In *ICCV*, 2017. 4

[49] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 3

[50] Li Xu and Jiaya Jia. Two-phase kernel estimation for robust motion deblurring. In *ECCV*, 2010. 3

[51] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural L0 sparse representation for natural image deblurring. In *CVPR*, 2013. 3

[52] Xiangyu Xu, Yongrui Ma, and Wenxiu Sun. Towards real scene super-resolution with raw images. In *CVPR*, pages 1723–1731, 2019. 3

[53] Xiangyu Xu, Jinshan Pan, Yu-Jin Zhang, and Ming-Hsuan Yang. Motion blur kernel estimation via deep learning. *TIP*, 27(1):194–205, 2018. 3

[54] Xiangyu Xu, Deqing Sun, Jinshan Pan, Yujin Zhang, Hanspeter Pfister, and Ming-Hsuan Yang. Learning to super-resolve blurry face and text images. In *ICCV*, pages 251–260, 2017. 3

[55] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *CVPR*, 2017. 3

[56] Xin Yang, Ke Xu, Yibing Song, Qiang Zhang, Xiaopeng Wei, and Rynson WH Lau. Image correction via deep reciprocating hdr transformation. In *CVPR*, pages 1798–1807, 2018. 3

[57] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016. 3

[58] Jiawei Zhang, Jinshan Pan, Jimmy Ren, Yibing Song, Linchao Bao, Rynson W.H. Lau, and Ming-Hsuan Yang. Dynamic scene deblurring using spatially variant recurrent neural networks. In *CVPR*, 2018. 3

[59] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *TIP*, 26(7):3142–3155, 2017. 3