Blind deblurring of images and videos

PhD thesis proposal

Keywords: blind deblurring, image and video restoration, deep learning, plug & play methods, variational methods, computer vision, diffusion models.

Contact

- Andrés Almansa andres.almansa@parisdescartes.fr

Supervising Team

- Academic Advisor: Andrés Almansa (CNRS & Université Paris Cité)
- Industrial Advisor: Eva Coupeté (GoPro)

Context & Objectives

Images, bursts and videos acquired by smartphones, action cameras, and even professional digital cameras are affected by blur of different sources. Even though some of these (intrinsic) sources can be calibrated, other (extrinsic) blur sources like motion, and defocus are unknown and need to be guessed from the degraded image itself. This blind deblurring problem is extremely ill-posed, and despite recent progress thanks to deep learning and generative AI techniques, the problem is still wide open for future improvements.
During the last three years, our team has significantly improved the state of the art performance in blind image deblurring. On one hand [Laroche2023] achieves reconstructions of unprecedented quality, but at a computational cost that is still not well-suited for real-time or embedded applications. On the other hand [Carbajal2023; Laroche2022] provide more flexibility at a lower cost but can only handle blur kernels of relatively small size. The aim of this thesis is to significantly improve the state of the art in blind image and video deblurring, in terms of image quality, real world applicability and computational speed by leveraging both mathematical models and cutting edge deep learning technologies.

Methodology

More precisely this thesis will explore the following research directions.

Part I- Diffusion models and bridges

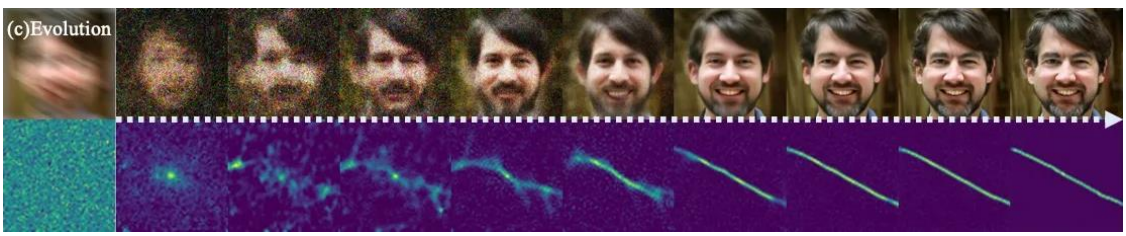

Figure 1: Diffusion-based blind image deblurring. Image credit [BlindDPS].

I.a- Accelerate via latent space diffusion

First of all the Diffusion EM blind deblurring method in [Laroche2023; BlindDPS] progressivley estimates the sharp image and blur kernel as shown in the figure above. It can be further accelerated in many ways to reach computational complexities closer to what can be executed in embedded devices: The most obvious one is to replace the standard diffusion model by a more lightweight one which runs in the compressed latent-space of an autoencoder like Stable Diffusion [StableDiffusion]. Early attempts to solve inverse problems with Stable

Diffusion, produced plausible images that are only loosely consistent with the measurements. To improve this consistency we need to:

1. find better trade-offs between compression and expressivity of the autoencoder, and
2. use more accurate approximations to compute the likelihood $p(y|z_t)$. This can be done by generalizing the Gaussian approximations proposed in [PiGDM], in combination with linearization and the use of the Fisher identity like in latent-space-ULA [Pereyra2022].

I.b- Accelerate further via diffusion bridges.

Diffusion models need a large number of steps because they start from a Gaussian white noise image x_T that is progressively transformed into an image from the posterior $x_0|y$. This is a sensible choice for generative models (without conditioning on the observations y). For conditional generative models where the distribution of the degraded observations y is relatively close to that of the clean image x_0, we can reduce the number of steps if the backward diffusion process starts from the distribution of the degraded observations y as evidenced by Image to Image Schrödinger Bridges [I2SB], or a deterministic variant called Inversion by Direct Iteration [InDI]. The similarities between these two approaches were highlighted in [CDDB], but need to be further explored in the context of blind deblurring. In particular, we shall address the following questions

1. I2SB and InDI may be trained for the blind deblurring task, without need for an explicit Expectation Maximisation step for kernel estimation like in [Laroche2023]. The question is whether these methods can handle a large family of blur kernels as effectively as explicit blur estimation methods like [Laroche2023] and [Carbajal2023].
2. In [CDDB] the authors suggest that I2SB and InDI do not sufficiently enforce data consistency, and suggest to incorporate a guidance mechanism (like in PiGDM) to improve this aspect. This suggests a negative answer to the previous question, and calls for the use (in the blind deblurring case) of an EM algorithm like [Laroche2023], where the diffusion model is replaced by a guided Schrödinger bridge. Such an algorithm is a promising alternative both in terms of speed and accuracy, that will be explored in this thesis.

I.c-Extend to spatially varying blur

So far we discussed improvements to the method in [Laroche2023], but this method has a salient shortcoming: it cannot deal in its present form with spatially varying blur. To deal with spatially varying blur we need to adapt both the score network (modeling the image prior) and the blur estimation module. The latter can handle spatially varying blur by employing the kernel prediction networks suggested in [Carbajal202]. To adapt the score network, we can employ an unrolled architecture like in [Laroche2022], or the decomposed diffusion sampler recently proposed in [DDS].
From a mathematical viewpoint the FastDiffusionEM method in [Laroche2023] is surprising: it is an accelerated version of a well-known Expectation Maximization algorithm, with known convergence properties. The convergence properties of the accelerated version are not known, and yet this accelerated version provides better performance than the original (slower) EM algorithm. This calls for an in-depth study of the accelerated algorithm to clarify why it works so well, and how it can be further improved or generalized.

Part II- Joint blur-prediction/deblurring architecture

The work in [Carbajal2023] is able to deal with spatially varying blur thanks to a kernel prediction network that infers spatially localized blur kernels. These local kernels are then deblurred by a non-blind deblurring network with an unrolled architecture [Laroche2022] as shown in the figure below. Both modules can be improved in order to deal with extreme blur.

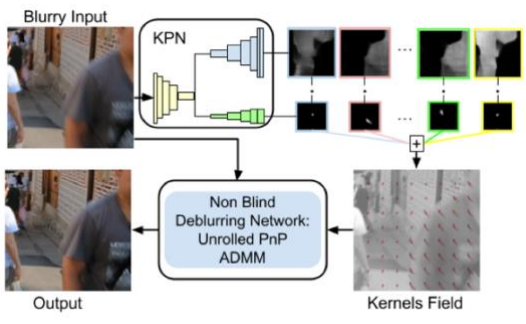

Figure 2: Joint blur estmation and deblurring. Image credit [Carbajal 2023]

II.a- Dealing with extreme blur

First of all, blur identification and deblurring performance can be improved if both kernel prediction and deblurring are performed progressively within a diffusion model as demonstrated in [Laroche2023] in the uniform blur

case. Doing so in the non-uniform blur setting requires to adapt the architecture and training mechanism of both the deblurring network (that should act now as a conditional score) and the blur estimation module (that should take as input an image with different levels of noise). Furthermore, the overall method can be accelerated by replacing the diffusion model by a diffusion bridge [InDI; I2SB; CDDB], or by a latent diffusion engine [StableDiffusion].

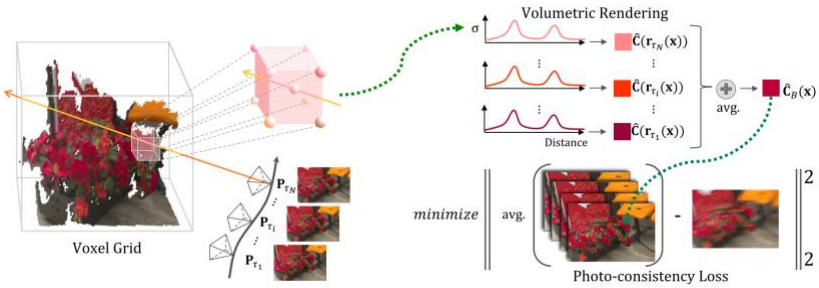

Figure 4: Extreme blur. Image credit [ExBlurRF]

Current attempts to deal with extreme blur are based on NeRFs [ExBluRF; DeblurNeRF] (see figure above) or on the projective motion blur model [Tai2011; Hirsch2011; Whyte2010]. These works need to assume that the scene is static, and that the focal length is sufficiently long or that the scene is far enough to be considered planar. In that case the deformations of the scene projected in the image plane are reduced to homographies, and the non-uniform blur can be globally modeled by a small set of (a few hundred) parameters. While these methods achieve impressive results when the planar assumptions hold, most blur scenes are not planar or contain moving objects, leading to unsatisfactory results.
In order to allow for both extreme blurs and moving objects in the scene, we can localize the projective motion blur model, by assuming that the scene is composed of a set of (possibly moving) planar patches. Each planar patch is then associated to a homographic motion blur model, and the global blur can be predicted by a siamese network similar to the KPN in [Carbajal2023]. Similarly the non-blind deblurring network (unrolled or diffusion model architecture) needs to be adapted from the KPN case to the piecewise-homographic blur representation. Finally both blur prediction and deblurring networks can be jointly trained as in [Carbajal2023].

II.b- Self-Supervised training

A major limitation of the work in [Carbajal2023] and the extension described above is that it requires triplets of sharp image, blurred image, blur kernel for supervised training. Even though a remarkable work has been done to build excellent synthetic datasets, there is still the risk of domain shift when applying these algorithms in situations where such training data is not available.
In such cases self-supervised training or fine-tuning can provide a sensible way of overcoming domain shifts. In particular, equivariant imaging [Tachella2023; Chen2023] provides a principled way of training without ground truth: they show that degraded observation suffices for training, as long as the prior distribution of the latent image satisfies some invariance properties.
In the case of blind deblurring we may safely assume that natural images follow a rotationally invariant distribution (and possibly also a scale invariant one). This invariance can be exploited via a data augmentation mechanism that leads to a self-supervised loss. For this procedure to work for a given degradation operator A, a necessary condition is that the stack of operators $[AR_1 \dots AR_n]$ (where R_i represent the different rotations) is full rank. For the case of motion deblurring where the kernels are locally directional, this condition is clearly satisfied.
As a first step toward fully self-supervised learning, we shall first consider the non-blind case, then the blind case with simple directional blurs, and finally the full model with non-uniform blur described in the previous paragraph.

Part III- motion blured videos

III.a- Plug & play methods and temporal upsampling.

So far, we dealt with the problem of blind deblurring of a single image. However, if we consider blind motion deblurring of videos, the temporal dimension introduces additional information that can be used to our advantage. More specifically, let's consider the problem of joint temporal upsampling and blind deblurring of

videos affected by motion blur, *i.e.* we want to recover an ideal high-framerate, sharp sequence $u(x,t)$ from the acquired low-framerate motion-blurred sequence $v(x,t)$. The relationship between both sequences can not only be expressed in closed form (1b), it does not require the knowledge of the unknown motion blur, and the corresponding proximal operator can be computed explicitly. This means that if we have a strong prior R_1 for u, that enforces both sharpness and temporal consistency, then we can solve this joint problem by computing the maximum a posteriori

$$\hat{u} = \text{argmin}_u D_1(u,v) + R_1(u) \tag{E1}$$

using alternate proximal descent algorithms like ADMM. The data-fitting term D_1 reflects equation (1b). The prior R_1 can be either hand-crafted ($e.g.$ spatio-temporal total variation) or implicit in a video denoiser algorithm as demonstrated in [Monod2023]. In the latter case we should prioritize SoTA video denoisers with a large temporal receptive field like [FastDVDNet; VRT; RVRT], possibly retrained on high-framerate videos. Alternatively, we could employ a video diffusion model as a prior as proposed in [Cherel2023].

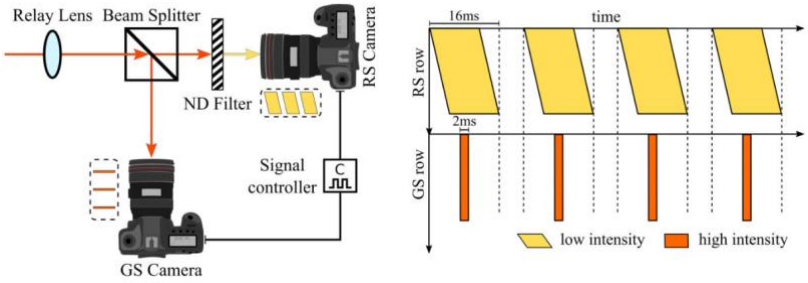

Figure 5: joint blind motion deblurring and upsampling. Image credit [Zhong2021]

The idea of joint motion deblurring and temporal upsampling of videos was already successfully explored in [NIRE, Zhong2021], by training a neural network specifically for this task in a supervised manner. To do so, sharp/blurred sequence pairs are required (see figure above) whereas the proposed plug & play approach only requires sharp sequences to train a powerful denoiser or diffusion model.
A potential shortcoming of the previous approach is that a large oversampling rate r is required for the Riemann sum (1b, that defines D_1) to accurately approximate the actual integral (1). In order to write a more accurate approximation with relatively low values of r, we can compute the optical flow ϕ of the sharp sequence [RIFE], and either

1. impose temporal regularity to the optical flow, or
2. use the optical flow to improve the temporal resolution of the data-fitting term D_1.

More precisely, the first idea amounts to considering the blind deblurring problem in a variational formulation as a joint minimization of an augmented energy of the form

$$\text{argmin}_{u,\phi} D_1(v,u) + R_1(u) + D_2(u,\phi) + R_2(\phi) \tag{E2}$$

where ϕ is the optical flow of the (unknown) sharp sequence, i.e. it minimizes a coupling term $D_2(u,\phi)$. By imposing temporal coherence to this vector field, we indirectly impose (via the coupling term D_2) temporal coherence to the sharp sequence, which is essential for sensible upsampling. For instance, we could use the following spatio-temporal regularization of the optical flow.
From an algorithmic viewpoint the variational formulation (E2) can be solved by alternate optimization splitting algorithms like ADMM, to separate implicit regularization R_1 and R_2 (as denoising algorithms on u and ϕ like in PnP approaches), from data-fitting and coupling terms like D_1 and D_2.
From a mathematical viewpoint, the non-trivial coupling term D_2 requires some careful attention to ensure its convergence. In [NikolovaTan2019] the convergence of a similar variational approach is studied in detail, where the coupling term is bi-convex. In our case only a linearized version of D_2 is bi-convex, so a generalization of that result shall be required if we want convergence guarantees for this algorithm. A further generalization shall be required to account for possibly non-convex implicit regularization terms $R_1(\varphi)$.

Funding & work environment

This PhD position is funded by GoPro Research Paris, via a CIFRE agreement with Université Paris Cité.
As a PhD student you will be part of both GoPro's research team in Boulogne, and of MAP5 lab in Paris 6ème arrondissement.

In addition to the academic advisors (Andrés Almansa – MAP5, and Eva Coupeté – GoPro) you will be part of an international collaboration team including Charles Laroche, Thomas Derbanne (GoPro), Pablo Musé & Guillermo Carbajal (ENS Paris-Saclay & UdelaR), Marcelo Pereyra (Heriot Watt University, Scotland), Pauline Tan (Sorbonne Université).

Candidate

The successful candidate is expected to hold a Master's degree in Applied Mathematics, Computer Science, Electrical Engineering, or a closely related discipline. They should have good programming skills in deep learning and scientific computing environments (*e.g.* Python, NumPy, PyTorch, TensorFlow, JAX).
In addition, the following qualifications are desired:

- Solid background in some of the following areas: Computational Imaging, Image Restoration, Variational Methods in Imaging, Bayesian Methods for Inverse Problems.
- Fluent in either French or English (written and spoken) and good communication and interpersonal skills.

Application

Candidates should apply by filling in the following form including at least

- CV and motivation letter
- Grades from your M2, M1
- One or two academic references (and their recommendation letters, if available)

Online form: https://forms.gle/5aQR5TTRBWiB3cVP6
You can contact andres.almansa@parisdescartes.fr to notify me about your application, or for any questions concerning the position.

References

[**Carbajal2023**] Carbajal, G., Vitoria, P., Lezama, J., & Musé, P. (2023). Blind Motion Deblurring with Pixel-Wise Kernel Estimation via Kernel Prediction Networks. IEEE Transactions on Computational Imaging, to appear. https://github.com/GuillermoCarbajal/J-MKPD/

[**Laroche2022**] Laroche, C., Almansa, A., & Tassano, M. (2023). Deep Model-Based Super-Resolution with Non-uniform Blur. In *(WACV) Winter Conference on Applications of Computer Vision* (pp. 1797–1808). Waikoloa, Hawaii: IEEE. https://doi.org/10.1109/WACV56688.2023.00184

[**Laroche2023**] Laroche, C., Almansa, A., & Coupete, E. (2023). Fast Diffusion EM: a diffusion model for blind inverse problems with application to deconvolution. In *WACV 2024* (to appear). http://arxiv.org/abs/2309.00287

[**StableDiffusion**] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2021). High-Resolution Image Synthesis with Latent Diffusion Models, 10684–10695. http://arxiv.org/abs/2112.10752

[**PiGDM**] Song, J., Vahdat, A., Mardani, M., & Kautz, J. (2023). Pseudoinverse-Guided Diffusion Models for Inverse Problems. In *(ICLR) International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=9_gsMA8MRKQ

[**Pereyra2022**] Pereyra, M., Vargas-Mieles, L. A., & Zygalakis, K. C. (2022). The split Gibbs sampler revisited: improvements to its algorithmic structure and augmented target distribution, 1–22. http://arxiv.org/abs/2206.13894

[**I2SB**] Liu, G.-H., Vahdat, A., Huang, D.-A., Theodorou, E. A., Nie, W., & Anandkumar, A. (2023). I2SB: Image-to-Image Schrödinger Bridge. In *(ICML) International Conference on Machine Learning*. Retrieved from http://arxiv.org/abs/2302.05872

[**InDI**] Delbracio, M., & Milanfar, P. (2023). Inversion by Direct Iteration: An Alternative to Denoising Diffusion for Image Restoration, 1–33. Retrieved from http://arxiv.org/abs/2303.11435

[**CDDB**] Chung, H., Kim, J., & Ye, J. C. (2023). Direct Diffusion Bridge using Data Consistency for Inverse Problems, 1–16. Retrieved from http://arxiv.org/abs/2305.19809

[**ExBluRF**] Lee, D., Oh, J., Lim, J., Cho, S., & Lee, K. M. (2023). ExBluRF: Efficient Radiance Fields for Extreme Motion Blurred Images, 17639–17648. http://arxiv.org/abs/2309.08957

[**DDS**] Chung, Hyungjin, Suhyeon Lee, and Jong Chul Ye. 2023. "Decomposed Diffusion Sampler for Accelerating Large-Scale Inverse Problems," March, 1–28. http://arxiv.org/abs/2303.05754.

[**DeblurNeRF**] Ma, L., Li, X., Liao, J., Zhang, Q., Wang, X., Wang, J., & Sander, P. V. (2022). Deblur-NeRF: Neural Radiance Fields from Blurry Images Camera Motion Blur Defocus Blur (a) Samples of Blurry Source Views (b) Novel Views from NeRF (c) Novel Views from Deblur-NeRF. *(CVPR) Conference on Computer Vision and Pattern Recognition*, 12861–12870.
[**Gupta2010**] Gupta, A., Joshi, N., Lawrence Zitnick, C., Cohen, M., & Curless, B. (2010). Single image deblurring using motion density functions. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6311 LNCS*(PART 1), 171–184. https://doi.org/10.1007/978-3-642-15549-9_13
[**Tai2011**] Tai, Y. W., Tan, P., & Brown, M. S. (2011). Richardson-Lucy deblurring for scenes under a projective motion path. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(8), 1603–1618. https://doi.org/10.1109/TPAMI.2010.222
[**Hirsch2011**] Hirsch, M., & Schuler, C. (2011). Fast removal of non-uniform camera shake. *(ICCV) International Conference on Computer Vision*. http://ieeexplore.ieee.org/document/6126276/
[**Whyte2010**] Whyte, O., Sivic, J., Zisserman, A., & Ponce, J. (2010). Non-uniform deblurring for shaken images. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 491–498. https://doi.org/10.1109/CVPR.2010.5540175
[**Chen2023**] Chen, D., Davies, M., Ehrhardt, M. J., Schonlieb, C.-B., Sherry, F., & Tachella, J. (2023). Imaging With Equivariant Deep Learning: From unrolled network design to fully unsupervised learning. *IEEE Signal Processing Magazine*, *40*(1), 134–147. https://doi.org/10.1109/MSP.2022.3205430
[**Tachella2023**] Tachella, J., Chen, D., & Davies, M. (2023). Sensing Theorems for Unsupervised Learning in Linear Inverse Problems. *(JMLR) Journal of Machine Learning Research*, *24*, 1–45. Retrieved from http://arxiv.org/abs/2203.12513
[**Monod2023**] Monod, A., Delon, J., Tassano, M., & Almansa, A. (2022). *Video Restoration with a Deep Plug-and-Play Prior*. https://doi.org/10.48550/arxiv.2209.02854
[**NikolovaTan2019**] Nikolova, Mila, and Pauline Tan. 2019. "Alternating Structure-Adapted Proximal Gradient Descent for Nonconvex Nonsmooth Block-Regularized Problems." *SIAM Journal on Optimization* 29 (3): 2053–78. https://doi.org/10.1137/17M1142624, https://hal.science/hal-01677456.
[**FastDVDNet**] Tassano, M., ...
[**VRT**] Liang, J., Cao, J., Fan, Y., Zhang, K., Ranjan, R., Li, Y., … Van Gool, L. (2022). VRT: A Video Restoration Transformer. https://github.com/jingyunliang/vrt, https://paperswithcode.com/paper/vrt-a-video-restoration-transformer
[**RVRT**] Liang, J., Cao, J., Fan, Y., Zhang, K., Ranjan, R., Li, Y., … Van Gool, L. (2022). VRT: A Video Restoration Transformer. https://github.com/jingyunliang/vrt, https://paperswithcode.com/paper/recurrent-video-restoration-transformer-with
[**NIRE**] Zhang, X., Huang, H., Jia, X., Wang, D., & Lu, H. (2023). Neural Image Re-Exposure. http://arxiv.org/abs/2305.13593
[**Zhong2021**] Zhong, Z., Zheng, Y., & Sato, I. (2021). Towards Rolling Shutter Correction and Deblurring in Dynamic Scenes. In *(CVPR) Conference on Computer Vision and Pattern Recognition* (pp. 9215–9224). IEEE. http://arxiv.org/abs/2104.01601
[**Cherel2023**] Nicolas Cherel, Yann Gousseau, Alasdair Newson, A. A. (2023). *Infusion: internal diffusion for video inpainting*. Preprint
[**RIFE**] Huang, Z., Zhang, T., Heng, W., Shi, B., & Zhou, S. (2020). Real-Time Intermediate Flow Estimation for Video Frame Interpolation. In *(ECCV) European Conference on Computer Vision* (Vol. 13674 LNCS, pp. 624–642). https://paperswithcode.com/paper/rife-real-time-intermediate-flow-estimation