DeblurGAN-v2: Deblurring (Orders-of-Magnitude) Faster and Better

Orest Kupyn[1, 3], Tetiana Martyniuk[1], Junru Wu[2], Zhangyang Wang[2]

[1] Ukrainian Catholic University, Lviv, Ukraine; [3] SoftServe, Lviv, Ukraine
{kupyn, t.martynyuk}@ucu.edu.ua

[2] Department of Computer Science and Engineering, Texas A&M University
{sandboxmaster, atlaswang}@tamu.edu

Abstract

We present a new end-to-end generative adversarial network (GAN) for single image motion deblurring, named DeblurGAN-v2, which considerably boosts state-of-the-art deblurring efficiency, quality, and flexibility. DeblurGAN-v2 is based on a relativistic conditional GAN with a double-scale discriminator. For the first time, we introduce the Feature Pyramid Network into deblurring, as a core building block in the generator of DeblurGAN-v2. It can flexibly work with a wide range of backbones, to navigate the balance between performance and efficiency. The plug-in of sophisticated backbones (e.g., Inception-ResNet-v2) can lead to solid state-of-the-art deblurring. Meanwhile, with light-weight backbones (e.g., MobileNet and its variants), DeblurGAN-v2 reaches 10-100 times faster than the nearest competitors, while maintaining close to state-of-the-art results, implying the option of real-time video deblurring. We demonstrate that DeblurGAN-v2 obtains very competitive performance on several popular benchmarks, in terms of deblurring quality (both objective and subjective), as well as efficiency. Besides, we show the architecture to be effective for general image restoration tasks too. Our codes, models and data are available at: `https://github.com/KupynOrest/DeblurGANv2`.

1. Introduction

This paper focuses on the challenging setting of single-image blind motion deblurring. Motion blurs are commonly found from photos taken by hand-held cameras, or low-frame-rate videos containing moving objects. Blurs degrade the human perceptual quality, and challenge subsequent computer vision analytics. The real-world blurs typically have unknown and spatially varying blur kernels, and are further complicated by noise and other artifacts.

The recent prosperity of deep learning has led to significant progress in the image restoration field [48, 28]. Specifically, generative adversarial networks (GANs) [9] often yield sharper and more plausible textures than classical feed-forward encoders and witness success in image super-resolution [23] and in-painting [53]. Recently, [21] introduced GAN to deblurring by treating it as a special image-to-image translation task [13]. The proposed model, called *DeblurGAN*, was demonstrated to restore perceptually pleasing and sharp images, from both synthetic and real-world blurry images. DeblurGAN was also 5 times faster than its closest competitor as of then [33].

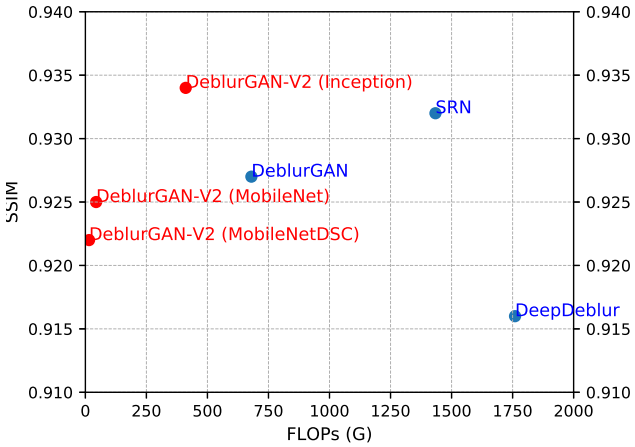


Figure 1: The SSIM-FLOPs trade-off plot on the GoPro dataset. Compared to three state-of-the-art competitors (in blue): DeblurGAN [21], DeepDeblur [33] and Scale-Recurrent Network (SRN) [45], DeblurGAN-v2 models (with different backbones, in red) are shown to achieve superior or comparable quality, and are much more efficient.

Built on the success of DeblurGAN, this paper aims to make another substantial push on GAN-based motion deblurring. We introduce a new framework to improve over DeblurGAN, called **DeblurGAN-v2** in terms of both deblurring performance and inference efficiency, as well as to enable high flexibility over the quality- efficiency spectrum. Our innovations are summarized as below[1]:

[1] An informal note: we quite like the sense of humor in [38], quoted

- **Framework Level:** We construct a new conditional GAN framework for deblurring. For the generator, we introduce the Feature Pyramid Network (FPN), which was originally developed for object detection [27], to the image restoration task for the first time. For the discriminator, we adopt a relativistic discriminator [16] with a least-square loss wrapped [30] inside, and with two columns that evaluate both global (image) and local (patch) scales respectively.
- **Backbone Level:** While the above framework is agnostic to the generator backbones, the choice would affect deblurring quality and efficiency. To pursue the state-of-the-art deblurring quality, we plug in a sophisticated Inception-ResNet-v2 backbone. To shift towards being more efficient, we adopt MobileNet, and further create its variant with depth-wise separable convolutions (MobileNet-DSC). The latter two become extremely compact in size and fast at inference.
- **Experiment Level:** We present very extensive experiments on three popular benchmarks to show the state-of-the-art (or close) performance (PSNR, SSIM, and perceptual quality) achieved by DeblurGAN-v2. In terms of the efficiency, DeblurGAN-v2 with MobileNet-DSC is **11 times** faster than DeblurGAN [21], over **100 times** faster than [33, 45], and has a model size of just **4 MB**, implying the possibility of real-time video deblurring. We also present a subjective study of the deblurring quality on real blurry images. Lastly, we show the potential of our models in general image restoration, as extra flexibility.

2. Related work

2.1. Image Deblurring

Single image motion deblurring is traditionally treated as a deconvolution problem, and can be tackled in either a blind or a non-blind manner. The former assumes a given or pre-estimated blur kernel [39, 52]. The latter is more realistic yet highly ill-posed. Earlier models rely on natural image priors to regularize deblurring [20, 36, 25, 5]. However, most handcrafted priors cannot well capture the complicated blur variations in real images.

Emerging deep learning techniques have boosted the breakthrough in image restoration tasks. Sun *et al*. [43] exploited a convolutional neural network (CNN) for blur kernel estimation. Gong *et al*. [8] used a fully convolutional network to estimate the motion flow. Besides those kernel-based methods, end-to-end kernel-free CNN methods were explored to restore a clean image from the blurry input directly, e.g., [33, 35]. The latest work by Tao *et al*. [45] extended the Multi-Scale CNN from [33] to a Scale-Recurrent CNN for blind image deblurring, with impressive results.

The success of GANs for image restoration has impacted single image deblurring as well since Ramakrishnan *et al*. [37] first solved image deblurring by referring to the image translation idea [13]. Lately, Kupyn *et al*. [21] introduced *DeblurGAN* that exploited Wasserstein GAN [2] with the gradient penalty [10] and the perceptual loss [15].

2.2. Generative adversarial networks

A GAN [9] consists of two models: a discriminator D and a generator G, that form a two-player minimax game. The generator learns to produce artificial samples and is trained to fool the discriminator, in a goal to capture the real data distribution. In particular, as a popular GAN variant, conditional GANs [31] have been widely applied to image-to-image translation problems, with image restoration and enhancement as special cases. They take the label or an observed image in addition to the latent code as inputs.

The minimax game with the value function $V(D,G)$ is formulated as the following [9] (fake-real labels set to $0-1$):

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}\big[\log D(x)\big] + \mathbb{E}_{z\sim p_z(z)}\big[\log(1-D(G(z)))\big]$$

Such an objective function is notoriously hard to optimize, and one needs to deal with many challenges, e.g., mode collapse and gradient vanishing/explosion, during the training process. To fix the vanishing gradients and stabilize the training, Least Squares GANs discriminator [30] tried to introduce a loss function that provides smoother and non-saturating gradient. The authors observe that the log-type loss in [9] saturates quickly as it ignores the distance between x to the decision boundary. In contrast, an $L2$ loss provides gradients proportional to that distance, so that fake samples more far away from the boundary receive larger penalties. The proposed loss function also minimizes the Pearson χ^2 divergence that leads to the better training stability. The LSGAN objective function is written as::

$$\begin{aligned}\min_D V(D) &= \frac{1}{2}\mathbb{E}_{x\sim p_{data}(x)}\big[(D(x)-1)^2\big] \\ &\quad + \frac{1}{2}\mathbb{E}_{z\sim p_z(z)}\big[D(G(z))^2\big] \\ \min_G V(G) &= \frac{1}{2}\mathbb{E}_{z\sim p_z(z)}\big[(D(G(z))-1)^2\big]\end{aligned} \tag{1}$$

Another relevant improvement to GANs is the Relativistic GAN [16]. It used a relativistic discriminator to estimate the probability that the given real data is more realistic than a randomly sampled fake data. As the author advocated, such would account for a priori knowledge that half of the

as: "We present some updates to YOLO. We made a bunch of little design changes to make it better. We also trained this new network that's pretty swell." – that well describes what we have done to DeblurGAN, too; although we consider DeblurGAN-v2 a non-incremental upgrade of DeblurGAN, with significant performance & efficiency improvements.

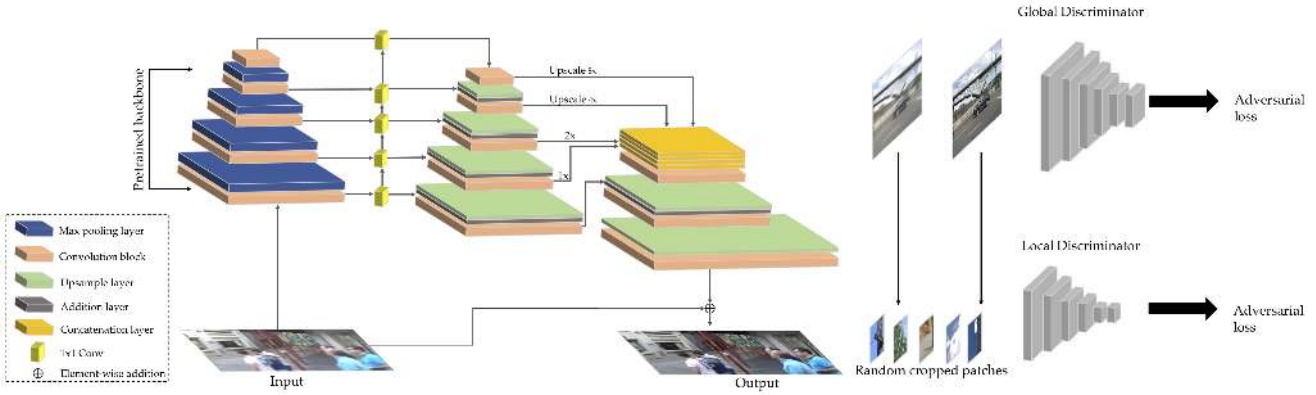


Figure 2: DeblurGAN-v2 pipeline architecture.

data in the mini-batch is fake. The relativistic discriminators show more stable and computationally efficient training in comparison to other GAN types, including WGAN-GP [10] that was used in DeblurGAN-v1.

3. DeblurGAN-v2 Architecture

The overview of DeblurGAN-v2 architecture is illustrated in Figure 2. It restores a sharp image I_S from a single blurred image I_B, via the trained generator.

3.1. Feature Pyramid Deblurring

Existing CNNs for image deblurring (and other restoration problems) [23, 33] typically refer to ResNet-like structures. Most state-of-the-art methods [33, 45] dealt with different levels of blurs, utilizing multi-stream CNNs with an input image pyramid at different scales. However, processing multiple scale images is time-consuming and memory-demanding. We introduce the idea of Feature Pyramid Networks [27] to image deblurring (more generally, the field of image restoration and enhancement), *for the first time to our best knowledge*. We treat this novel approach as a lighter-weight alternative to incorporate multi-scale features.

The FPN module was originally designed for object detection [27]. It generates multiple feature map layers which encode different semantics and contain better quality information. FPN comprises a bottom-up and a top-down pathway. The bottom-up pathway is the usual convolutional network for feature extraction, along which the spatial resolution is downsampled, but more semantic context information is extracted and compressed. Through the top-down pathway, FPNs reconstructs higher spatial resolution from the semantically rich layers. The lateral connections between the bottom-up and top-down pathways supplement high-resolution details and help localize objects.

Our architecture consists of an FPN backbone from which we take five final feature maps of different scales as the output. Those features are later upsampled to the same $\frac{1}{4}$ input size and concatenated into one tensor which contains the semantic information on different levels. We additionally add two upsampling and convolutional layers at the end of the network to restore the original image size and reduce artifacts. Similar to [21, 29], we introduce a direct skip connection from the input to the output, so that the learning focuses on the residue. The input images are normalized to [-1 1]. We also use a *tanh* activation layer to keep the output in the same range. In addition to the multi-scale feature aggregation capability, FPN also strikes a balance between accuracy and speed: please see experiment parts.

3.2. Choice of Backbones: Trade-off between Performance and Efficiency

The new FPN-embeded architecture is agnostic to the choice of feature extractor backbones. With this plug-and-play property, we are entitled with the flexibility to navigate through the spectrum of accuracy and efficiency. By default, we choose ImageNet-pretrained backbones to convey more semantic-related features. As one option, we use **Inception-ResNet-v2** [44] to pursue strong deblurring performance, although we find other backbones such as SE-ResNeXt [12] to be similarly effective.

The demands of efficient restoration model have recently drawn increasing attentions due to the prevailing need of mobile on-device image enhancement [54, 50, 47]. To explore this direction, we choose the **MobileNet** V2 backbone [40] as one option. To reduce the complexity further, we try another more aggressive option on top of DeblurGAN-v2 with MobileNet V2, by replacing all normal convolutions in the *full network* (including those not in backbone) with Depthwise Separable Convolutions [6]. The resulting model is denoted as **MobileNet-DSC**, and can provide extremely lightweight and efficient image deblurring.

To unleash this important flexibility to practitioners, in our codes, we have implemented the switch of backbones as a simple *one-line command*: it can be compatible with

many state-of-the-art pre-trained networks.

3.3. Double-Scale RaGAN-LS Discriminator

Instead of the WGAN-GP discriminator in DeblurGAN [21], we suggest several upgrades in DeblurGAN-v2. We first adopt the relativistic "wrapping" [16] on the LSGAN [30] cost function, creating a new *RaGAN-LS* loss:

$$L_D^{RaLSGAN} = \mathbb{E}_{x\sim p_{data}(x)}\big[(D(x) - \mathbb{E}_{z\sim p_z(z)}D(G(z)) - 1)^2\big] + \mathbb{E}_{z\sim p_z(z)}\big[(D(G(z)) - \mathbb{E}_{x\sim p_{data}(x)}D(x) + 1)^2\big] \quad (2)$$

It is observed to make training notably faster and more stable compared to using the WGAN-GP objective. We also empirically conclude that the generated results possess higher perceptual quality and overall sharper outputs. Correspondingly, the adversarial loss L_{adv} for the DeblurGAN-v2 generator will be optimizing (2) w.r.t. G.

Extending to Both Global and Local Scales. Isola *et al*. [13] propose to use a PatchGAN discriminator which operates on the images patches of size 70×70, that proves to produce sharper results than the standard "global" discriminator that operates on the full image. The PatchGAN idea was adopted in DeblurGAN [21].

However, we observed that for highly non-uniform blurred images, especially when complex object movements are involved, the "global" scales are still essential for discriminators to incorporate full spatial contexts [14]. To take advantage of both global and local features, we propose to use a double-scale discriminator, consisting of one local branch that operates on patch levels like [13] did, and the other global branch that feeds the full input image. We observe that to allow DeblurGAN-v2 to better handle larger and more heterogeneous real blurs.

Overall Loss Function For training image restoration GANs, one needs to compare the images on the training stage the reconstructed and the original ones, under some metric. One common option is the pixel-space loss L_P, e.g., the simplest L_1 or L_2 distance. As [23] suggested, using L_p tends to yield oversmoothened pixel-space outputs. [21] proposed to use the perceptual distance [15], as a form of "content" loss L_X. In contrast to the L_2, it computes the Euclidean loss on the VGG19 [41] *conv3_3* feature maps. We incorporate those prior wisdoms and use a hybrid three-term loss for training DeblurGAN-v2:

$$L_G = 0.5 * L_p + 0.006 * L_X + 0.01 * L_{adv}$$

The L_{adv} terms contains both global and local discriminator losses. Also, we choose mean-square-error (MSE) loss as L_p: although DeblurGAN did not include an L_p term, we find it to help correct color and texture distortions.

3.4. Training Datasets

The **GoPro** dataset [33] uses the GoPro Hero 4 camera to capture 240 frames per second (fps) video sequences, and generate blurred images through averaging consecutive short-exposure frames. It is a common benchmark for image motion blurring, containing 3,214 blurry/clear image pairs. We follow the same split [33], to use 2,103 pairs for training and the remaining 1,111 pairs for evaluation.

(a) (b) (c) (d)

Figure 3: Visual comparison of synthesized blurry images, without interpolation (a,c) and with interpolation (b,d).

The **DVD** dataset [42] collects 71 real-world videos captured by various devices such as iPhone 6s, GoPro Hero 4 and Nexus 5x, at 240 fps. The author then generated 6708 synthetic blurry and sharp pairs by averaging consecutive short-exposure frames to approximate a longer exposure [46]. The dataset was initially used for video deblurring but was later also brought to the image deblurring field.

The **NFS** dataset [17] was initially proposed to benchmark visual object tracking. It consists of 75 videos captured with high-frame rate cameras from iPhone 6 and iPad Pro. Additionally, 25 sequences are collected from YouTube captured at 240 fps from a variety of different devices. It covers variety of scenes including sport, skydiving, underwater, wildlife, roadside, and indoor scenes.

Training data preparation: Conventionally, the blurry frames are averaged from consecutive clean frames. However, we notice unrealistic ghost effects when observing the directly averaged frames, as in Figure 3(a)(c). To alleviate that, we first use a video frame interpolation model [34] to increase the original 240-fps videos to 3840 fps, then perform average pooling over the same time window (but now with more frames). It leads to smoother and more continuous blurs, as in Figure 3(b)(d). Experimentally, this data preparation did not noticeably impact PSNR/SSIM but was observed to improve the visual quality results.

4. Experimental evaluation

4.1. Implementation Details

We implemented all of our models using PyTorch [1]. We compose our training set by selecting each second frame from the GoPro and DVD datasets, and every tenth frame from the NFS dataset, with the hope to reduce overfitting to any specific dataset. We then train DeblurGAN-v2 on the resulting set of approximately 10,000 image

Table 1: Performance and efficiency comparison on the GoPro test dataset, All models were tested on the *linear* image subset.

	Sun *et al*. [43]	Xu *et al*. [51]	DeepDeblur [33]	SRN [45]	DeblurGAN [21]	Inception-ResNet-v2	MobileNet	MobileNet-DSC
PSNR	24.64	25.10	29.23	**30.10**	28.70	29.55	28.17	28.03
SSIM	0.842	0.890	0.916	0.932	0.927	**0.934**	0.925	0.922
Time	20 min	13.41s	4.33s	1.6s	0.85s	0.35s	0.06s	**0.04s**
FLOPS	N/A	N/A	1760.04G	1434.82G	678.29G	411.34G	43.75G	**14.83G**

Table 2: PSNR and SSIM comparison on the Kohler dataset.

Method	Sun [43]	DeepDeblur [33]	SRN [45]	DeblurGAN [21]	Inception-ResNet-v2	MobileNet	MobileNet-DSC
PSNR	25.22	26.48	26.75	26.10	26.72	26.36	26.35
SSIM	0.773	0.807	0.837	0.816	0.836	0.820	0.819

pairs. Three backbones are evaluated: Inception-ResNet-v2, MobileNet, and MobileNet-DSC. The former targets at high-performance deblurring, while the latter two are more suited for resource-constrained edge applications. Specifically, the extremely lightweight DeblurGAN-v2 (MobileNet-DSC) costs 96% fewer parameters than DeblurGAN-v2 (Inception-ResNet-v2).

All models were trained on a single Tesla-P100 GPU, with Adam [18] optimizer and the learning rate of 10^{-4} for 150 epochs, followed by another 150 epochs with a linear decay to 10^{-7}. We freeze the pre-trained backbone weights for 3 epochs, and then we unfreeze all weights and continue the training. The un-pre-trained parts are initialized with random Gaussian. The training takes 5 days to converge. The models are fully convolutional, thus can be applied to the images of arbitrary size.

4.2. Quantitative Evaluation on GoPro Dataset

We compare our models with a number of state-of-the-arts: one of is a traditional method by Xu *et al*. [51], while the rest are deep learning-based: [43] by Sun *et al*., DeepDeblur [33], SRN [45], and DeblurGAN [21]. We compare on both standard performance metrics (PSNR, SSIM), and inference efficiency (averaged running time per image measured on a single GPU). Results are summarized in Table1.

In terms of PSNR/SSIM, DeblurGAN-v2 (Inception-ResNet-v2) and SRN are ranked top-2: DeblurGAN-v2 (Inception-ResNet-v2) has slightly lower PSNR, which is not surprising since it was not trained under pure MSE loss; but it outperforms SRN in SSIM. However, we are very encouraged to observe that DeblurGAN-v2 (Inception-ResNet-v2) takes **78% less** inference time than SRN. Moreover, two of our light-weight models, DeblurGAN-v2 (MobileNet) and DeblurGAN-v2 (MobileNet-DSC), show SSIMs (0.925 and 0.922) on par with the other two latest deep deblurring methods, DeblurGAN (0.927) and DeepDeblur (0.916), while being up to **100 times faster**.

In particular, MobileNet-DSC only costs 0.04s per image, which even enables near real-time video frame deblurring, for 25-fps videos. To our best knowledge, DeblurGAN-v2 (MobileNet-DSC) is the only deblurring method so far that can simultaneously achieve (reasonably) high performance and that high inference efficiency.

Table 3: Results on DVD dataset

	PSNR	SSIM	Inference Time	Resolution
WFA	28.35	N/A	N/A	N/A
DVD (single)	28.37	0.913	1.0s	960 x 540
DeblurGAN-v2 (MobileNet)	**28.54**	**0.929**	**0.06s**	1280 x 720

4.3. Quantitative Evaluation on Kohler dataset

The Kohler dataset [19] consists of 4 images, each blurred with 12 different kernels. It is a standard benchmark for evaluating blind deblurring algorithms. The dataset was generated by recording and analyzing real camera motion, which was then played back on a robot platform such that a sequence of sharp images was recorded sampling the 6D camera motion trajectory.

The comparison results are reported in Table 2. Similarly to GoPro, SRN and DeblurGAN-v2 (Inception-ResNet-v2) remain to be the best two PSNR/SSIM performers, but this time SRN is marginally superior in both. However, please be reminded that, similarly to the GoPro case, this "almost tie" result was achieved while DeblurGAN-v2 (Inception-ResNet-v2) costs only 1/5 of SRN's inference complexity. Moreover, both DeblurGAN-v2 (MobileNet) and DeblurGAN-v2 (MobileNet-DSC) outperform DeblurGAN on the Kohler dataset in both SSIM and PSNR: that is impressive given the former two's much lighter weights.

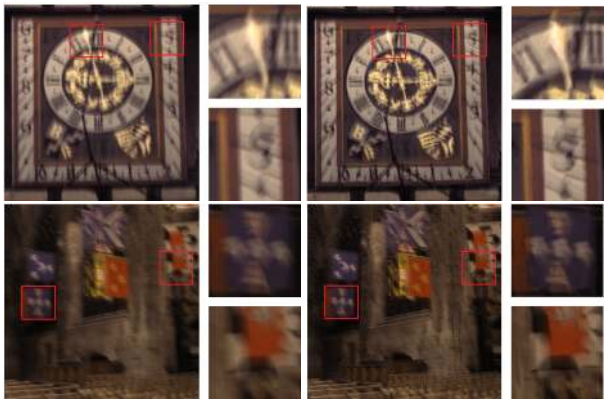
Figure 4 displays visual examples on the Kohler dataset. DeblurGAN-v2 effectively restores the edges and textures, without noticeable artifacts. SRN for this specific example shows some color artifacts when zoomed in.

4.4. Quantitative Evaluation on DVD dataset

We next test DeblurGAN-v2 on the DVD testing set used in [42], but with a *single-frame* setting (treating all frames as individual images) without using multiple frames together. We compare with two strong video deblurring meth-

Table 4: Average subjective scores of deblurring results on the Lai dataset [22].

Blurry	Krishnan *et al*. [20]	Whyte *et al*. [49]	Xu *et al*. [51]	Sun *et al*. [43]	Pan *et al*. [36]
1	1.08	0.57	0.77	0.64	0.91
DeepDeblur [33]	SRN [45]	DeblurGAN [21]	DeblurGAN-v2 (Inception-ResNet-v2)	DeblurGAN-v2 (MobileNet)	DeblurGAN-v2 (MobileNet-DSC)
1.08	1.68	1.29	**1.74**	1.44	1.32



(a) Blurry (b) DeepDeblur [33]

(c) SRN [45] (d) DeblurGAN [21]

(e) DeblurGAN-v2 (Inception-ResNet-v2) (f) DeblurGAN-v2 (MobileNet)

Figure 4: Visual comparison on the Kohler dataset.

ods: WFA [7], and DVD [42], For the latter, we adopt the authors' self-reported results when using a single frame as the model input (denoted as "single"), for a fair comparison. As shown in Table 6, DeblurGAN-v2 (MobileNet) outperforms WFA and DVD (single), while being at least 17 times faster (DVD was tested on a reduced resolution of 960×540, while DeblurGAN-v2 is on 1280 x 720).

While not specifically optimized for video deblurring, DeblurGAN-v2 shows good potential, and we will extend it to video deblurring as future work.

4.5. Subjective Evaluation on Lai dataset

The Lai dataset [22] has real-world blurry images of different qualities and resolutions collected in various types of scenes. Those real images have no clean/sharp counterparts, making a full-reference quantitative evaluation impossible. Following [22], we conduct a subjective survey to compare the deblurring performance on those real images.

We fit a Bradley-Terry model [3] to estimate the subjective score for each method so that they can be ranked, with the identical routine following the previous benchmark work [24, 26]. Each blurry image is processed with each of the following algorithms: Krishnan *et al*. [20], Whyte *et al*. [49], Xu *et al*. [51], Sun *et al*. [43], Pan *et al*. [36], DeepDeblur [33], SRN [45], DeblurGAN [21]; and the three DeblurGAN-v2 variants (Inception-ResNet-v2, MobileNet, MobileNet-DSC). The eleven deblurring results, together with the original blurry image, are sent for pairwise comparison to construct the winning matrix. We collect the pair comparison results from 22 human raters. We observed good consensus and small inter-person variances among raters, which makes scores reliable.

The subjective scores are reported in Table 4. We did not normalize the scores due to the absence of ground-truth: as a result, it is the score rank rather than the absolute score value that matters here. It can be observed that deep learning-based deblurring algorithms, in general, have more favorable visual results than traditional methods (some even making visual quality worse than the blurry input). DeblurGAN [21] outperforms DeepDeblur [33], but lags behind SRN [45]. With the Inception-ResNet-v2 backbone, DeblurGAN-v2 demonstrates clearly superior perceptual quality over SRN, making it the top performer in terms of subjective quality. DeblurGAN-v2 with MobileNet and MobileNet-DSC backbones have minor performance degradations compared to the Inception-ResNet-v2

(a) Blurred photo (b) Whyte *et al*. [49] (c) Krishnan *et al*. [20] (d) Sun *et al*. [43]

(e) Xu *et al*. [51] (f) Pan *et al*. [36] (g) DeepDeblur [33] (h) SRN [45]

(i) DeblurGAN [21] (j) DeblurGAN-v2 (Inception-ResNet-v2) **[Best visual quality]** (k) DeblurGAN-v2 (MobileNet) **[High efficiency]** (l) DeblurGAN-v2 (MobileNet-DSC) **[Highest efficiency]**

Figure 5: Qualitative comparison on the "face2" test image of the Lai dataset [22]. DeblurGAN-v2 models are artifact-free, in contrast to other neural and non-CNN algorithms, producing smoother and visually more pleasing results.

version. However, both are still preferred by subjective raters, compared to DeepDeblur and DeblurGAN, while being 2-3 orders-of-magnitude faster.

Figure 5 displays visual comparison examples on deblurring the "face2" image. DeblurGAN-v2 (Inception-ResNet-v2) (5j) and SRN (5h) are the top-2 most favored results, both balancing well between edge-sharpness and overall smoothness. By zooming in, SRN is found to still generate some ghost artifacts on this example, e.g., the white "intrusion" from the collar to the bottom right face region. In

(a) Degraded photo (b) DeblurGAN (c) DeblurGAN-v2 (Inception-ResNet-v2) (d) Clean photo

Figure 6: Visual comparison example on the Restore Dataset.

comparison, DeblurGAN-v2 (Inception-ResNet-v2) shows artifact-free deblurring. Besides, DeblurGAN-v2 (MobileNet) and DeblurGAN-v2 (MobileNet-DSC) results are also smooth and visually better than DeblurGAN, though less sharper than DeblurGAN-v2 (Inception-ResNet-v2).

4.6. Ablation Study and Analysis

We perform an ablation study on the effect of specific components of the DeblurGAN-v2 pipeline. Starting from the original DeblurGAN (ResNet G, local-scale patch D, WGAN-GP + perceptual loss), we gradually inject our modifications on the generator (adding FPN), discriminator (adding global-scale), and the loss (replacing WGAN-GP loss with RaGAN-LS, and adding an MSE term). The results are summarized in Table 6. We can see that all our proposed components steadily improve both PSNR and SSIM. In particular, the FPN module contributes most significantly. Also, adding either MSE or perceptual loss benefits both training stability and final results.

Table 5: Ablation Study on the GoPro dataset, based on DeblurGAN-v2 (Inception-ResNet-v2).

	PSNR	SSIM
DeblurGAN (starting point)	28.70	0.927
+ FPN	29.26	0.931
+ FPN + Global D	29.29	0.932
+ FPN + Global D + RaGAN-LS	29.37	0.933
DeblurGAN-v2 (FPN + Global D + RaGAN-LS + MSE Loss)	**29.55**	**0.934**
Removing perceptual loss (replace 0.5 with 0 in L_G)	28.81	0.924

As an extra baseline for the efficiency of FPN, we tried to create a "compact" version of SRN, with roughly the same FLOPs (456 GFLOPs) to match DeblurGAN-v2 Inception-ResNet-v2 (411 GFLOPs). We reduced the numbers of ResBlocks by 2/3 in each EBlock/DBlock while keeping their 3-scale recurrent structure. We then compare with DeblurGAN-v2 (Inception-ResNet-v2) on GoPro, where that "compact" SRN only achieved PSNR = 28.92 dB and SSIM = 0.9324. We also tried channel pruning [11] to reduce SRN FLOPs and the result was no better.

Table 6: PSNR/SSIM comparison on Restore Dataset.

	PSNR	SSIM
Degraded	22.056	0.873
DeblurGAN	26.435	0.892
DeblurGAN-v2 (Inception-ResNet-v2)	**26.916**	**0.894**
DeblurGAN-v2 (MobileNet-DSC)	25.412	0.891

4.7. Extension to General Restoration

Real-world natural images commonly go through multiple kinds of degradations (noise, blur, compression, etc.) at once, and a few recent works were devoted to such join enhancement tasks [32, 55] We study the effect of DeblurGAN-v2 on the task of general image restoration. While **NOT** being the main focus of this paper, we intend to show the general architecture superiority of DeblurGAN-v2, especially for modifications made w.r.t. DeblurGAN.

We synthesize a new challenging *Restore Dataset*. We take 600 images from GoPro, and 600 images from DVD, both with motion blurs already (same as above). We then use the *albumentations* library [4] to further add Gaussian and speckle Noise, JPEG compression, and up-scaling artifacts to those images. Eventually, we split 8000 images for training and 1200 for testing. We train and compare DeblurGAN-v2 (Inception-ResNet-v2), DeblurGAN-v2 (MobileNet-DSC), and DeblurGAN. As shown in Table 6 and Fig. 6, DeblurGAN-v2 (Inception-ResNet-v2) achieves the best PSNR, SSIM, and visual quality.

5. Conclusion

This paper introduces DeblurGAN-v2, a powerful and efficient image deblurring framework, with promising quantitative and qualitative results. DeblurGAN-v2 enables to switch between different backbones, for flexible trade-offs between performance and efficiency. We plan to extend DeblurGAN-v2 for real-time video enhancement, and for better handling mixed degradations.

Acknowledgements: O. Kupyn was supported by SoftServe, T. Martyniuk - by Let's Enhance and Eleks. J. Wu and Z. Wang were supported by NSF Award RI-1755701. The authors thank Arseny Kravchenko, Andrey Luzan and Yifan Jiang for constructive discussions, and Igor Krashenyi and Oles Dobosevych for computational resources.

References

[1] PyTorch. http://pytorch.org.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[3] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[4] Alexander Buslaev, Alex Parinov, Eugene Khvedchenya, Vladimir I Iglovikov, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *arXiv preprint arXiv:1809.06839*, 2018.

[5] Chia-Feng Chang and Jiunn-Lin Wu. A new single image deblurring algorithm using hyper laplacian priors. In *ICS*, pages 1015–1022, 2014.

[6] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[7] Mauricio Delbracio and Guillermo Sapiro. Burst deblurring: Removing camera shake through fourier burst accumulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2385–2393, 2015.

[8] Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian 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. 2016.

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. June 2014.

[10] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[11] Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1389–1397, 2017.

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arxiv*, 2016.

[14] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *arXiv preprint arXiv:1906.06972*, 2019.

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

[16] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

[17] Hamed Kiani Galoogahi, Ashton Fagg, Chen Huang, Deva Ramanan, and Simon Lucey. Need for speed: A benchmark for higher frame rate object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1125–1134, 2017.

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[19] 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 *Proceedings of the 12th European Conference on Computer Vision - Volume Part VII*, ECCV'12, pages 27–40, Berlin, Heidelberg, 2012. Springer-Verlag.

[20] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *CVPR 2011*, pages 233–240. IEEE, 2011.

[21] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8183–8192, 2018.

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

[23] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[24] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, 2019.

[25] Lerenhan Li, Jinshan Pan, Wei-Sheng Lai, Changxin Gao, Nong Sang, and Ming-Hsuan Yang. Learning a discriminative prior for blind image deblurring. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[26] Siyuan Li, Iago Breno Araujo, Wenqi Ren, Zhangyang Wang, Eric K Tokuda, Roberto Hirata Junior, Roberto Cesar-Junior, Jiawan Zhang, Xiaojie Guo, and Xiaochun Cao. Single image deraining: A comprehensive benchmark analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3838–3847, 2019.

[27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[28] Ding Liu, Zhaowen Wang, Yuchen Fan, Xianming Liu, Zhangyang Wang, Shiyu Chang, and Thomas Huang. Robust video super-resolution with learned temporal dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2507–2515, 2017.

[29] Ding Liu, Bihan Wen, Xianming Liu, Zhangyang Wang, and Thomas S Huang. When image denoising meets high-level vision tasks: a deep learning approach. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 842–848. AAAI Press, 2018.

[30] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, and Zhen Wang. Least squares generative adversarial networks, 2016. cite arxiv:1611.04076.

[31] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[32] Janne Mustaniemi, Juho Kannala, Jiri Matas, Simo Särkkä, and Janne Heikkilä. Lsd_2 - joint denoising and deblurring of short and long exposure images with convolutional neural networks. *CoRR*, abs/1811.09485, 2018.

[33] Seungjun Nah, Tae Hyun, Kim Kyoung, and Mu Lee. Deep Multi-scale Convolutional Neural Network for Dynamic Scene Deblurring. 2016.

[34] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 261–270, 2017.

[35] Mehdi Noroozi, Paramanand Chandramouli, and Paolo Favaro. Motion Deblurring in the Wild. 2017.

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

[37] Sainandan Ramakrishnan, Shubham Pachori, Aalok Gangopadhyay, and Shanmuganathan Raman. Deep generative filter for motion deblurring. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2993–3000, 2017.

[38] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[39] 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 *Advances in Neural Information Processing Systems*, pages 295–305, 2018.

[40] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *arXiv e-prints*, page arXiv:1801.04381, Jan 2018.

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

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

[43] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a Convolutional Neural Network for Non-uniform Motion Blur Removal. 2015.

[44] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[45] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8174–8182, 2018.

[46] Jacob Telleen, Anne Sullivan, Jerry Yee, Oliver Wang, Prabath Gunawardane, Ian Collins, and James Davis. Synthetic shutter speed imaging. In *Computer Graphics Forum*, 2007.

[47] Yue Wang, Tan Nguyen, Yang Zhao, Zhangyang Wang, Yingyan Lin, and Richard Baraniuk. Energynet: Energy-efficient dynamic inference. *NeurIPS CDNNRIA Workshop*, 2018.

[48] Zhangyang Wang, Ding Liu, Shiyu Chang, Qing Ling, Yingzhen Yang, and Thomas S Huang. D3: Deep dual-domain based fast restoration of jpeg-compressed images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2764–2772, 2016.

[49] Oliver Whyte, Josef Sivic, Andrew Zisserman, and Jean Ponce. Non-uniform deblurring for shaken images. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 491–498. IEEE, 2010.

[50] Junru Wu, Yue Wang, Zhenyu Wu, Zhangyang Wang, Ashok Veeraraghavan, and Yingyan Lin. Deep k-means: Retraining and parameter sharing with harder cluster assignments for compressing deep convolutions. In *International Conference on Machine Learning*, pages 5359–5368, 2018.

[51] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural L0 Sparse Representation for Natural Image Deblurring. 2013.

[52] Xiangyu Xu, Jinshan Pan, Yu-Jin Zhang, and Ming-Hsuan Yang. Motion blur kernel estimation via deep learning. *IEEE Transactions on Image Processing*, 27(1):194–205, 2018.

[53] Raymond A. Yeh, Chen Chen, Teck-Yian Lim, Mark Hasegawa-Johnson, and Minh N. Do. Semantic image inpainting with perceptual and contextual losses. *CoRR*, abs/1607.07539, 2016.

[54] Ke Yu, Chao Dong, Liang Lin, and Chen Change Loy. Crafting a toolchain for image restoration by deep reinforcement learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2443–2452, 2018.

[55] Xinyi Zhang, Hang Dong, Zhe Hu, Wei-Sheng Lai, Fei Wang, and Ming-Hsuan Yang. Gated fusion network for joint image deblurring and super-resolution. In *BMVC*, 2018.