A factor of integer polynomials with minimal integrals

A NOVEL METHOD AND DATASET FOR DEPTH-GUIDED IMAGE DEBLURRING FROM SMARTPHONE LIDAR

Antonio Montanaro, Diego Valsesia

Politecnico di Torino, Italy

ABSTRACT

Modern smartphones are equipped with Lidar sensors providing depth-sensing capabilities. Recent works have shown that this complementary sensor allows to improve various tasks in image processing, including deblurring. However, there is a current lack of datasets with realistic blurred images and paired mobile Lidar depth maps to further study the topic. At the same time, there is also a lack of blind zero-shot methods that can deblur a real image using the depth guidance without requiring extensive training sets of paired data. In this paper, we propose an image deblurring method based on denoising diffusion models that can leverage the Lidar depth guidance and does not require training data with paired Lidar depth maps. We also present the first dataset with real blurred images with corresponding Lidar depth maps and sharp ground truth images, acquired with an Apple iPhone 15 Pro, for the purpose of studying Lidar-guided deblurring. Experimental results on this novel dataset show that Lidar guidance is effective and the proposed method outperforms state-of-the-art deblurring methods in terms of perceptual quality.

***Index Terms*—** Image deblurring, Depth maps, Lidar

1. INTRODUCTION

Modern smartphones are increasingly becoming multimodal imaging devices with the inclusion of Lidar sensors in recent consumer devices, such as the Apple iPhone. Despite its limited resolution due to space and cost constraints, the active Lidar instrument can serve as complementary source of information to passive optical cameras, even in tasks that do not explicitly deal with 3D reconstruction. In fact, promising results have been shown in terms of enhanced rate-distortion performance for Lidar-guided image compression [1] and improved image quality for Lidar-guided deblurring [2].

Focusing on deblurring, the active nature of the Lidar sensor can be useful in low-light scenarios where motion blur is easily introduced in photos, while depth information can provide sharp object boundaries. While the work in [2] has shown that smartphone Lidar data can substantially enhance image deblurring, it has done so via simulation of blurred images, albeit with realistic kernels. In fact, currently, the main limitation towards a deeper development of the topic is the lack of curated data with real blurred images under low-light conditions, paired with Lidar depth maps. The only existing dataset with paired mobile Lidar depth maps is ARKitScenes [3], which does not have real blurred images. Another limitation of [2] is the supervised training approach, which requires a large dataset with paired data, including blurred image, registered Lidar depth maps and sharp ground truth. It would be desirable to devise a zero-shot method that learns a general "sharp image prior" and side-information fusion operator that allows to avoid the dependency on large quantities of images with associated Lidar depth maps.

In this paper, we address both of the aforementioned limitations. First, we present a novel dataset of real low-light images affected by motion blur, captured with an Apple iPhone 15 Pro smartphone, with registered Lidar depth maps and sharp ground truth images. Moreover, we present a novel blind zero-shot method for Lidar-guided image deblurring, called ZSLDB (Zero-Shot Lidar DeBlur) that leverages denoising diffusion generative models to avoid the need for extensive paired datasets required by supervised learning. The model is blind to the real degradation kernel, which is estimated during inference. It also allows for conditioning based on the Lidar depth maps without explicit training with paired data having them. Experimental results on the novel dataset of real blurred images with mobile Lidar depth maps show that the depth information improves the quality of the reconstructed images and that ZSLDB outperforms state-of-the-art algorithms in terms of perceptual image quality.

Our novel contributions can be therefore summarized as:

- we present a novel blind zero-shot method for image deblurring guided by Lidar depth maps, called ZSLDB;
- we introduce a novel dataset of real smartphone images affected by motion blur, with registered Lidar depth maps and sharp ground truth images;
- experimental results show that thanks to Lidar depth information ZSLDB achieves state-of-the-art perceptual

This study was carried out within the "AI-powered LIDAR fusion for next-generation smartphone cameras (LICAM)" project – funded by European Union – Next Generation EU within the PRIN 2022 program (D.D. 104 - 02/02/2022 Ministero dell'Università e della Ricerca), CUP E53D23000790006. This manuscript reflects only the authors' views and opinions and the Ministry cannot be considered responsible for them.

quality;

- code and dataset are available at `https://github.com/diegovalsesia/ZSLDB`.

2. BACKGROUND

Image restoration, including deblurring, is a longstanding problem in the image processing field. Several solutions have been developed over the years, including optimization-based methods carefully modeling priors suitable to describe natural images. Deep learning has then shown that neural networks could effectively capture more sophisticated representations that led to improved reconstruction performance. At a high level, deep learning approaches generally either use a supervised learning strategy, directly learning the inverse mapping between degraded observations and reconstructions, or a generative model learning a data prior, including learned denoisers in plug-and-play approaches [4], variational autoencoders (VAEs) [5], generative adversarial networks [6], denoising diffusion models [7], etc.

The former approach has proved to be very successful with recent models like Restormer [8], NAFNet [9] achieving impressive results on a variety of inverse problems. Concerning blind image deblurring, where the blurring operator is not known in advance, the recent J-MKPD model [10] can be regarded as the state-of-the-art. The latter approach based on generative models is particularly suitable when image restoration has to face with real data, and it is difficult to source large quantities of paired data for supervised training. Morevoer, by modeling the real data distribution, methods based on generative models favour better perceptual quality in the perception-distrotion tradeoff. Examples for the solutions of inverse problems include works like PULSE [11], based on GAN inversion to seek a solution to the problem in the latent space of the generative model. DeblurGAN [12] is another well-known work focusing on blind motion deblurring using a Wasserstein GAN.

Since the rise of denoising diffusion models as the new state of the art in image generation, zero-shot methods exploiting these priors have become even more popular. In DPS [13], the authors extend diffusion solvers to efficiently handle general noisy (non)linear inverse problems via approximation of the posterior sampling; DDRM [14] reformulates the diffusion process to be guided by the degraded observation; in DDNM [15], the authors decompose the restored signal in two algebraic parts and exploit the diffusion process to fill up the information of one part; [16] integrates a traditional plug-and-play method into the diffusion sampling framework; the authors in [17] are the first who consider latent diffusion model rather than pixel-space diffusion models, and they define a new correction term to DPS. Other works also utilize backpropagation along the sampling chain for image restoration [18]. Some limitations of the previous works that challenge them on real-world data are the often-used assumption that the degradation operator is known (non-blind problem), or the use pixel-space diffusion models trained on simple datasets like Imagenet or faces FFHQ, which do not scale well to realistic natural images.

Moreover, works in the literature typically do not assume the availability of side information to condition the reconstruction process, which is the case of our setting where the Lidar depth map should provide guidance about edges. Before the advent of deep learning, an effective solution towards side information guidance was the use of the guided filter [19]. More recently, ControlNet [20] has shown a general way to condition the neural networks used in denosing diffusion models with side information. To the best of our knowledge, there is no work about blind image restoration using a diffusion model guided by depth maps.

3. METHOD

The overall goal of this paper is to investigate the problem of image deblurring from smartphone cameras, when side information in the form of a depth map acquired by the Lidar on the same device is available. Compared to [2] that worked on synthetically degraded data, we focus on real data by both devising a novel depth-guided zero-shot deblurring method based on diffusion models, which can deblur our target images without requiring training data with blurred images and Lidar depth maps, as well as a novel dataset presented in Sec. 4.

3.1. Problem setting, diffusion models and conditional control

Image deblurring can be regarded as a linear inverse problem:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n} \tag{1}$$

where the original sharp image $\mathbf{x}$ has been corrupted by a forward operator $\mathbf{A}$ and, possibly, an additive random noise $\mathbf{n}$ to generate the blurred image $\mathbf{y}$. In its simplest form, the forward operation is a convolution with a low-pass filter, representing a spatially-invariant degradation. In more complex models, a spatially-variant filter is used to capture nonstationary degradations.

The classical formulation to solve this ill-posed problem is optimization as a Maximum a Posteriori (MAP) estimation problem:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda R(\mathbf{x}),$$

A regularizer function $R(\mathbf{x})$ captures the data prior by encoding the knowledge about the sharp original images.

With the advent of powerful generative models, strong image priors are available to regularize inverse problems. In essence, many generative models learn a map G from a simple

distribution on a latent space $\mathcal{Z}$ to the real data distribution. This allows to solve an inverse problem via inversion to the latent space, i.e., seeking the point in the model latent space (thus guaranteeing a realistic generation) that best fits the degraded observations. More formally, the reconstructed image $\hat{\mathbf{x}}$ is obtained as:

$$\hat{\mathbf{x}} = G(\hat{\mathbf{z}}) \tag{2}$$

$$\hat{\mathbf{z}} = \arg\min_{\mathbf{z}} \|\mathbf{y} - \mathbf{A}G(\mathbf{z})\|_2^2 \quad \text{s.t.} \quad z \in \mathcal{Z}. \tag{3}$$

Denoising Diffusion Probabilistic Models (DDPMs) [7] are generative models that utilize a noise diffusion process to model the image distribution starting from random noise of the same dimension. The model is trained to progressively denoise data with different levels of noise, allowing the model to generate new data from pure noise by reverting the forward process; this is denoted as backward process and can be formulated as:

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$$
$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1} \mid \mu_\theta(\mathbf{x}_t, t), \sigma_t^2 I). \tag{4}$$

where $\mu_\theta(\mathbf{x}_t, t)$ is a function of a denoising neural network parametrized by θ. A widely used neural network architecture for the denoiser is a UNet with self-attention blocks [21]. Even if this process is Markovian, the work of [22] shows that is possible to construct a non-Markovian process defining a faster deterministic sampler (DDIM) that is compatible with a pretrained model. So starting from $p_\theta(\mathbf{x}_{0:T})$, it is possible to sample $\mathbf{x}_{t-1}$ using:

$$x_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{x_t - \sqrt{1-\alpha_t}\hat{\epsilon}_t}{\sqrt{\alpha_t}} \right)$$
$$+ \sqrt{1 - \alpha_{t-1} - \sigma_t(\eta)^2} \cdot \hat{\epsilon}_t + \sigma_t(\eta)\varepsilon_t \tag{5}$$

where $\sigma_t(\eta) = \eta \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}} \sqrt{\frac{1-\alpha_t}{\alpha_{t-1}}}$, ε_t is a normalized gaussian variable and $\epsilon_t(\cdot)$ is the denoiser neural network estimating the noise realization. $\eta \in (0, 1)$ is a parameter controlling the forward process. When $\eta = 0$ the sampling becomes deterministic, enabling the inversion approach described in Eq. (2). Stable Diffusion is a latent diffusion model [23] that uses a Variational Auto-Encoder (VQ-VAE [24]) to reduce the image dimensions and run the diffusion process in the reduced space encoded by the VAE encoder E, and after the final denoising step, let the VAE decoder D recover the full image.

Diffusion models can be easily extended to model $p(\mathbf{x}|\mathbf{s})$, where $\mathbf{s}$ is a conditioning signal, such as an image caption, category, semantic maps etc., either by providing an additional input to the denoising neural network, which is typically used for text-conditioned generation, or by means of subnetworks that modulate the features of the denoiser. The latter approach has been shown by ControlNet [20] to allow the integration of various forms of side information without the need to retrain the denoiser of the generative model.

In the next section, we show how we use these concepts to design a zero-shot Lidar-guided deblurring method.

3.2. Zero-Shot Lidar DeBlur (ZSLDB)

We propose a method called Zero-Shot Lidar DeBlur (ZSLDB) to leverage a pretrained diffusion model to deblur real images acquired by smartphones while conditioning the results on the information of the Lidar depth map captured by the same device. We frame the problem as a blind deblurring problem, as the blur kernel $\mathbf{A}$ is unknown in real images. Our experiments are based on using StableDiffusion (SD) as generative model, but the formulation extends to any latent diffusion model. We denote the Lidar depth map to be used as side information as $\mathbf{d}$ and the mapping performed by the diffusion model through the iterated denoiser function as G.

Conceptually, ZSLDB casts the deblurring problem as the solution to the following optimization problem, following the inversion principle outlined in Eq. (2):

$$\hat{\mathbf{z}}, \hat{\mathbf{A}} = \arg\min_{\mathbf{z},\mathbf{A}} \|\mathbf{y} - \mathbf{A}G(\mathbf{z}, \mathbf{d})\|_2^2, \quad \hat{\mathbf{x}} = G(\hat{\mathbf{z}}, \mathbf{d}). \tag{6}$$

Once the optimization finds the desired latent noise $\hat{\mathbf{z}}$, this is used by the model to generate the deblurred image $\hat{\mathbf{x}}$. Notice that the optimization is both over the noise latent $\mathbf{z}$ of the SD model and the blur kernel $\mathbf{A}$. The diffusion model G is a function of the depth signal by means of a ControlNet, mapping a resampled version of the depth map to match the image resolution, to residual features that are added to the denoiser features.

The optimization in Eq. (6) is performed as follows. We start from the blurred image $\mathbf{y}$ and we use DDIM (50 timesteps) to invert this image back to the SD noise latent space at $T = 0$. This inversion has no information about the depth map yet. Then, we generate an image starting from this latent by adding $\mathbf{d}$ as input to the ControlNet and we run the DDIM sampler with only 10 timesteps to generate a guided version of $\mathbf{y}$. Finally, we optimize the latent $\mathbf{z}$ by backpropagation through the sampling chain, adding some regularizers to basic idea of Eq. 6 and by measuring consistency with the observations in the latent space of the VAE model:

$$\hat{\mathbf{z}}, \hat{\mathbf{A}} = \arg\min_{\mathbf{z},\mathbf{A}} \Big[\|E(\mathbf{y}) - E(\mathbf{A}G(\mathbf{z}, \mathbf{d}))\|_2^2$$
$$+ \gamma L_{\text{aesthetic}}(G(\mathbf{z}, \mathbf{d})) + \lambda \cdot \text{LPIPS}(\mathbf{y}, G(\mathbf{z}, \mathbf{d})) \Big]. \tag{7}$$

We found that optimization is more stable and effective by measuring fidelity with the blurred observations both as L2 norm in the VAE latent space, i.e., after the image encoder E, and with the LPIPS [25] perceptual loss in the pixel space. Additionally, $L_{\text{aesthetic}}$ is an aesthetic differentiable

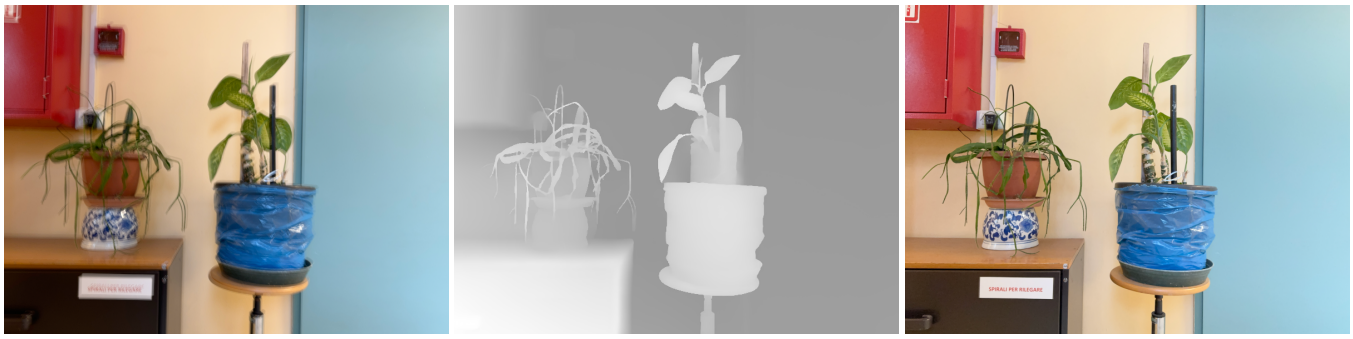

Fig. 1. Sample scene from the novel smartphone Lidar-guided deblurring dataset. Left to right: blurred scene, Lidar depth map, sharp ground truth. Acquired with an Apple iPhone 15 Pro.

reward function as formulated in [26]. This metric was devised by training a multi-layer perceptron operating on CLIP embeddings of 176,000 images to regress human ratings of perceptual quality. The hyperparameters γ and λ control the strength of the two regularizations and are set to 0.1 and 1.5, respectively. Finally, we remark that backpropagation through the sampling chain of the diffusion process is made computationally feasible by means of the commonly-used gradient checkpointing [26, 27] technique to lower memory requirements. We remark that the first inversion process uses a higher number of timesteps to ensure an accurate estimation of the latent noise, while subsequent DDIM sampling uses fewer timesteps for the purpose of acceleration, without significantly reducing image quality.

4. DATASET

Current research efforts into the investigation of the use of smartphone Lidar sensors for the regularization of inverse problems in imaging are limited by the lack of available data. In particular, the only existing dataset with paired smartphone images and Lidar depth maps captured by the same device is ARKitScenes [3]. However, this dataset was created for 3D reconstruction and understanding tasks and not for image restoration. As such, it does not have images affected by realistic degradations such as motion blur.

In this paper, we present a novel dataset acquired with an Apple iPhone 15 Pro of low-light images affected by motion blur. We use the onboard Lidar sensor to also obtain registered depth maps. A sharp ground truth acquired by the same device in a more stable manner is also available and registered to the blurred image and depth map. The dataset is composed of 45 scenes, mostly indoors. We argue that indoor usage is the scenario where the Lidar sensor can be most effective since objects with detectable boundaries will be in range of the instrument. The dataset also comprises acquisitions of the same scenes from a DSLR camera. However, these are only coarsely registered due to differences in focal length but can be used for distribution-level quality assessment rather than paired metrics.

Table 1. Quantitative deblurring results

Method	**LPIPS** $\downarrow$
DeblurGAN [12]	0.1884
Restormer [8]	0.1783
J-MKPD [10]	0.1819
ZSLDB (ours)	**0.1643**

Table 2. Impact of depth information

	ZSLDB no depth	**ZSLDB**
LPIPS $\downarrow$	0.1821	**0.1643**

An example of a blurred scene, Lidar depth map and sharp ground truth is shown in Fig. 1. The full dataset will be made publicly available upon publication.

5. EXPERIMENTAL RESULTS

5.1. Experimental setting and main result

In this section we present deblurring results using the new dataset presented in Sec. 4. We remark that since our focus is on real images and Lidar depth maps, this is the only existing dataset that allows such investigation.

The proposed ZSLDB uses a pretrained Stable Diffusion with ControlNet model as image prior with depth map guidance. Notice that pretraining of SD only used generic natural images, while pretraining of ControlNet used a mixture of edges, sketches, poses, etc.. Critically, real Lidar depth maps were not needed for the pretraining process, rendering ZSLDB "zero-shot" in the sense that it does not require training data with real blurred images paired with depth maps and ground truths, and can be directly used on test data. We center crop and rescale the scenes in the new dataset to a standard 512×512 resolution, compatible with the SD backbone.

We compared the proposed ZSLDB with state-of-the-art and well-known approaches to blind image deblurring. In particular, we use DeblurGAN [12] as a classic baseline us-

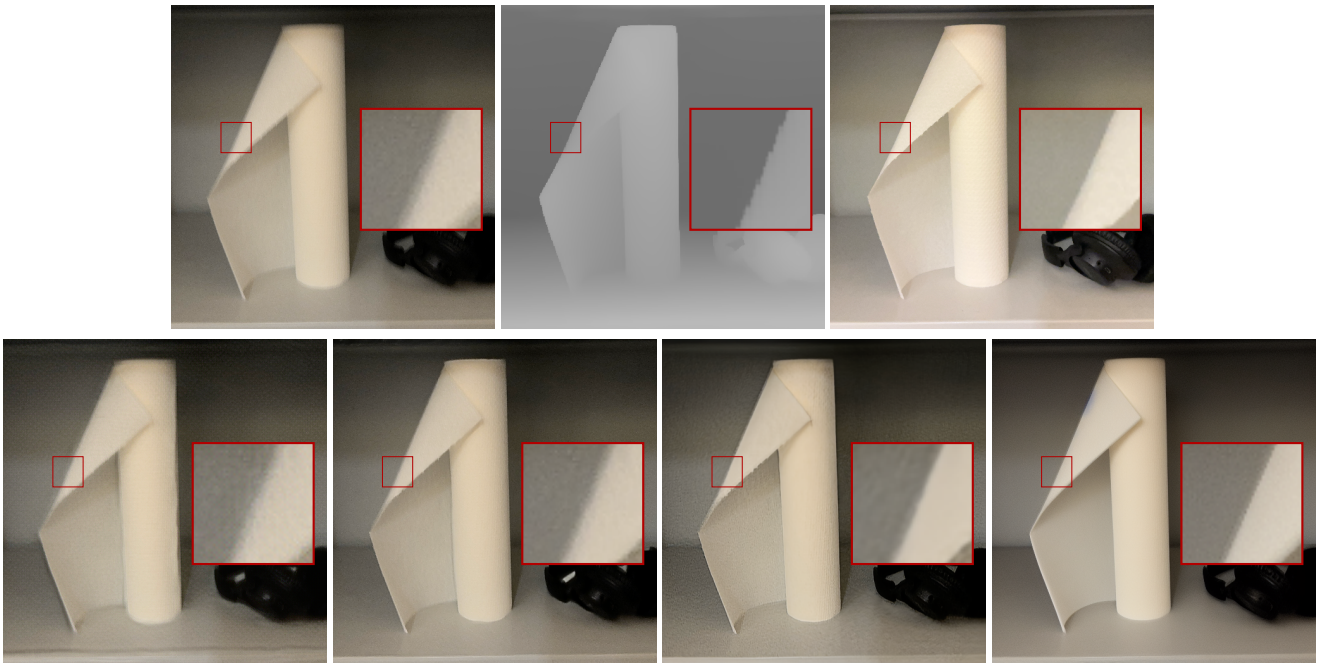

Fig. 2. Qualitative results. Top row left to right: blurred image, depth map, sharp image. Bottom row left to right: DeblurGAN, Restormer, J-MKPD, ZSLDB.

Table 3. Conditioning on depth v. edge map

	ZSLDB (depth)	ZSLDB (edge)
LPIPS ↓	0.1643	0.1682

ing generative models for deblurring, and Restormer [8] and J-MKPD [10] as state-of-the-art models. Quality of the deblurred images is measured by means of the LPIPS metric, which is a perceptual quality metric. We remark that a pixelwise metric like PSNR is not sufficiently robust due to imperfect alignment at a single-pixel level with respect to the ground truth. Table 1 reports the main results. We can see that ZSLDB outperforms the state-of-the-art methods by providing better image quality. Qualitative results are also shown in Fig. 2, where it appears that the edge information provided by the Lidar depth map allows sharper reconstruction.

5.2. Ablation: impact of depth information

In this section, we present an experiment to assess the impact of the depth information on the deblurring performance. In order to do this, we modify the proposed ZSLDB method by removing the ControlNet subnetwork providing depth guidance. The results are reported in Table 2. It can be noticed that ZSLDB without depth guidance is competitive with the state-of-the-art methods reported in Table 1. However, the inclusion of depth guidance results in a significant improvement in perceptual quality, highlighting its importance for image deblurring.

5.3. Ablation: conditioning on depth edges

Since the depth information is most useful in regularizing the deblurring problem by detecting object boundaries, we investigate whether it is more effective to provide guidance via the depth map itself or an edge map extracted from it. In particular, we use a well-known edge detection algorithm (Canny edge detector) on the depth map to derive an edge map that is provided as input to the ControlNet subnetwork. Results are reported in Table 3. It can be noticed that the edge map is still effective at providing a regularizing effect but not to the same extent as using the depth map itself. We conjecture that this might mean that the depth map is providing more information than just the edges, e.g., depth gradients.

6. CONCLUSIONS

We presented an investigation into image deblurring guided by smartphone Lidar depth information with a focus on two contributions: the first dataset of real blurred smartphone photos with registered depth information and a novel method for zero-shot Lidar-guided deblurring based on denoising diffusion models. Results on the new dataset show that the depth information is effective at better regularizing the deblurring problem and that the proposed method outperforms state-of-the-art methods in terms of perceptual quality.

7. REFERENCES

[1] Alessandro Gnutti, Stefano Della Fiore, Mattia Savardi, Yi-Hsin Chen, Riccardo Leonardi, and Wen-Hsiao Peng, "Lidar depth map guided image compression model," in *2024 IEEE International Conference on Image Processing (ICIP)*, 2024, pp. 1890–1896.

[2] Ziyao Yi, Diego Valsesia, Tiziano Bianchi, and Enrico Magli, "Deep lidar-guided image deblurring," *arXiv preprint arXiv:2412.07262*, 2024.

[3] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman, "ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[4] Ulugbek S Kamilov, Charles A Bouman, Gregery T Buzzard, and Brendt Wohlberg, "Plug-and-play methods for integrating physical and learned models in computational imaging: Theory, algorithms, and applications," *IEEE Signal Processing Magazine*, vol. 40, no. 1, pp. 85–97, 2023.

[5] Diederik P Kingma, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[7] Yang Song and Stefano Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[8] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5728–5739.

[9] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun, "Simple baselines for image restoration," in *European conference on computer vision*. Springer, 2022, pp. 17–33.

[10] Guillermo Carbajal, Patricia Vitoria, José Lezama, and Pablo Musé, "Blind motion deblurring with pixel-wise kernel estimation via kernel prediction networks," *IEEE Transactions on Computational Imaging*, 2023.

[11] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin, "Pulse: Self-supervised photo upsampling via latent space exploration of generative models," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2020, pp. 2437–2445.

[12] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jivrí Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8183–8192.

[13] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye, "Diffusion posterior sampling for general noisy inverse problems," *arXiv preprint arXiv:2209.14687*, 2022.

[14] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song, "Denoising diffusion restoration models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23593–23606, 2022.

[15] Yinhuai Wang, Jiwen Yu, and Jian Zhang, "Zero-shot image restoration using denoising diffusion null-space model," *arXiv preprint arXiv:2212.00490*, 2022.

[16] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool, "Denoising diffusion models for plug-and-play image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1219–1229.

[17] Litu Rout, Negin Raoof, Giannis Daras, Constantine Caramanis, Alex Dimakis, and Sanjay Shakkottai, "Solving linear inverse problems provably via posterior sampling with latent diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[18] Bowen Song, Soo Min Kwon, Zecheng Zhang, Xinyu Hu, Qing Qu, and Liyue Shen, "Solving inverse problems with latent diffusion models via hard data consistency," *arXiv preprint arXiv:2307.08123*, 2023.

[19] Kaiming He, Jian Sun, and Xiaoou Tang, "Guided image filtering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 6, pp. 1397–1409, 2012.

[20] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[22] Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising Diffusion Implicit Models," in *International Conference on Learning Representations*, 2020.

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684–10695.

[24] Aaron Van Den Oord, Oriol Vinyals, et al., "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[25] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[26] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki, "Aligning text-to-image diffusion models with reward backpropagation," *arXiv preprint arXiv:2310.03739*, 2023.

[27] Audrunas Gruslys, Rémi Munos, Ivo Danihelka, Marc Lanctot, and Alex Graves, "Memory-efficient backpropagation through time," *Advances in neural information processing systems*, vol. 29, 2016.