Learning Blind Motion Deblurring

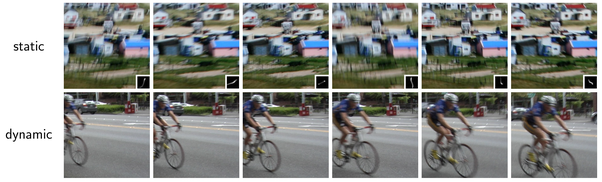
P. Wieschollek[1,2], M. Hirsch[1], B. Schölkopf[1], H. P.A. Lensch[2],

[1] Max Planck Institute for Intelligent Systems, Tübingen, [2] University of Tübingen

In a nutshell: Taking a quick shot with a camera frequently yields a blurry result due to moving objects in the scene and/or unwanted camera shake during recording. Removing these artifacts from the blurry recordings is a highly ill-posed problem as neither the sharp image nor the blur information is known. We propose a data-driven approach to restore the sharp image from a sequence of blurry observations of a *dynamic* scene that contains moving objects, i.e. the input image sequence suffers from blur due to both ego and object motion.

Problem Formulation

Given a sequence $\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3, \dots$ of blurry observations from a video stream we aim at reconstructing the latent sharp frames $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \dots$. Hereby, the blur from camera shake ([6, 3]) is superimposed with blur artifacts due to object motion.



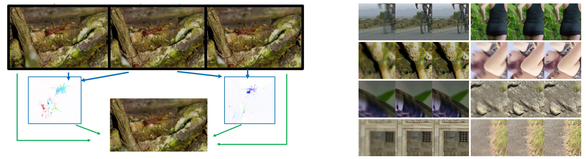
Comparison of inputs only featuring camera shake blur (top, [1, 2, 3, 5, 6]) and inputs which are further superimposed by motion blur due to moving objects in dynamic scenes (bottom, [4, 7, 8, 9])

Generating Realistic Training Data

A realistic training data should feature two aligned versions for each video frame:

- a blurry version serving as the input, and
- an associated sharp version serving as ground-truth

Obtaining this data is challenging as any recorded sequence might suffer from the described blur effects itself.
We propose to crawl YouTube videos with sharp frames and synthesize both motion blur and blur from camera shake. We employ the estimated optical flow to warp between two frames creating subframes and average those.



Generation of synthetic motion blur (left) and a snapshot from training (right) with one input, network prediction and sharp ground truth.

Our training data set consists of 5.43 hour training content without manual capturing effort. Note, [7] captured 6min video content (71 videos with 3-5sec) for training similar to [8, 9].

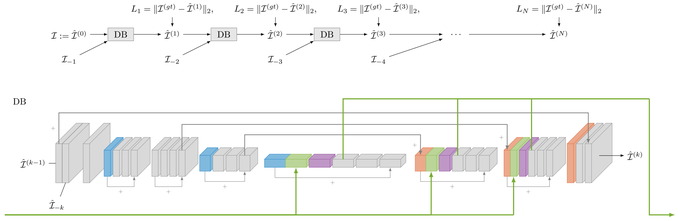
approach	**ours**	[7, 8, 9]
source of frames	YouTube videos	manually recorded
corpus training size	5 hours	a few minutes
recording device types	many different	a few devices
video content	highly diverse content	constrained by capturing effort

Predicting Sharp Frames

Ideally, an approach should incorparate inputs of arbitrary *spatial* ($H \times W \times C$) and *temporal* size (T).

method	[6, 7]	[2, 5]	convGRU	ours
temporal information propagation	✓	✘	✓	✓
arbitrary temporal input size	✘	✓	✓	✓
easy to train	✓	✓	✘	✓
small memory footprint	✘(entire stack on GPU)	✘	✓	✓(at most 2 frames)
input shape	[B,H,W,C*T]	[B*T,H,W,C]	[B,T,H,W,C]	[B,T,H,W,C]

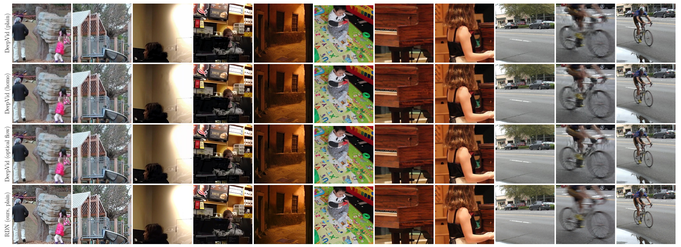
We propose to iteratively apply fully-convolutional *DeblurBlocks* (DBs) on an input sequence. Each DB is trained to predict a sharper version of $\mathcal{I}$ given the current estimate $\hat{\mathcal{I}}^{(k)}$ and an additional observation $\mathcal{I}_{-k}$.



Our proposed network architecture. All DeblurBlocks (top) share the same parameters and are trained to improve the current prediction. These are applied iteratively with temporal skip-connections (green) using an encoder-decoder network.

Results Dynamic Scenes

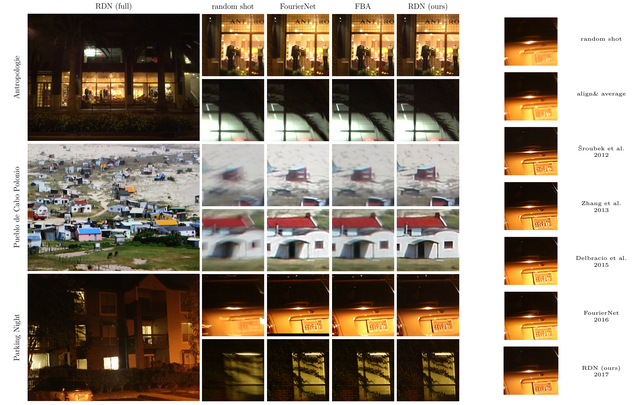
Without the need of applying expensive pre-processing steps our approach is competitive to recent methods like the network of [7] which also has to deal with artifacts from the alignment-step using optical flow:



Evaluation against previous methods on dynamic scenes.

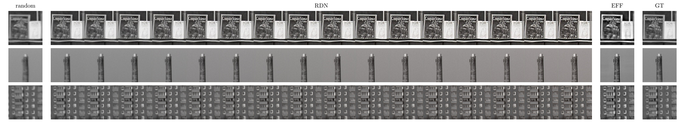
Results Static Scenes

Each DB iteration takes 0.57sec on GTX Titan X. Hence, our approach has competitive speed to FBA and outperforms the FourierNet method of [6] that takes 15min without sacrificing image quality.



Evaluation against previous methods on static scenes.

Although, we exclusively trained on input sequences of 5 frames, our network is applicable to a sequence with *arbitrary* temporal length and spatial size.



We applied our approach on sequences up to 16 inputs and get competitive results to [1].

References

[1] M. Hirsch, S. Sra, B. Schölkopf and S. Harmeling. *Efficient filter flow for space-variant multiframe blind deconvolution.*, CVPR 2010.
[2] C. Schuler, M. Hirsch, S. Harmeling, and B. Schölkopf. *Learning to Deblur.*, NIPS 2014.
[3] M. Delbracio and G. Guillermo. *Burst deblurring: Removing camera shake through fourier burst accumulation.*, CVPR 2015.
[4] M. Delbracio and G. Guillermo. *Hand-held video deblurring via efficient fourier aggregation.*, Trans. on Comp. Imaging 2015.
[5] A. Chakrabarti. *A Neural Approach to Blind Motion Deblurring.*, ECCV 2016.
[6] P. Wieschollek, B. Schölkopf, H. Lensch and M. Hirsch. *End-to-End Learning for Image Burst Deblurring.*, ACCV 2016.
[7] S. Su, M. Delbracio, J. Wang, G. Guillermo, W. Heidrich, and O. Wang. *Deep Video Deblurring.*, CVPR 2017.
[8] M. Noroozi, P. Chandramouli, and P. Favaro. *Motion Deblurring in the Wild.*, Arxiv 2017.
[9] T. Kim, K. Lee, B. Schölkopf and M. Hirsch. *Online Video Deblurring via Dynamic Temporal Blending Network.*, ICCV 2017.