Single-Image Deblurring, Trajectory and Shape Recovery of Fast Moving Objects with Denoising Diffusion Probabilistic Models

Radim Spetlik[1] Denys Rozumnyi[1,2] Jiří Matas[1]
[1]Czech Technical University in Prague, Faculty of Electrical Engineering
[2]Department of Computer Science, ETH Zurich

Abstract

Blurry appearance of fast moving objects in video frames was successfully used to reconstruct the object appearance and motion in both 2D and 3D domains. The proposed method addresses the novel, severely ill-posed, task of single-image fast moving object deblurring, shape, and trajectory recovery – previous approaches require at least three consecutive video frames. Given a single image, the method outputs the object 2D appearance and position in a series of sub-frames as if captured by a high-speed camera (i.e. temporal super-resolution). The proposed SI-DDPM-FMO method is trained end-to-end on a synthetic dataset with various moving objects, yet it generalizes well to real-world data from several publicly available datasets. SI-DDPM-FMO performs similarly to or better than recent multi-frame methods and a carefully designed baseline method.

1. Introduction

Deblurring of fast moving objects (FMO), *i.e.* objects that move over a distance larger than their size for the duration of camera exposure of a single image has recently gained significant attention [10, 11, 21–26]. Given an image of an FMO with its estimated background, the methods showed that the 2D and 3D appearance and trajectory reconstruction of FMOs is possible.

The FMO deblurring task is challenging: (i) the blurry appearance is ambiguous when the object has complex shape or texture, (ii) it is impossible to determine the direction of the FMO trajectory, and (iii) in many cases, the appearance of an FMO can hardly be distinguished from the background.

Classical deblurring methods that assume global blur effects such as defocus or camera-induced motion blur are not applicable to the problem of FMO 2D appearance and trajectory estimation [24]. Recently, [30] dealt with local blurring, but required two consecutive frames of a video as input, and together with [7] assumed only a small local blur.

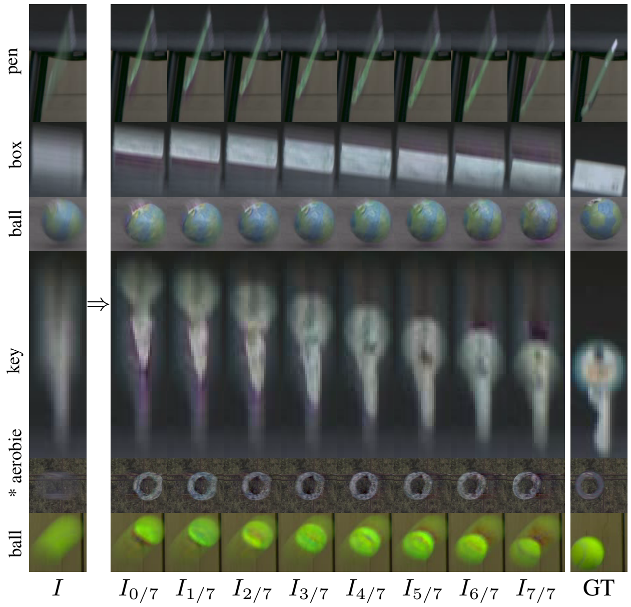


Figure 1. **Single-image temporal super-resolution.** Given a single input image I with a blurred fast moving object (first column), SI-DDPM-FMO decomposes the image into a series of deblurred sub-frames $I_{0/7}$ to $I_{7/7}$ as if captured by a high-speed camera.

All existing FMO methods alleviate the difficulty of the deblurring task by providing an estimation of the background, which greatly reduces the ambiguity of the foreground/background separation task. These methods [10, 22, 24] compute the background of the current frame as a median of the previous three or more frames, removing all FMOs. However, in many real-world cases, this is either infeasible, *e.g.*, only a single input image is available, or unstable, *e.g.*, the video is not static, and the background is changing. Thus, it greatly reduces their applicability.

We propose SI-DDPM-FMO – a method that solves this more general task with only a *single* RGB image as input (Fig. 1), in contrast to at least *three* consecutive frames of a *static-background* video, which is the limitation of previous approaches. If, for whatever reason, only a single image with a blurred object is available, we do not expect the same reconstruction quality as when several consecutive frames are available. We investigate the limits of the single-image approach. We adopt the denoising diffusion probabilistic model [27] conditioned on an image of an FMO to gradually denoise a white Gaussian noise tensor into a sequence of residual sub-frames, each accompanied by a corresponding alpha mask (Fig. 2). Adding residual sub-frames to the image conditional reconstructs the FMO's appearance, thresholding the masks reveals its shape, and the mask's center of mass estimates its trajectory. The network is trained end-to-end on a synthetic dataset with complex and highly textured objects, generated using a procedure described in [24].

To summarize our contributions, we:

(1) introduce a novel problem of single-image FMO 2D appearance and trajectory recovery,

(2) propose a novel approach setting a new state of the art in the single-image FMO 2D appearance and trajectory recovery problem,

(3) show that a customized conditional denoising diffusion probabilistic model is successfully applied to the proposed problem.

The codes and models are publicly available at https://github.com/radimspetlik/SI-DDPM-FMO.

2. Related work

The task of FMO 2D appearance and trajectory recovery is vaguely related to the task of classical deblurring methods. However, in the following paragraphs, we refer mainly to the literature specific to FMO, because there are major differences between the two areas of research in: (i) the *image formation model*, and (ii) the *task*.

Classical deblurring methods Many methods for classic image deblurring have been proposed, *e.g.*, [13]. However, as shown in [24], such methods fail on fast moving objects, and domain-specific methods are required. Similarly, existing methods for video extraction from a single blurred image [7, 14, 18, 20, 30, 31] do not assume fast motion, or require additional inputs such as more than one input frame, *e.g.*, [30], or explicit motion guidance, *e.g.*, [31].

Image formation model Classical deblurring methods describe a single color image as

$$I = H * B + N, \tag{1}$$

assuming that the entire image B is convolved with a blur kernel H, optionally with the addition of Gaussian noise N (*cf.* Eq. (2)). This definition is explicitly in, *e.g.*, [30]. It is worth mentioning that in some recent works, more advanced models are used to better capture real-world blur, *e.g.*, affine transformations [14], or neural networks [29]. Still, the blur is assumed to be present in the whole image, as in the case of blur induced by camera motion. The classical literature on deblurring also assumes a much lower level of blur than is commonly present in images of FMO.

Our method aims specifically to reconstruct the appearance and trajectory of an FMO defined in [23]. In the blurring and matting (blatting) equation [10, 11], a 2D image of the fast moving object F is introduced as

$$I = H * F + (1 - H * M)B, \tag{2}$$

where I is a single color image, H is a blur kernel, M is the shape of the object (indicator function of F), and B is the background. We use the blatting equation (2) to clearly distinguish the task of FMO 2D trajectory and appearance recovery to the task of classical deblurring methods manifested in Eq. (1).

Task FMO methods belong to a group of *temporal super-resolution* methods. In temporal super-resolution methods, the appearance is reconstructed as K images, each being exposed for $\frac{1}{K}$ of the original exposure time, forming a temporally consistent sequence $I_1, \ldots, I_K$ such that

$$\frac{1}{K}(I_1^{-1} + \cdots + I_K^{-1}) = I^{-1} \tag{3}$$

holds[1], where I^{-1} is the image I in the linear image space as captured by a camera[2].

The task of the classical deblurring methods is to reconstruct the *sharp* image. This formulation is adopted in [14], where the datasets designed to measure the performance of FMO methods [9, 10] are used to evaluate the classical deblurring task – the reconstruction of a single sharp image. The authors [14] specifically mention that "both [FMO] datasets [9, 10] do not have a sharp ground truth image." There are no *sharp* ground truth images, *i.e.* images with no blur present, in the FMO benchmarks, because FMO methods perform temporal super-resolution (Eq. 3). Two recent deblurring works [7, 30] also perform temporal super-resolution. In [30], two images are required as input,

[1] It should be noted that this image formation model does not take into account the gap between exposures of consecutive frames. When shutter is closed, some parts of object motion are not observed by a camera, which manifests in the exposure gap. Also note that in [31] the term "blur decomposition" is used to describe the same procedure.

[2] This linear space is commonly unavailable due to a highly non-linear processing introduced by camera manufactures and is therefore only approximated by applying inverse gamma calibration to RGB images.

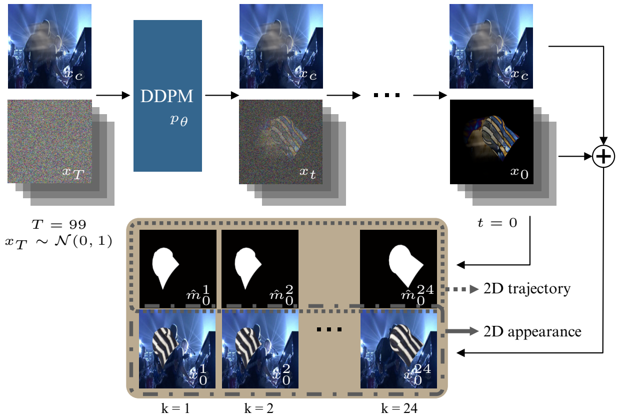


Figure 2. **SI-DDPM-FMO method**. A white Gaussian noise tensor is gradually denoised with a DDPM [27] conditioned on a single image of a fast moving object (FMO), resulting in 24 residual images with alpha masks. When added to the image of an FMO, the residual images form a series of sub-frames as if captured by a high-speed camera (*i.e.* temporal super-resolution). The trajectory of FMO is estimated as the center of mass of the alpha masks.

in both works only a small blur is assumed and neither of them provides a trajectory estimation. The 2D trajectory estimation provided by the FMO methods allows for tracking. We use the image formation model in Eq. (3) to construct our training and validation datasets.

The FMO research area is represented by a series of publications [9–11, 22–24]. In [23], detection and tracking of FMO was introduced. Their work is limited by assuming that object F must: (i) not change its appearance within an image (no rotation or rotational symmetry), (ii) have a high contrast to the background and no contact with other objects, and (iii) travel in a 2D plane parallel to the camera plane. The limitations are partially lifted by a method called Tracking by Deblatting (TbD) introduced in [10], which solves a joint deblurring and matting problem. In [21], the TbD trajectories are improved with non-causal post-processing. A special case with a planar FMO that rotates only within a 2D plane parallel to the camera plane was studied. In [32], improved motion blur prior to FMOs was proposed.

The limitations on object appearance were partially lifted by the TbD-3D [22] which assumes a piece-wise constant formation model

$$I = \sum_i H_i * F_i + (1 - \sum_i H_i * M_i)B, \qquad (4)$$

where the index i denotes one part of the complete trajectory $H = \sum_i H_i$ that traveled within a frame I, and the appearances of the sub-frames F_i and the alpha masks M_i account for the potentially changing appearance of an object (*cf.* the blatting equation Eq. (2)). TbD-3D is highly dependent on initial trajectory estimation from external module. Moreover, TbD-3D only works with objects of trivial shape (*e.g.*, a sphere) or appearance (*e.g.*, uniform color).

In [24], the first learning-based method was presented for FMO deblurring, DeFMO. The generalized formation model

$$I = \int_0^1 F_i \cdot M_i \mathrm{d}i + (1 - \int_0^1 M_i \mathrm{d}i) \cdot B, \qquad (5)$$

where the appearance of the object F_i and the alpha mask M_i is modeled by an encoder-decoder network, which generalizes to objects with a 2D appearance that can change arbitrarily along the trajectory.

The two recent papers [25, 26] explore the 3D reconstruction of FMOs. Unlike [25], the [26] method produces: (i) more complex trajectories, (ii) temporally consistent predictions, and (iii) more completely reconstructed 3D shape models.

All published FMO methods require an estimation of the background B. We are the first to study the problem of the single-image FMO reconstruction.

3. Method

In the following paragraphs, we describe a novel single-image method for reconstruction of the 2D appearance and trajectory of fast moving objects (SI-DDPM-FMO).

Simply put, we adopted the generative model [27] conditioned on an image of an FMO to gradually denoise a white Gaussian noise tensor into a sequence of 24 residual sub-frames accompanied by alpha masks (Fig. 2). The appearance is recovered by adding the residuals to the image conditional, the shape is recovered by thresholding the masks, and the trajectory is estimated as the center of mass of the masks. Our method can produce arbitrary time interpolation, as it may be applied recursively.

Denoising Diffusion Probabilistic Model (DDPM) [27] defines a diffusion process that transforms a random white Gaussian noise tensor $x_T \sim \mathcal{N}(0, 1)$ into a ground-truth residual image x_0 in T time steps and vice versa. Each step in the forward direction is given by

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t \mathbf{I}), \qquad (6)$$

where β_t is the variance of the Gaussian noise in the timestep t. The sample x_t is obtained by adding *i.i.d.* Gaussian noise with variance β_t and scaling the previous sample x_{t-1} with $\sqrt{1 - \beta_t}$ according to a variance schedule.

The DDPM, represented by a neural network with parameters θ, is trained to reverse the process in Eq. (6). Conditioned on x_c (which is I in Eq. (2)), the network predicts

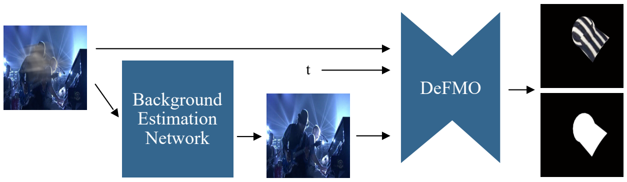


Figure 3. **Baseline – SI-DeFMO method**. A single image of a fast moving object (FMO) is fed to a background estimation network. The FMO image and estimated background, together with the time index $t \in (0,1)$, are passed to the DeFMO method [24]. The output is a deblurred object appearance with an alpha mask at time index t.

the parameters $\mu_\theta(x_t, t \mid x_c)$ and $\Sigma_\theta(x_t, t \mid x_c)$ of a Gaussian distribution

$$p_\theta(x_{t-1} \mid x_t, x_c) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t \mid x_c), \Sigma_\theta(x_t, t \mid x_c)). \tag{7}$$

The learning objective L_{vlb} for the model in Eq. (7) is derived from the variational lower bound and can be decomposed [17] as

$$\mathbb{E}_q \left[L_T + \sum_{t \geq 1} L_{t-1} - L_0 \right] \tag{8}$$

where

$$L_T = D_{\text{KL}} \left(q(x_T \mid x_0) \mid\mid p(x_T) \right), \tag{9}$$

$$L_{t-1} = D_{\text{KL}} \left(q(x_{t-1} \mid x_t, x_0) \mid\mid p_\theta(x_{t-1} \mid x_t, x_c) \right), \tag{10}$$

$$L_0 = \log p_\theta(x_0 \mid x_1, x_c). \tag{11}$$

Note that L_T does not depend on θ, L_0 is calculated according to [17], and L_{t-1} is a KL divergence between two Gaussian distributions and can therefore be evaluated in closed form.

The predicted mean $\mu_\theta(x_t, t \mid x_c)$ in Eq. (7) may be parametrized in many ways. In our training, we experimentally selected the parameterization in which the neural network predicts x_0. Accordingly, μ_θ is computed as

$$\mu_\theta(x_t, t) = \frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t}{1 - \bar{\alpha}_t} \hat{x}_0 + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_t, \tag{12}$$

where $\hat{x}_0$ is the network prediction of x_0 parametrized by θ, $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$ is the total noise variance.

Following [6], we use a simplified training objective

$$L_{\text{simple}} = \mathbb{E}_{t,(x_0, x_c)} \left[\|x_0 - \hat{x}_0\|^2 \right]. \tag{13}$$

We learn the variance $\Sigma_\theta(x_t, t | x_c)$ in Eq. (7) of the reverse process, as it reduces the number of sampling steps by an order of magnitude [17]. Using the independence property of the noise added at each step in Eq. (6), we rewrite Eq. (6) as

$$q(x_t | x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t)\mathbf{I}) \tag{14}$$

allowing efficient sampling of training data to train the reverse transition step in Eq. (7). During training, the time step t is selected with a loss-aware sampling procedure defined in [17].

We experimentally selected the number of denoising steps T=100. The input of the DDPM is a tuple of K=24 random white Gaussian tensors $x_T = (x_T^0, \dots, x_T^{K-1}) \in \mathbb{R}^{K \times (4,256,256)}$ concatenated with a single RGB image of an FMO x_c. The output $x_0 = (\hat{x}_0, \hat{m}_0)$, where $\hat{x}_0 = (\hat{x}_0^0, \dots, \hat{x}_0^{K-1}) \in \mathbb{R}^{K \times (3,256,256)}$ is trained to be a temporally consistent sequence of K sub-frame residuals with alpha masks $\hat{m}_0 = (\hat{m}_0^0, \dots, \hat{m}_0^{K-1}) \in \mathbb{R}^{K \times (1,256,256)}$ such that

$$\dot{x}^k = x_c + \dot{m}\hat{x}_0^k \tag{15}$$

is the final k-th sub-frame as if captured by a high-speed camera, where $\dot{m} = \frac{1}{K}\sum_{i=0}^{K-1} \hat{m}_0^i$. Note that the network predicts alpha masks $\hat{m}_0^k$ for each sub-frame k, but the predicted $\hat{x}_0^k$ is a residual for the whole area of the blurred FMO. The final loss is computed as

$$L = \lambda L_{\text{vlb}} + L_{\text{simple}}, \tag{16}$$

where $\lambda = 0.001$ in our experiments to make the L_{simple} main source of influence on $\mu_\theta(x_t, t \mid x_c)$ and to keep L_{vlb} guiding the training process of $\Sigma_\theta(x_t, t \mid x_c)$.

Baseline method (Fig. 3), the single-image DeFMO (SI-DeFMO), is a modified version of DeFMO [24]. It was carefully designed to provide the most fair comparison possible with our approach. DeFMO is an encoder-decoder network that requires an image of an FMO and an estimation of a background. To provide the estimation, we experimented extensively with various architectures (Residual Networks [5], Fast Fourier Convolutions [28], Visual Transformers [4]). We searched exhaustively for both the architecture and the parameterization of the network. The best results for the background estimation were produced by a DDPM conditioned on an input image. The DDPM that performs background estimation in the baseline is conditioned on a single image of an FMO I in Eq. (2) and produces background estimation $\hat{B}$. The estimated background is fed to the DeFMO network together with I. We discuss more about the training of the baseline method and our design choices in Sec. 5.

Training settings The neural networks were trained with the AdamW [16] optimizer with fixed learning rate 10^{-4}.

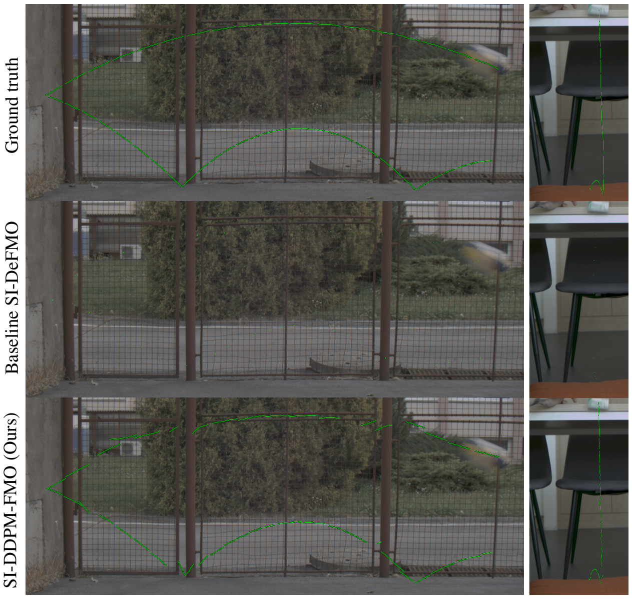


Figure 4. **Trajectory estimation** on sequences from the *Falling Objects* dataset [9] (right) and the *TbD-3D* dataset [22] (left). Best viewed zoomed in. Compare almost-static trajectory predictions (almost dots) of baseline with the proposed SI-DDPM-FMO.

The benchmark (see Sec. 4) requires the reconstructed sub-frame images to have a size of 320×240 px. We design the baseline method SI-DeFMO to work at the same resolution and resize the 256×256 px outputs of the SI-DDPM-FMO method to the requested resolution with bicubic interpolation. Both methods are implemented in PyTorch [19] and trained on a cluster of 12 to 48 NVIDIA V100 40GB GPUs.

4. Datasets

Fast moving object deblurring benchmark [24] (FMO benchmark) is a publicly available benchmark[3] that consists of three real-world datasets and a Python code designated to easily test novel methods. The TbD dataset [10] is made up of 12 high-speed videos captured at 240 fps in raw format with full exposure (total 471 frames). Low-speed videos at 30 fps were created by temporal averaging. Sharp appearances, masks, and full trajectory are provided. The dataset contains only sport videos with mostly spherical objects with little change in appearance over time. 10 sequences (total 516 frames) of objects that significantly change their appearance within a low-speed video frame are recorded in the TbD-3D dataset [22], which addresses the mentioned shortcomings of the TbD dataset. The Falling Objects dataset [9] is the first to contain objects with non-trivial shapes, *i.e.* box, marker, pen, key, cell, eraser (6 sequences, total of 94 frames). Originally, only high-speed videos were available in the dataset. Later, the authors of the benchmark also added trajectories. Note that the FMO dataset [23] is not included in the benchmark, as it contains neither trajectories nor ground-truth appearances.

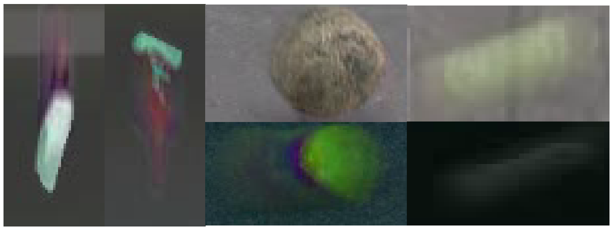


Figure 5. **Failure cases of SI-DDPM-FMO.** We show three groups, six samples in total, of the most common failures of the proposed method. Samples from FMO benchmark (Sec. 4).

We created **the synthetic training dataset** following the procedure described in [24]. An *object* was rendered with 6D *trajectory* on a *background* frame using Blender Cycles [3] resulting in a set of sub-frame renderings. To generate the low-speed frame showing the blurred FMO, the image formation model in Eq. (5) is applied.

Objects are sampled from 3D models of the 50 largest classes of the ShapeNet [1] dataset, each class being represented uniformly. There are 1600 DTD [2] textures used in the training. *Trajectories* are sampled uniformly as in [24]. *Backgrounds* are sampled from the VOT [12] training sequences. In total, there are 50000 training images.

5. Experiments

The three datasets from the FMO Benchmark (Sec. 4) were used to evaluate SI-DDPM-FMO and compare it with the SI-DeFMO baseline method and the multi-frame methods. Similarly to [24], we use PSNR (peak signal-to-noise ratio), SSIM (structural similarity index measure), and TIoU (intersection over union averaged along the trajectory) [10] as evaluation metrics. We compare to all multi-frame, or background-requiring, FMO methods that reconstruct 2D appearance and trajectory, namely TbD [10], TbD-3D [22], and DeFMO [24]. We do not compare with recent 3D shape and motion reconstruction methods [25,26] since our method only produces a 2D appearance and position.

The number of sub-frames that our method produces is fixed to 24. Since the high-speed videos in the benchmark are available at 8 times higher frame rate than the low-speed videos, we average every three sub-frames produced by SI-DDPM-FMO to generate full exposure temporal super-resolution to match the high-speed frames. SI-DDPM-FMO performs temporal super-resolution from a single input image. Therefore, the direction of time is ambiguous. To account for that, we compute metrics for both directions and

[3] https://github.com/rozumden/fmo-deblurring-benchmark

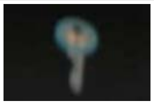
Dataset	Typical Object	Score	Inputs		Background-Requiring Methods			Baseline	Proposed
			B	*I*	TbD [10]	TbD-3D [22]	DeFMO [24]	SI-DeFMO	SI-DDPM-FMO
Falling [9]		TIoU ↑	-	-	0.545	0.545	0.703 ①	0.506 ③	0.563 ②
		PSNR ↑	19.68	23.73	22.09	23.01	26.80 ①	24.16 ③	24.77 ②
		SSIM ↑	0.459	0.593	0.617	0.695 ②	0.752 ①	0.632	0.645 ③
TbD-3D [22]		TIoU ↑	N/A	N/A	0.539	0.539	0.850 ①	0.666 ③	0.770 ②
		PSNR ↑	19.81	24.80 ②	19.67	21.99	26.23 ①	23.63	24.57 ③
		SSIM ↑	0.425	0.640 ③	0.483	0.621	0.699 ①	0.599	0.653 ②
TbD [10]		TIoU ↑	-	-	0.601 ①	0.601 ①	0.583 ②	0.407	0.551 ③
		PSNR ↑	21.51	25.07 ③	23.99	24.57	25.54 ①	24.46	25.26 ②
		SSIM ↑	0.467	0.569	0.612 ②	0.678 ①	0.599 ③	0.547	0.587
Runtime (on 240 × 320)			-	-	0.01fps	0.001fps	20fps	4 fps	5 fps

Table 1. **Evaluation on the Fast Moving Object Deblurring Benchmark [24].** Metrics: peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and intersection over union averaged along the trajectory (TIoU) [10]. The datasets are sorted by decreasing difficulty: *Falling Objects* [9] – arbitrarily shaped and textured, *TbD-3D* [22] – significantly textured but mostly spherical, *TbD* [10] – mostly spherical and uniformly colored objects. Runtime is the output capacity, *i.e.*, how many images per second the method generates.

report the best one (for all methods). Sub-frame trajectory (Fig. 4) is estimated as the center of mass of the generated alpha masks $\hat{m}_0$ (Sec. 3).

Note that we follow the same evaluation methodology as [24]. We conduct the same quantitative examination and repeat selected qualitative experiments, because we want to make the comparison with the relevant area of research as easy and as fair as possible.

5.1. Baseline – SI-DeFMO

In this section, we describe the baseline method. For a fair and equitable comparison of the two closely related problems, we decided to modify the existing DeFMO [24] method. DeFMO requires as input a single image of an FMO and, in contrast to the proposed SI-DDPM-FMO (Fig. 2), an estimation of a background.

The easiest way to provide an estimate of a background from a single image of an FMO is to train a neural network to do so, which we did (Fig. 3). If we assume a simple FMO image formation model in Eq. (2), the task is a combination of reflection removal and inpainting.

We could not use a state-of-the-art inpainting method because the task of inpainting is to fill a specified part of an image with some visually plausible content. Inpainting assumes that there is no signal to be recovered. By definition, an FMO does not completely occlude a background – the signal of the background is attenuated. Therefore, we did not use inpainting, and instead we trained a designated network.

We also could not use a state-of-the-art reflection removal method – reflection removal methods assume that the reflection is uniformly blended with the background, forming an image akin to a double-exposure photo. However, the FMO is spatially localized in the image. Note that there exists a line of reflection removal research (*e.g.*, [8]) where a non-uniform blend of the foreground and background layers is studied. Still, it is tailored to glass-reflection setting, does not assume appearance of fast moving objects, and cannot be used to estimate a background without an FMO.

Therefore, we searched exhaustively for the best architecture and configuration of the background-estimation network. The best results were achieved with a DDPM model similar to that used in our method. In our SI-DDPM-FMO method, the DDPM performs a temporal super-resolution. The task of DDPM in the baseline method is a simultaneous segmentation and removal of an FMO. The input of the baseline DDPM network is a single RGB image with an FMO, the output is the background estimation required by the DeFMO.

The baseline DDPM network was trained on a synthesized dataset with FMOs from the dataset described in Sec. 4 and backgrounds from the *train2017* subset of MS-COCO [15]. The background-estimation network was trained with a simple ℓ_2 loss.

5.2. FMO benchmark

We compared the proposed method against the multi-frame FMO methods, which require a background estimation, on the FMO benchmark (Sec. 4). The results are shown in Tab. 1. When available, we include metrics also for the input of the methods: a background estimation B and an image of an FMO I.

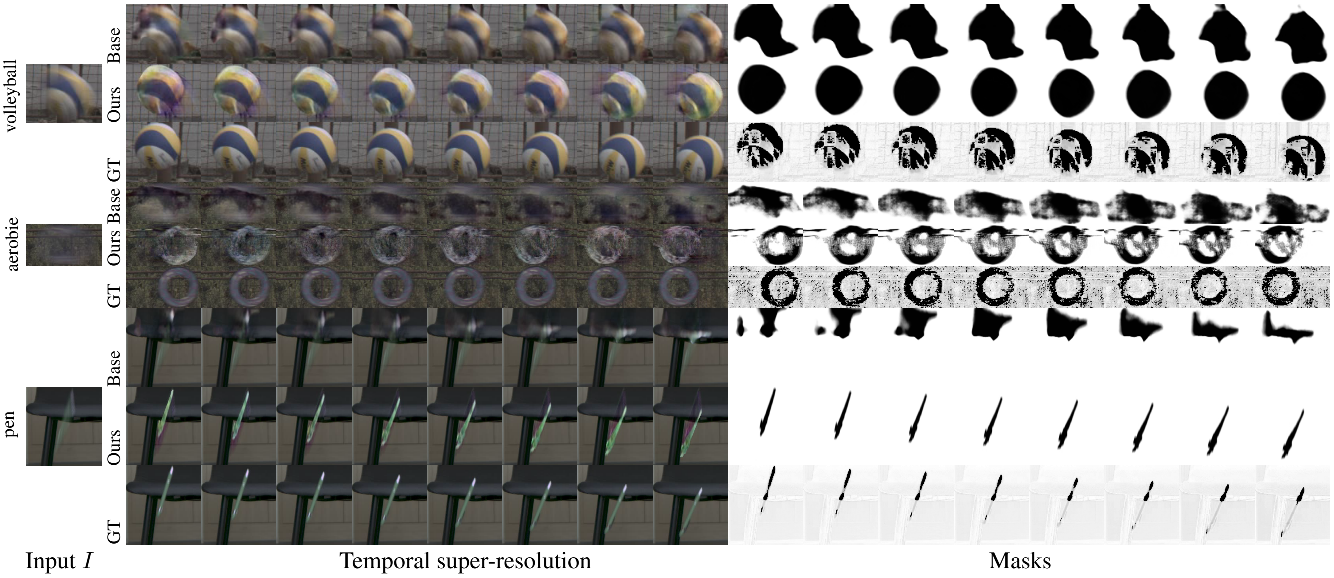


Figure 6. **Temporal super-resolution, by a factor of 8, on selected sequences from test datasets.** The SI-DeFMO baseline method (Sec. 3) compared with the ground truth of high-speed footage (GT). Ground-truth masks were computed as a difference image between the GT sub-frames and the background. The proposed SI-DDPM-FMO and baseline method generates all outputs just from a single input image I on the left.

Falling Objects dataset [9] is a challenging dataset; for results, see the top block of Tab. 1. the proposed method outperforms two multi-frame methods and scores similarly as the baseline method. The clear winner in 2D appearance and trajectory estimation is the multi-frame approach [24].

TbD-3D dataset [22] results are shown in the middle block of Tab. 1. Here, DeFMO also performs the best in all metrics. However, this time the performance of the proposed method is much closer to that of the best method than in the Falling Objects dataset.

TbD dataset [10] is the most challenging dataset for learning-based methods. It consists mostly of spherical objects with a constant appearance. We find that the multi-frame DeFMO method performs in a way comparable to our single-frame SI-DDPM-FMO method. We believe that the results of the two methods are caused by specifics of the training dataset (Sec. 4) – mostly spherical objects with constant appearance are not included, and therefore we cannot expect the models to generalize well in this setting. Commonly, objects in the TbD dataset occupy only a few per cent of an image. We believe that the performance of the proposed SI-DDPM-FMO would be improved by introducing the aforementioned spherical objects in our datasets. It is also worth noting that the TbD(-3D) methods are designed to address specifics of the two respective datasets. Still, SI-DDPM-FMO gives similar or better results than these tailored methods.

Discussion Overall, the proposed method performs well. In most cases, it gives better results than the multi-frame TbD and TbD-3D methods in the FMO benchmark. The performance gap between the single-image and multi-frame methods shows the importance of having a good estimation of background. In multi-frame methods, the background is estimated as a median of three consecutive frames with the middle frame being an input to a method. The baseline method, which replaces the multi-frame median background estimation by a single-frame dedicated DDPM background estimation, produces worse results than our proposed method in all cases. Therefore, the proposed method sets a new state of the art in the novel task of *single-image* FMO 2D appearance and trajectory estimation.

We provide a qualitative study in Fig. 6. Alpha masks produced by our method greatly resemble the ground-truth masks.

To complete the image of the single-frame versus multi-frame FMO deblurring methods, we reconstruct the comparison presented in [24]. The results are in Fig. 7. Note that we did not cherry pick the results; we exactly reproduced the work of [24] and added the output of the baseline and proposed methods accordingly. In the SI-DDPM-FMO column, we find very clear masks and spatially precise 2D appearance reconstructions. Even compared to the best-in-benchmark, multi-frame DeFMO method, our reconstructions seem to have similar visual quality.

More examples and videos are available in the supplementary material.

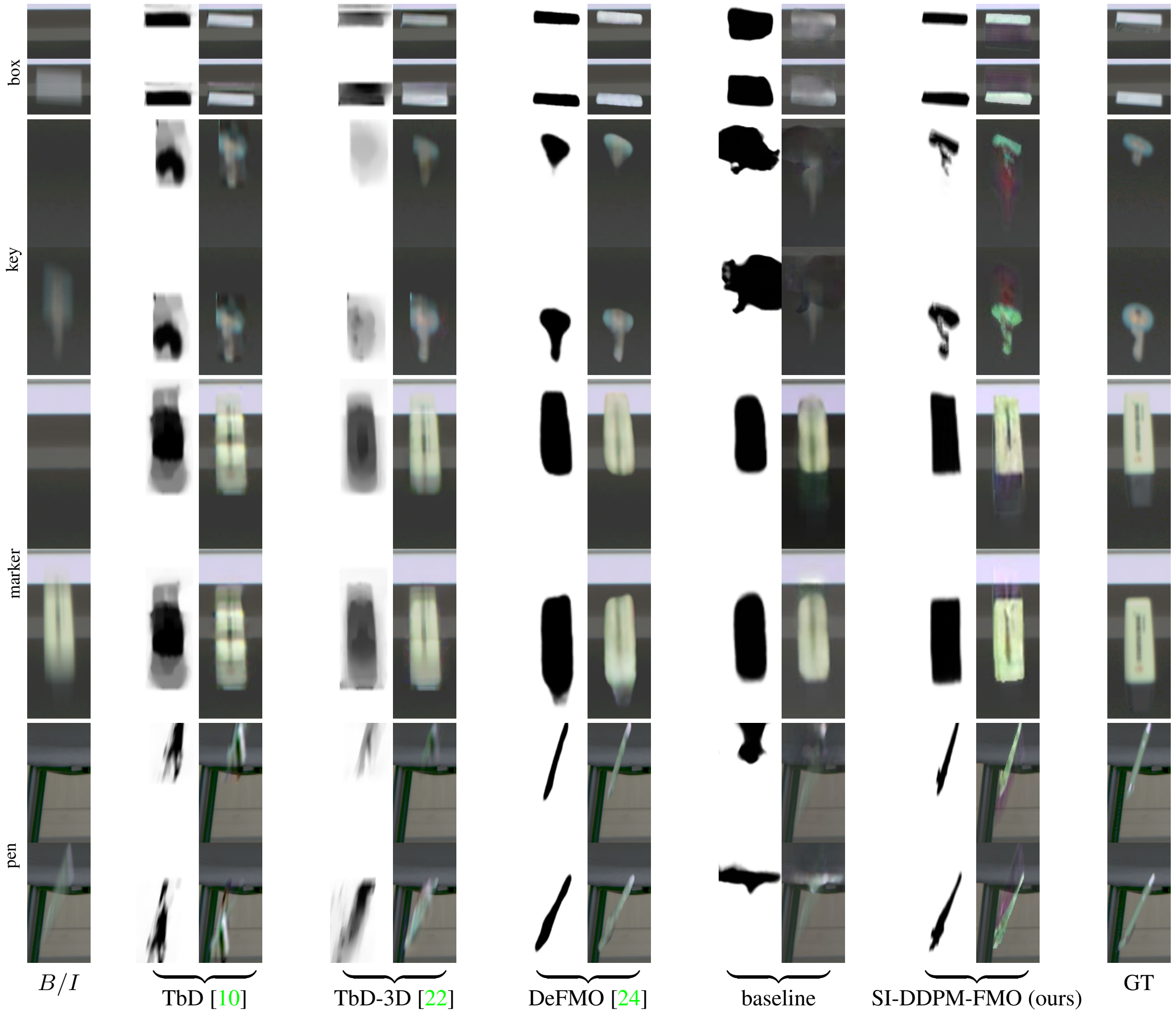


Figure 7. **Comparison on the Falling Objects dataset [24]** with the multi-frame methods: TbD [10], TbD-3D [10], TbD-3D [22], and with the SI-DeFMO baseline (Sec. 3). For each method, we show the estimated alpha mask and the first and the last sub-frame of the generated temporal super-resolution sequence of length 8.

5.3. Failure Cases

Single-image temporal super-resolution is a severely ill-posed problem (Sec. 1). Therefore, we did not expect to over-perform existing multi-frame methods. Our goal was to get as close to their performance as possible. Fig. 5 exemplifies three of the most common failure groups of our SI-DDPM-FMO method. We hypothesize that the failures could be solved, at least partially, by introducing more diverse training data, specifically images of backgrounds similar to those we see in the FMO benchmark.

Furthermore, the proposed SI-DDPM-FMO cannot recover FMOs that are similar in color to the background (see the marker cap in Fig. 7). The method is also not designed for transparent objects such as a bottle or glass.

6. Conclusions

We studied a novel, severely ill-posed, problem of single-image fast moving object 2D appearance and trajectory estimation. We proposed a method based on denoising diffusion probabilistic models, which performs better than a carefully designed baseline built on a recent state-of-the-art method. Experimental results show that the proposed method handles real-world fast moving objects with complex shapes and significant appearance changes well.

Acknowledgements. This research was supported by the Czech Technical University in Prague grant No. SGS23/173/OHK3/3T/13, by BMK, BMAW, and the State of Upper Austria within the SCCH competence center INTEGRATE (grant no. 892418) part of the FFG COMET Competence Excellent Technologies Programme, and the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254).

References

[1] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository, Dec. 2015. arXiv:1512.03012 [cs]. 5

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing Textures in the Wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, June 2014. ISSN: 1063-6919. 5

[3] Blender Online Community. Blender - a 3D modelling and rendering package, 2023. 5

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Oct. 2020. 4

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*, Dec. 2015. arXiv: 1512.03385. 4

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. 4

[7] Meiguang Jin, Givi Meishvili, and Paolo Favaro. Learning to Extract a Video Sequence from a Single Motion-Blurred Image. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6334–6342, June 2018. ISSN: 2575-7075. 1, 2

[8] N. Kong, Y. Tai, and J. S. Shin. A Physically-Based Approach to Reflection Separation: From Physical Modeling to Constrained Optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):209–221, Feb. 2014. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. 6

[9] Jan Kotera, Jiri Matas, and Filip Sroubek. Restoration of fast moving objects. *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, PP, Aug. 2020. 2, 3, 5, 6, 7

[10] Jan Kotera, Denys Rozumnyi, Filip Šroubek, and Jiří Matas. Intra-Frame Object Tracking by Deblatting. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2300–2309, Oct. 2019. ISSN: 2473-9944. 1, 2, 3, 5, 6, 7, 8

[11] Jan Kotera and Filip Šroubek. Motion Estimation and Deblurring of Fast Moving Objects. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2860–2864, Oct. 2018. ISSN: 2381-8549. 1, 2, 3

[12] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomáš Vojíř, Roman Pflugfelder, Gustavo Fernández, Georg Nebehay, Fatih Porikli, and Luka Čehovin. A Novel Performance Evaluation Methodology for Single-Target Trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2137–2155, Nov. 2016. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. 5

[13] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. DeblurGAN-v2: Deblurring (Orders-of-Magnitude) Faster and Better. pages 8878–8887, 2019. 2

[14] Daoyu Li, Liheng Bian, and Jun Zhang. High-Speed Large-Scale Imaging Using Frame Decomposition From Intrinsic Multiplexing of Motion. *IEEE Journal of Selected Topics in Signal Processing*, 16(4):700–712, June 2022. Conference Name: IEEE Journal of Selected Topics in Signal Processing. 2

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context. *arXiv:1405.0312 [cs]*, Feb. 2015. arXiv: 1405.0312. 6

[16] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. *arXiv:1711.05101 [cs, math]*, Jan. 2019. arXiv: 1711.05101. 4

[17] Alexander Quinn Nichol and Prafulla Dhariwal. Improved Denoising Diffusion Probabilistic Models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8162–8171. PMLR, July 2021. ISSN: 2640-3498. 4

[18] Wenjia Niu, Kaihao Zhang, Wenhan Luo, and Yiran Zhong. Blind Motion Deblurring Super-Resolution: When Dynamic Spatio-Temporal Learning Meets Static Image Understanding. *IEEE Transactions on Image Processing*, 30:7101–7111, 2021. Conference Name: IEEE Transactions on Image Processing. 2

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 5

[20] Kuldeep Purohit, Anshul Shah, and A. N. Rajagopalan. Bringing Alive Blurred Moments. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6823–6832, June 2019. ISSN: 2575-7075. 2

[21] Denys Rozumnyi, Jan Kotera, Filip Šroubek, and Jiří Matas. Non-causal Tracking by Deblatting. In Gernot A. Fink, Simone Frintrop, and Xiaoyi Jiang, editors, *Pattern Recognition*, Lecture Notes in Computer Science, pages 122–135, Cham, 2019. Springer International Publishing. 1, 3

[22] Denys Rozumnyi, Jan Kotera, Filip Šroubek, and Jiří Matas. Sub-Frame Appearance and 6D Pose Estimation of Fast Moving Objects. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6777–6785, June 2020. ISSN: 2575-7075. 1, 3, 5, 6, 7, 8

[23] Denys Rozumnyi, Jan Kotera, Filip Šroubek, Lukáš Novotný, and Jiří Matas. The World of Fast Moving Objects. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4838–4846, July 2017. ISSN: 1063-6919. 1, 2, 3, 5

[24] Denys Rozumnyi, Martin R. Oswald, Vittorio Ferrari, Jiří Matas, and Marc Pollefeys. DeFMO: Deblurring and Shape

Recovery of Fast Moving Objects. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3455–3464, June 2021. ISSN: 2575-7075. 1, 2, 3, 4, 5, 6, 7, 8

[25] Denys Rozumnyi, Martin R. Oswald, Vittorio Ferrari, and Marc Pollefeys. Shape from Blur: Recovering Textured 3D Shape and Motion of Fast Moving Objects. In *Advances in Neural Information Processing Systems*, volume 34, pages 29972–29983. Curran Associates, Inc., 2021. 1, 3, 5

[26] Denys Rozumnyi, Martin R. Oswald, Vittorio Ferrari, and Marc Pollefeys. Motion-From-Blur: 3D Shape and Motion Estimation of Motion-Blurred Objects in Videos. pages 15990–15999, 2022. 1, 3, 5

[27] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 2256–2265, Lille, France, July 2015. JMLR.org. 2, 3

[28] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust Large Mask Inpainting with Fourier Convolutions. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3172–3182, Jan. 2022. ISSN: 2642-9381. 4

[29] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Björn Stenger, Wei Liu, and Hongdong Li. Deblurring by Realistic Blurring. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2734–2743, June 2020. ISSN: 2575-7075. 2

[30] Zhihang Zhong, Mingdeng Cao, Xiang Ji, Yinqiang Zheng, and Imari Sato. Blur Interpolation Transformer for Real-World Motion from Blur. Nov. 2022. 1, 2

[31] Zhihang Zhong, Xiao Sun, Zhirong Wu, Yinqiang Zheng, Stephen Lin, and Imari Sato. Animation from Blur: Multimodal Blur Decomposition with Motion Guidance. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pages 599–615, Cham, 2022. Springer Nature Switzerland. 2

[32] Filip Šroubek and Jan Kotera. Motion Blur Prior. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 928–932, Oct. 2020. ISSN: 2381-8549. 3