Learning to Understand Image Blur

Shanghang Zhang[†⋆], Xiaohui Shen[‡], Zhe Lin[‡], Radomír Měch[‡], João P. Costeira[⋆], José M. F. Moura[†]
[†]Carnegie Mellon University [‡]Adobe Research [⋆]ISR - IST, Universidade de Lisboa
{shanghaz, moura}@andrew.cmu.edu, {zlin, xshen, rmech}@adobe.com, jpc@isr.ist.utl.pt

Abstract

While many approaches have been proposed to estimate and remove blur in a photo, few efforts were made to have an algorithm automatically understand the blur desirability: whether the blur is desired or not, and how it affects the quality of the photo. Such a task not only relies on low-level visual features to identify blurry regions, but also requires high-level understanding of the image content as well as user intent during photo capture. In this paper, we propose a unified framework to estimate a spatially-varying blur map and understand its desirability in terms of image quality at the same time. In particular, we use a dilated fully convolutional neural network with pyramid pooling and boundary refinement layers to generate high-quality blur response maps. If blur exists, we classify its desirability to three levels ranging from good to bad, by distilling high-level semantics and learning an attention map to adaptively localize the important content in the image. The whole framework is end-to-end jointly trained with both supervisions of pixel-wise blur responses and image-wise blur desirability levels. Considering the limitations of existing image blur datasets, we collected a new large-scale dataset with both annotations to facilitate training. The proposed methods are extensively evaluated on two datasets and demonstrate state-of-the-art performance on both tasks.

1. Introduction

Image blur is very common in natural photos, arising from different factors such as object motion, camera lens out-of-focus, and camera shake. In many cases it is undesired, when important regions are affected and become less sharp; while in other cases it is often desired, when the background is blurred to make the subject pop out, or motion blur is added to give the photo artistic look. Many research efforts have been made to either detect the undesired blur and subsequently remove it [22, 11, 37, 4], or directly estimate the desired blur and then enhance it [2, 38, 23, 8, 21]. However, there are rather limited efforts to have an algorithm automatically understand whether such blur is desired or not in the first place, which would be very useful to help users categorize photos and make corresponding edits, especially with the dramatic growth in the number of personal photos nowadays. It can also be used to estimate photo quality and applied in photo curation [31], photo collage creation [20], image quality and aesthetics [15], and video summarization [16].

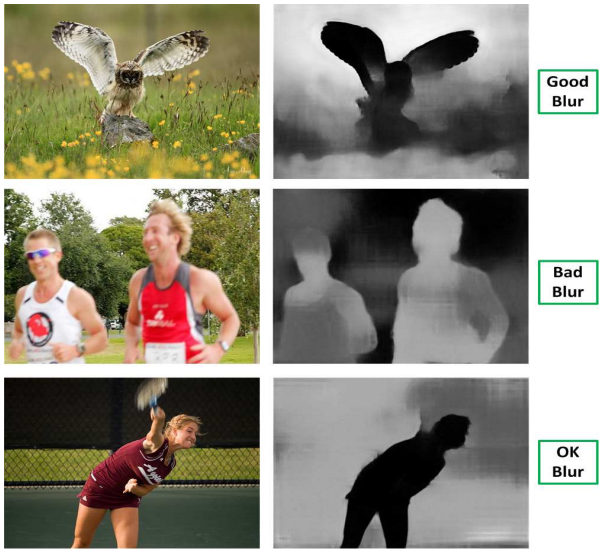

Figure 1. Problem statement. Given the natural photos in the left column, we generate their corresponding blur maps and estimate if the blur is desirable. Brighter color indicates higher blur amount.

Understanding blur desirability in terms of image quality nevertheless is not trivial and in many cases very challenging, as it not only requires accurate spatially-varying blur amount estimation, but also needs to understand if the blurry regions are important from the perspective of image content and sometimes user's intent when capturing the photo. Take the examples in Fig.1 for instance, both images in the first and second row are with depth-of-field effect. Yet the first one is regarded as a good photo while the second one is considered bad by most people, only because we think the blurry runners are the subject intended to be captured and more important than other content in the scene. The blur desirability in the third example is somewhere in between, as even though the tennis racket and the right arm of the player are blurred, her major body and face are clear,

which conveys the most important information in the photo.

Motivated by this observation, we propose a novel algorithm for image blur understanding by fusing low-level blur estimation and high-level understanding of important image content at the same time. Given an image, our approach can automatically determine if blur exists in the image, and if exists, can accurately estimate spatially-varying blur amount and categorize the blur desirability in terms of image quality to three levels: Good, OK, and Bad, as shown in Fig.1. Specifically, we propose a unified **ABC-FuseNet**, a deep neural network that jointly learns the attention map (A), blur map (B), and content feature map (C), and fuses them together to detect if there is blur on important content and estimate the blur desirability. The pixel-wise blur map estimation is based on a dilated fully convolutional network (FCN) with specifically designed global pyramid pooling mechanism. The local and global cues together make the blur map estimation more reliable in homogeneous regions and invariant to multiple object scales. The entire network is end-to-end jointly trained on both pixel-wise blur map estimation and image-level blur categorization.

Solving such a problem is in need of a large dataset with both pixel-level blur amount annotation and image-level blur category supervision. Considering the limitations of existing blur image dataset in both quality and quantity, we collect a new dataset SmartBlur, containing $10,000$ natural photos with elaborate human annotations of both pixel-level blur amount and image-level blur categories, to facilitate our training and evaluation. Contributions of this paper are summarized as follows:

- To the best of our knowledge, our work is the first attempt to detect spatially-varying blur and understand image blur in terms of image quality at the same time. In particular,we propose an end-to-end trainable neural network ABC-FuseNet to jointly estimate blur map, attention map, and content feature map, which are fused together to understand important content in the image and perform final blur desirability estimation.
- We collect a large-scale blur image dataset SmartBlur, containing $10,000$ natural photos with annotations of both pixel-level blur amount and image-level blur desirability, which we plan to release in the future. Besides the tasks addressed in the paper, SmartBlur can serve as a versatile benchmark for various tasks such as blur magnification and image deblur. Data is released at https://github.com/Lotuslisa/Understand_Image_Blur.
- The proposed approach is extensively evaluated on SmartBlur as well as a public blur image dataset [23]. Experimental results show it significantly outperforms the state-of-the-art baseline methods on both blur map estimation and blur desirability categorization.

2. Related Work

Most existing work focused on local blur detection, assuming the users already know the blur category (desired or undesired) [8]. Different cues and hand-craft features are used to estimate blur amount, such as image gradients [38], local filters [23], sparse representation [24], local binary patterns [33], and relevance to similar neighboring regions [29]. Nevertheless, those hand-craft features are error-prone as they are not robust to various conditions and are lack of semantic information. In recent years, neural networks have proved their superiority to the conventional counterparts [12, 27, 32, 6]. Park et al. [21] improve the accuracy of defocus blur estimation by combining handcrafted features with deep features from a convolutional neural network (CNN). This work limits its application to defocus blur estimation, and often fails when detecting blurs caused by camera shake. In addition, all the above-mentioned methods do not estimate whether the detected blur is desired or not in terms of image quality.

More recently, Yu et al.[34] learn a deep neural network to detect photographic defects, including undesired blur. However, there is no explicit understanding on the image content in their learning. As a result, the model sometimes still mis-classifies good depth-of-field effects into undesired defects. It also suffers from low accuracy due to limited training data in terms of both annotation quality and quantity. Although image blur analysis has been an active research area for recent years, we found that there are very limited number of high-quality blur image datasets [19, 1]. The most widely used blur image dataset-CUHK [23] only has pixel-level binarized annotations. The scale of CUHK is also small (1000 images).

3. The SmartBlur Dataset

To train and evaluate the proposed ABC-FuseNet, we need a large-scale dataset with both pixel-level blur amount and image-level blur desirability annotations. However, existing datasets only contain limited number of images with coarsely-annotated blur amount, and no annotations on blur desirability, as shown in Table 1. Therefore, we collect a new dataset SmartBlur, which contains $10,000$ natural photos with elaborate human annotations of both pixel-level blur amount and image-level blur desirability to supervise the blur map estimation and blur desirability classification. SmartBlur provides a reliable training and evaluation platform for blur analysis, and can serve as a versatile benchmark for various tasks such as blur magnification and image deblur. In this section, we describe the data collection and annotation with detailed statistics. More details can be found in the supplementary material. SmartBlur will be publicly available to promote research in blur analysis.

Dataset	CUHK[23]	CERTH[19]	Portland[18]	**SmartBlur**
# of Images	1000	2450	2976	**10,000**
Blur Type	1,2	1,2,3	3	**1,2,3**
Blur Amount	Pixel-wise binary	Image-wise binary	Image-wise binary	**Pixel-wise multi-level**
Blur Desirability	X	X	X	✓
Image Source	Natural	Natural+Synthetic	Synthetic	**Natural**

Table 1. Comparison of blur image datasets. For Blur Type, 1, 2, 3 indicates motion blur, defocus, and camera shake respectively.

3.1. Data Collection

To collect a large and varied set of natural photos, we download $75,000$ images from Flickr which carry a Creative Commons license. Then we select $10,000$ images for further annotation. When selecting these $10,000$ photos, we try to balance the number of images of different image blur desirability levels: Good blur, OK blur, Bad blur, and No blur (if there is no blur in the image). We also try to have photos with different blur types: object motion, camera shake, and out-of-focus. These $10,000$ images are captured by various camera models in different shooting conditions, and cover different scenes. Image resolution ranges from 500×300 to 1024×720. To our knowledge, SmartBlur is the largest blur image dataset with richest annotations.

3.2. Data Annotation

For each image in SmartBlur, we have two levels of annotations: pixel-level blur amount and image-level blur desirability. We train professional annotators on both labeling tasks. Each image is labeled by 3 annotators, and we check and merge the final annotations to make sure they are correct. As shown in Fig. 2, for pixel-level blur amount annotation, we label each region in the image with four blur amounts: No Blur, Low Blur, Medium Blur, and High blur. This is distinctly different from the existing datasets, which only indicate the pixel-level or image-level blur existence. We classify them based on the visual appearance with predefined criteria: No blur - no visible blur; Low - the blur is visible, but people can still see the details in blurred region; Medium - the details are not clear anymore; High - not only details are missing, the textures are largely changed, and the shapes are distorted. The boundary of each region is annotated based on the blur amount, instead of object semantics. For image-level blur desirability, we label each image with four categories: good-blur, ok-blur, bad-blur, or no-blur. Good-blur indicates the blur is manipulated by photographers to create visually pleasing effects. The blur in good-blur images often appears on the background or unimportant objects. Ok-blur indicates the blur is on some small or unimportant regions, or with negligible small amount. Such blur is not created on purpose, and is usually generated due to imperfect capture conditions or limited expertise of the photographer. Bad-blur indicates the blur is on the important objects with non-negligible amount. Such blur is not desirable and significantly degrade the image quality. No-blur indicates the whole image is sharp, with no blur in it. Annotation samples are shown in Fig. 2.

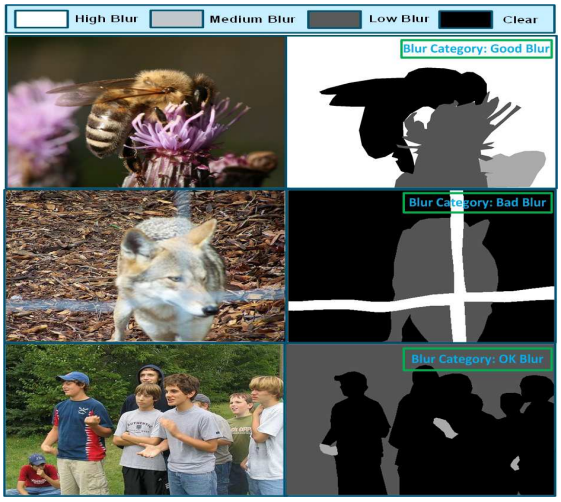

Figure 2. Annotation Samples from SmartBlur.

	Bad-Blur	Ok-Blur	Good-Blur	No-Blur	Total
Training	1568	1583	3777	1422	8400
Validation	200	200	200	200	800
Testing	200	200	200	200	800
Total	1968	1983	4177	1822	**10,000**

Table 2. Dataset split and image amount for different categories.

SmartBlur consists of $1,822$ no-blur images, $1,968$ bad-blur images, $1,983$ ok-blur images, and $4,177$ good-blur images, making it with $10,000$ images in total. We randomly split it into three portions: training, validation, and testing. The image amount for each set, as well as for each category is described in Table 2. For evaluation and validation, we random select the same amount of images from each blur type to balance the data of different categories.

Compared with existing datasets, SmartBlur has the following advantages: 1. It is the first dataset that has pixel-level blur amount annotations with multiple levels, from low, medium to high. 2. It is the first dataset that has image-level blur desirability annotation in terms of image quality. 3. It is the largest blur image dataset, with all natural photos.

4. Proposed Approach

In this paper, we introduce the problem of automatically understanding image blur in terms of image quality. Such a task not only relies on low-level visual features to detect blur regions, but also requires high-level understanding of the image content and user intent. In this section, we propose ABC-FuseNet, a unified framework to jointly estimate spatially-varying blur map and understand its effect on image quality to classify blur desirability.

4.1. Approach Overview

The architecture of ABC-FuseNet is provided in Fig. 3. ABC-FuseNet is a novel network to fuse low-level blur es-

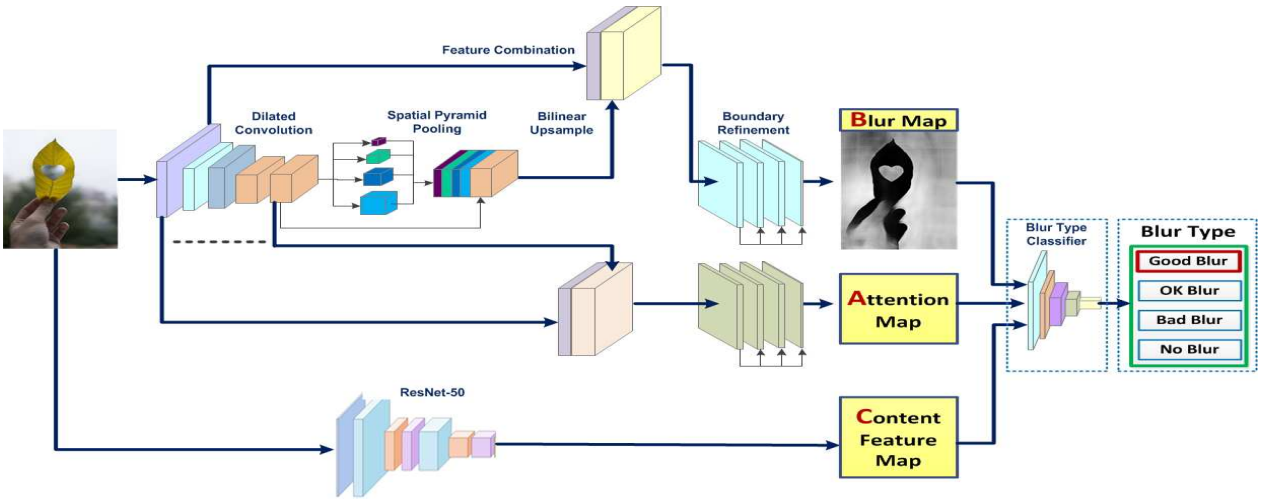

Figure 3. Architecture of ABC-FuseNet. It jointly learns the blur map, attention map, and content feature map, and fuses them together to detect if there is blur on important content and estimate the blur desirability.

timation and high-level understanding of important image content. Given an image, our approach automatically determine if blur exists in the image. If blur exists, we accurately estimate spatially-varying blur amount and classify its blur desirability into three categories ranging from good to bad, by distilling high-level semantics and learning an attention map to adaptively attend to important regions. In particular, ABC-FuseNet jointly learn the attention map, blur map, and content feature map, and fuse them together for blur desirability classification. We use a dilated fully convolutional neural network (upper branch in Fig. 3) with pyramid pooling and boundary refinement module to generate high-quality blur response maps. The local and global features together make the blur map estimation more reliable in homogeneous regions and invariant to multiple object scales. Attention map estimation is based on the fully convolutional network (middle branch in Fig. 3). The entire network is end-to-end trained on both pixel-level blur map estimation and image-level blur desirability categorization.

4.2. Blur Map Estimation

The blur map is estimated with fully convolutional neural networks (FCN), building on top of Inception-V2 [28][1]. Accurate blur map estimation is faced with two main challenges. First, it is difficult to detect blurs in small regions, because the feature map resolution is reduced by the repeated combination of max-pooling and downsampling (striding) performed at consecutive layers in the CNN, which is originally designed for image classification. To effectively enlarge the receptive fields without sacrificing much spatial resolutions, we remove the downsampling operator and replace the regular convolution in $Inception\_4a$ with dilated convolutions [5]. In addition, we combine the high-level semantic features with the low-level features after the first convolution layer to keep spatial resolution and make better estimation of blurs in the small regions. Specifically, the high-level features are upsampled by bilinear interpolation and then concatenated with the low-level features along the channel dimension. To further obtain better blur region boundaries, several boundary refinement layers with dense connections are appended after upsampling.

The second challenge is to detect blurs in multiple scale objects and in the homogeneous regions, which show almost no difference in appearance when they are sharp or blurred. A standard way to deal with the challenge of variable scales is to re-scale the CNN for the same image and then aggregate the feature or score maps [14, 7], which significantly increases computation cost. Inspired by [36], we adopt a pyramid pooling module to combine the local and global clues together to make the final blur detection more reliable in the homogeneous regions and invariant to multiple object scales. Such strategy provide hierarchical global prior, containing information with different scales and varying among different sub-regions . To be specific, we pool four-level features from $Inception\_5b$: 1×1, 2×2, 3×3, 6×6. To maintain the weight of global feature, we use 1×1 convolution layer after each pyramid level to reduce the dimension of context representation to $1/4$ of the original one. Then we upsample each pooled feature map into the same size as $Inception\_5b$ and concatenate them together as the final pyramid pooling feature.

4.3. Blur Desirability Classification

As understanding image blur relies on both low-level visual features to estimate blur responses map, and high-level understanding of the image content and user intent. We further learn content feature map to facilitate blur desirability

[1]While other networks such as ResNet [9] and VGGNet [25] can also be utilized as the backbone network, we choose Inception-V2 for its relatively smaller model size.

classification. Specifically, we extract semantic feature map from $res5c$ of ResNet-50 [9] with pretained weights (lower branch in Fig. 3). To understand if blur is on the important content in the image, we estimate an attention map at the same time to adaptively localize the important content. The attention map estimation is based on the fully convolutional networks. We pre-train the attention map branch with salient object segmentation datasets [35] to obtain the initial weights.

After learning the blur map (B_m), attention map (A_m), and content feature map (C_m), we fuse these three maps together and feed them to a light classifier to estimate the image blur category. Here we propose a dual attention mechanism to extensively exploit the blur responses and high-level semantics when concatenating these three maps together. To be specific, we stack $B_m \times A_m$, $B_m \times (1 - A_m)$, and C_m in the channel direction to form the final input of the blur category classifier, which contains two convolution layers, two dropout layers, and one fully connected layer [2]. The whole ABC-FuseNet is end-to-end trainable, in which the blur map estimation and blur desirability classification are jointly trained with both supervisions. We conduct extensive ablation study in Section 5 to verify the efficacy of the proposed mechanisms.

For blur map estimation, we apply sigmoid function on the last layer output of blur map estimation branch. Then, we compute the $L2$ loss between the estimated blur map and the ground truth blur map. As the blur amount for each pixel is annotated with four different levels in SmartBlur, we normalize these amounts into 0, $1/3$, $2/3$, and 1 respectively. The loss function of the blur map estimation is:

$$L_{B_m} = \frac{1}{2N} \sum_{i=1}^{N} \sum_{p=1}^{P} \left\| \frac{1}{1 + exp(-b_i(p;\Theta))} - b_i^0(p) \right\|_2^2 \tag{1}$$

where $b_i(p;\Theta)$ is the estimated blur amount for pixel p in image i, and Θ indicates the parameters of the blur estimation branch. $b_i^0(p)$ is the ground truth blur amount for pixel p in image i.

For the image blur desirability classification, we convert each blur category label into an one-hot vector to generate the ground truth supervision of each training image. The loss of the blur desirability classification L_{B_c} is computed by the softmax cross-entropy loss. We note that, there is no supervision for the attention map estimation. The attention region in each image is estimated by the weakly supervised learning from the image blur category. To this end, the total loss of the ABC-FuseNet is:

$$L = L_{B_m} + \lambda L_{B_c} \tag{2}$$

[2]Detailed architectures are described in the supplementary material.

5. Experiments

To verify the efficacy of ABC-FuseNet for both blur map estimation and image blur type classification, we extensively evaluate the proposed methods on two datasets, CUHK [23] and SmartBlur. In this section, we discuss the experiments and results: 1. We first evaluate and compare ABC-FuseNet with the state-of-the-art methods on CUHK [23] for the task of blur map estimation. Experimental protocol and implementation details are provided. Here we show our proposed method significantly outperforms the existing methods in terms of both quantitative and qualitative results regardless of the blur sources (object motion, camera shake, or defocus). 2. We then evaluate the proposed methods on the SmartBlur dataset for both blur map estimation and image blur type classification. We compare with the state-of-the-art methods and conduct thorough ablation studies to verify the efficacy of ABC-FuseNet.

Implementation details. To train the ABC-FuseNet, we first pretrain the blur map estimation and attention map estimation branches with salient object segmentation dataset [35] to obtain the initial weights. Afterwards, we further train the blur map estimation branch with the SmartBlur dataset. The loss function is optimized via batch-based Adam [13] and backpropagation. The hyperparameters, including initial learning rate, weight decay penalty multiplier, and dropout rate are selected by cross-validation, and are set to be 0.001, 0.05, and 0.5 respectively. The batch size is 12 images for training. Then we test the performance of blur map estimation on two datasets, CUHK and SmartBlur. Detailed results are described in Sec. 5.1 and Sec. 5.2 respectively. After obtaining the initial weights of blur map and attention map estimation branches, we jointly train the network with both blur map supervision and blur desirability supervision. The hyperparameters, including the coefficient of blur type classification loss λ, initial learning rate, weight decay penalty multiplier, and dropout rate are selected by cross-validation, and are set to be 0.1, 0.01, 0.01, and 0.5 respectively. The batch size is 4 images for training. To improve the generalization and robustness of the network, we apply various data augmentation techniques to all the training processes: 1. horizontal flip, 2. random crop, 3. random brightness, 4. and random contrast.

5.1. Evaluations on CUHK Dataset

Experiment Settings. We first verify the reliability and robustness of our algorithm on a public blur detection dataset CUHK [23]. It contains $1,000$ images with human labeled blur regions, among which 296 images are partially motion-blur and 704 images are defocus-blur. It was the most widely used blur image dataset with pixel-level binary annotations (1 indicates blur, and 0 indicates clear). As most of the existing blur detection methods are not learning based and do not have training images from CUHK,

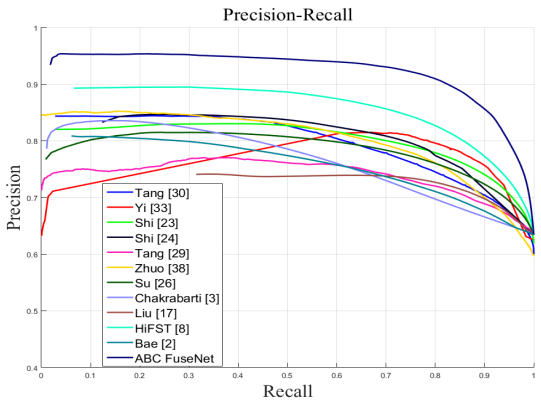

Figure 4. Quantitative Precision-Recall comparison on CUHK for different methods, tested on all blur types.

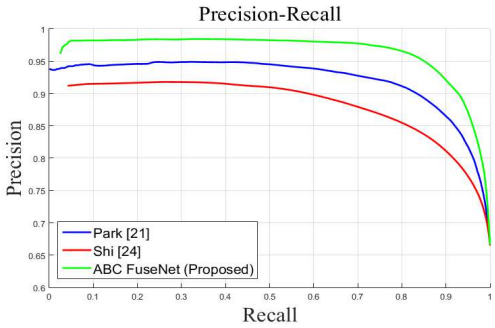

Figure 5. Quantitative Precision-Recall comparison on CUHK for different methods, tested on defocus blur.

for a fare comparison with the baselines, we only train the ABC-FuseNet on our collected SmartBlur dataset and directly test the trained model on the $1,000$ images of the CUHK dataset, without finetuning on the CUHK dataset at all. Such treatment also guarantees that our method is evaluated on the same amount of testing set as the baselines.

Experimental Results. We extensively compare the performance of our method with the state-of-the-art baseline methods [2, 3, 17, 23, 24, 26, 29, 30, 33, 38, 21, 8], using publicly released implementations. While most of the baselines use hand-crafted visual features, work [21] combined hand-crafted features with deep features to estimate the defocus blur map. The quantitative performance is evaluated using the precision-recall curve.

Fig. 4 and Fig. 5 show the quantitative Precision-Recall comparison on CUHK for different methods. Fig. 4 is the precision-recall curve tested on $1,000$ blur images, including both motion blur and defocus blur. Fig. 5 is the precision-recall curve tested on 704 defocus blur images. Note that baseline Park et al. [21] is designed for the defocus blur detection. From the comparison we can see that, for the performance tested on the $1,000$ images with different blur sources, our method consistently outperforms all the state-of-the-art baselines by a large margin, which verifies its efficacy in detecting blur from different levels and sources. For the results tested on 704 defocus blur images, our model also significantly outperforms Park et al. [21] and Shi et al. [24]. The average precision on CUHK before/after joint training are 0.869 and 0.868, respectively. Joint training would focus the blur map estimation on more important semantic regions, which might not be reflected in average precision uniformly evaluated over the entire images. However, it could significantly improve blur desirability classification (Fig. 9).

For qualitative comparison, we show visual results of some challenging images in CUHK for different methods [23, 24, 38, 21, 8] in Fig.6. We can see that the estimated blur maps of our method are the most accurate and closest to the ground truth. It works with different blur types (object motion in the first three rows, defocus in last four rows), and with complex scenes and multiple objects (second, fourth, seventh, and eighth rows). For the homogeneous regions, baselines show some erroneous estimation results due to the insufficient textures in such regions, while our method avoid this problem by estimating blur map with multiple scale features using the pyramid pooling module. More visual results comparison will be shown in the supplementary material.

5.2. Evaluations on SmartBlur Dataset

Experiment Settings. We now evaluate the performance of ABC-Fusenet on our SmartBlur dataset for the tasks of both blur map estimation and blur desirability classification. As described in Section 3, SmartBlur is a large-scale blur image dataset containing $10,000$ blur images from different blur sources and blur levels, with the annotations of both pixel-levle blur amount and image-levle blur type.

Experimental Results on Blur Map Estimation. The experiments on SmartBlur dataset including two tasks: blur map estimation and image blur type classification. We compare the performance of the first task using blur map estimation branch before joint training with the state-of-the-art baseline methods [23, 24, 21]. For quantitative comparison, we utilize the average precision (AP) by averaging the precision over all recall levels. As most of the baselines are designed for blur existence estimation (without estimating blur severity), for a fair comparison, we binarize the ground truth blur map and compute the precision-recall by varying the threshold for all the methods. The AP for our method and baselines are 0.822, 0.616, 0.607, and 0.785 respectively. Our method outperforms all the baseline methods with a large margin, verifying the efficacy of ABC-FuseNet to detect blurs from different levels and sources. For qualitative comparison, We show visual results of some challenging images in SmartBlur for ABC-FuseNet and the baseline methods [23, 24, 21] in Figure. 7. These images have blurs from different sources (defocus, camera shake,

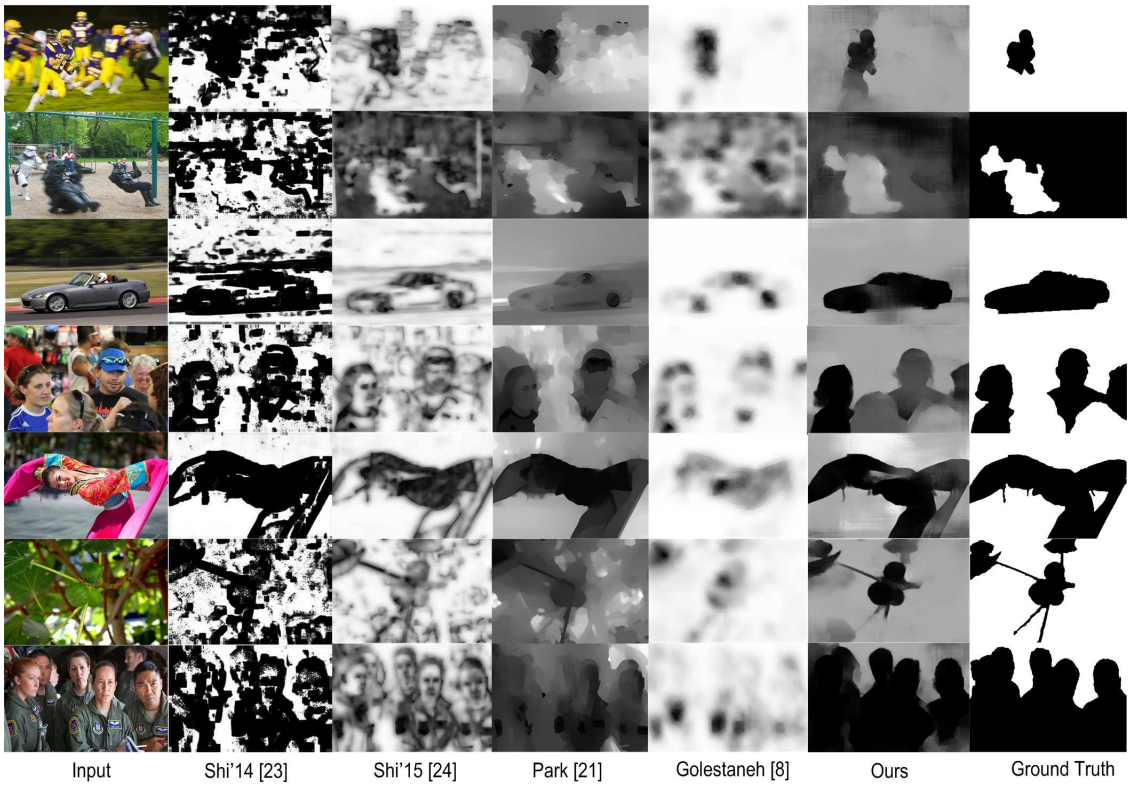

Figure 6. Visual comparison of blue map estimation on CUHK. The blurred regions have higher intensities than the clear ones.

or object motion) and amounts (low, medium, or high). The results further demonstrate that our method can produce high-quality blur maps with accurate boundaries. Furthermore, our method can estimate different blur amounts that are consisent with ground-truth annotations (third row). An interesting observation is that for the image blur from camera shake (second row), all the baselines fail to detect the uniform blur over the whole image. Baselines [3, 23, 21] tends to output high responses based on the object features, instead of blur amount. Baseline [24] mistakenly estimate the whole image as a clear one. By contrast, our method is robust to different blur sources and can detect the uniform camera-shake blurs over the whole image.

Baseline Methods for Image Blur Classification. To verify the effectiveness of the proposed methods, we extensively compare ABC-FuseNet with the state-of-the-art methods and conduct thorough ablation studies. Here we introduce the baselines: *Baseline 1: Direct classification with CNN* [34]. Yu. et al [34] build a classifier based on GoogLeNet [10] to directly classify if the image has undesired blur. Considering our ABC-FuseNet extracts content features from $ResNet-50$, for a fair comparison, we following the idea in [34] and replace the base net of *Baseline* 1 with $ResNet-50$. We finetune the network with blur category supervision from SmartBlur. Detailed network architecture is in the supplementary material.

To verify the efficacy of fusing low-level blur estimation and high-level understanding of important image content for the image blur categorization, we build another four baselines based on the different combinations of the blur map (B_m), saliency map (S_m), and content feature map (C_m) to conduct extensive ablation studies. Take *Baseline 5* as an example, we show its framework in Fig. 8. Other baselines share the same pipeline with different combination of the blur map, saliency map, and content feature map. The combined maps are fed to a light network to perform the final image blur categorization. Here we summarize the configuration of different baselines: *Baseline 2*: B_m; *Baseline 3*: B_m+C_m; *Baseline 4*: B_m + S_m; *Baseline 5*: B_m+C_m + S_m. All the baselines separately generate blur map, saliency map, or content feature map, and then perform blur type classification. Such two-stage treatment is to provide a comparison with the proposed end-to-end trainable ABC-FuseNet. To be specific, saliency map is generated by training the attention map estimation branch of ABC-FuseNet on the salient object segmentation datasets [35]. Blur map is generated by training the blur map estimation branch of ABC-FuseNet on the SmartBlur dataset, with the initial weights pretrained on the salient object segmentation datasets [35]. Content feature map is extracted from $res5c$ of $ResNet-50$ [9].

Experimental Results for Image Blur Classification. For quantitative analysis, we compare the classification accuracy of ABC-FuseNet and baselines in Fig. 9. From the

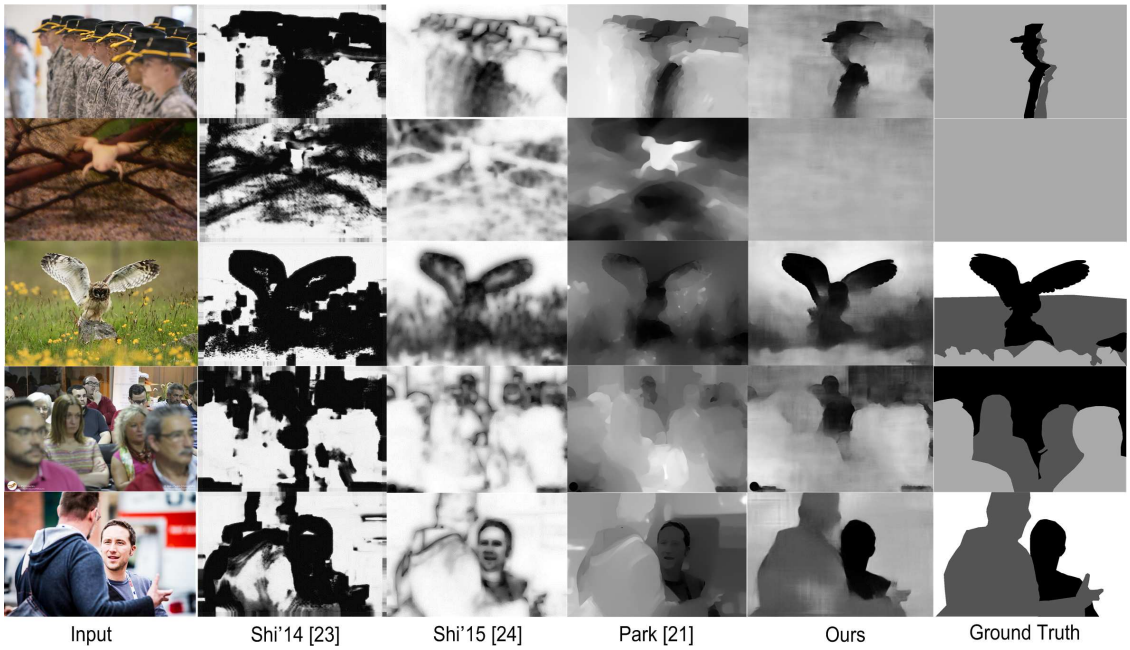

Figure 7. Visual comparison of blue map estimation on SmartBlur. The blurred regions have higher intensities than the unblurred ones.

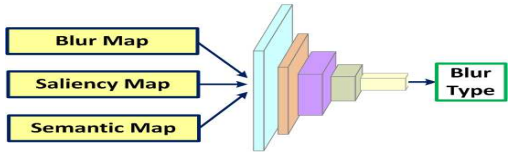

Figure 8. Framework of *Baseline 5*.

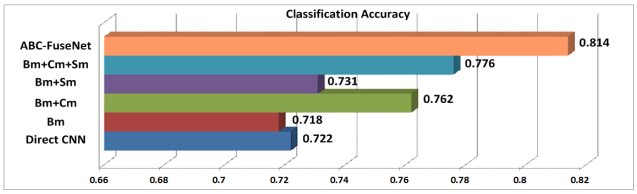

Figure 9. Comparison of image blur classification accuracy.

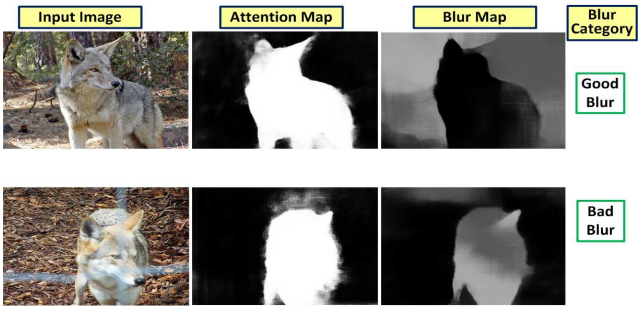

Figure 10. Results visualization of ABC-FuseNet.

results we see that, ABC-FuseNet achieves the accuracy of 0.814, outperforming all the baselines by a large margin. The poor performance of *Baseline 1: Direct CNN* and *Baseline 2*: B_m implies the necessity to combine low-level blur responses with high-level semantics for image blur categorization. When combining B_m and C_m together, the performance obtain large improvement, from around 0.72 to 0.762. *Baseline 4*: $B_m + S_m$ is more accurate than *Baseline 2*: B_m, verifying that the salient map helps better localize the important content and understand the image blur. *Baseline 5*: $B_m + C_m + S_m$ outperforms *Baseline 1* to *Baseline 4*, but it is less accurate than ABC-FuseNet, proving that joint the training of the whole network significantly improve the blur classification accuracy. For qualitative analysis, we visualize the estimated blur map and attention map, and the classification results in Fig. 10. Our model correctly classified the desirability in both cases, because of its understanding on the important content in the image, as demonstrated in the attention maps.

6. Conclusions

In this paper, we introduce the problem of automatically understanding image blur in terms of image quality and decompose this problem into two steps: generating spatially-variant blur responses, and understanding if such responses are desired by distilling high-level image semantics. We propose an end-to-end trainable ABC-FuseNet to jointly estimate blur map, attention map, and semantic map, and fuse three maps to perform final classification. We also propose a new dataset-SmartBlur, containing 10,000 natural photos with elaborate human annotations of both pixel-level blur amount and image-level blur desirability. The proposed methods significantly outperform all the baselines for the tasks of both blur map estimation and blur classification.

References

[1] A. Agrawal and R. Raskar. Optimal single image capture for motion deblurring. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2560–2567. IEEE, 2009. 2

[2] S. Bae and F. Durand. Defocus magnification. In *Computer Graphics Forum*, volume 26, pages 571–579. Wiley Online Library, 2007. 1, 6

[3] A. Chakrabarti, T. Zickler, and W. T. Freeman. Analyzing spatially-varying blur. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2512–2519. IEEE, 2010. 6, 7

[4] J. Chen, L. Yuan, C.-K. Tang, and L. Quan. Robust dual motion deblurring. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 1

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016. 4

[6] J. Du, S. Zhang, G. Wu, J. M. Moura, and S. Kar. Topology adaptive graph convolutional networks. *arXiv preprint arXiv:1710.10370*, 2017. 2

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 4

[8] S. A. Golestaneh and L. J. Karam. Spatially-varying blur detection based on multiscale fused and sorted transform coefficients of gradient magnitudes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 6

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 5, 7

[10] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015. 7

[11] N. Joshi, S. B. Kang, C. L. Zitnick, and R. Szeliski. Image deblurring using inertial measurement sensors. In *ACM Transactions on Graphics (TOG)*, volume 29, page 30. ACM, 2010. 1

[12] N. Joshi, R. Szeliski, and D. J. Kriegman. Psf estimation using sharp edge prediction. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 2

[13] D. Kinga and J. B. Adam. A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 5

[14] I. Kokkinos. Pushing the boundaries of boundary detection using deep learning. *arXiv preprint arXiv:1511.07386*, 2015. 4

[15] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision (ECCV)*, 2016. 1

[16] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1346–1353. IEEE, 2012. 1

[17] R. Liu, Z. Li, and J. Jia. Image partial blur detection and classification. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 6

[18] L. Mai and F. Liu. Kernel fusion for better image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 371–380, 2015. 3

[19] E. Mavridaki and V. Mezaris. No-reference blur assessment in natural images using fourier transform and spatial pyramids. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 566–570. IEEE, 2014. 2, 3

[20] A. L. Mendelson and Z. Papacharissi. Look at us: Collective narcissism in college student facebook photo galleries. *The networked self: Identity, community and culture on social network sites*, 1974:1–37, 2010. 1

[21] J. Park, Y.-W. Tai, D. Cho, and I. S. Kweon. A unified approach of multi-scale deep and hand-crafted features for defocus estimation. *arXiv preprint arXiv:1704.08992*, 2017. 1, 2, 6, 7

[22] Q. Shan, J. Jia, and A. Agarwala. High-quality motion deblurring from a single image. In *Acm transactions on graphics (tog)*, volume 27, page 73. ACM, 2008. 1

[23] J. Shi, L. Xu, and J. Jia. Discriminative blur detection features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2965–2972, 2014. 1, 2, 3, 5, 6, 7

[24] J. Shi, L. Xu, and J. Jia. Just noticeable defocus blur detection and estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 657–665, 2015. 2, 6, 7

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[26] B. Su, S. Lu, and C. L. Tan. Blurred image region detection and classification. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1397–1400. ACM, 2011. 6

[27] S. Suwajanakorn, C. Hernandez, and S. M. Seitz. Depth from focus with your mobile phone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2015. 2

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 4

[29] C. Tang, C. Hou, and Z. Song. Defocus map estimation from a single image via spectrum contrast. *Optics letters*, 38(10):1706–1708, 2013. 2, 6

[30] C. Tang, J. Wu, Y. Hou, P. Wang, and W. Li. A spectral and spatial approach of coarse-to-fine blurred image region detection. *IEEE Signal Processing Letters*, 23(11):1652–1656, 2016. 6

[31] Y. Wang, Z. Lin, X. Shen, R. Mech, G. Miller, and G. W. Cottrell. Recognizing and curating photo albums via event-specific image importance. *arXiv preprint arXiv:1707.05911*, 2017. 1

[32] L. Xu and J. Jia. Two-phase kernel estimation for robust motion deblurring. In *European Conference on Computer Vision*, pages 157–170. Springer, 2010. 2

[33] X. Yi and M. Eramian. Lbp-based segmentation of defocus blur. *IEEE Transactions on Image Processing*, 25(4):1626–1638, 2016. 2, 6

[34] N. Yu, X. Shen, Z. Lin, R. Mech, and C. Barnes. Learning to detect multiple photographic defects. *arXiv preprint arXiv:1612.01635*, 2016. 2, 7

[35] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech. Minimum barrier salient object detection at 80 fps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1404–1412, 2015. 5, 7

[36] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. *arXiv preprint arXiv:1612.01105*, 2016. 4

[37] L. Zhong, S. Cho, D. Metaxas, S. Paris, and J. Wang. Handling noise in single image deblurring using directional filters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 612–619, 2013. 1

[38] S. Zhuo and T. Sim. Defocus map estimation from a single image. *Pattern Recognition*, 44(9):1852–1858, 2011. 1, 2, 6