Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Few-Step Cold Diffusion with Blur-Adaptive Richardson–Lucy Data Consistency for Endoscopic Image Deblurring

**Kyosuke Miura**
Shizuoka University
**Shin Usuki**
usuki@shizuoka.ac.jp

Shizuoka University

**Research Article**

**Keywords:** Image deblurring, Cold diffusion, Richardson–Lucy deconvolution, Data consistency, Endoscopy

**Posted Date:** November 24th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11672495/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Few-Step Cold Diffusion with Blur-Adaptive Richardson–Lucy Data Consistency for Endoscopic Image Deblurring

Kyosuke Miura[1*] and Shin Usuki[1]

[1*]Shizuoka University, 3-5-1 Johoku, Chuo-ku, Hamamatsu, 432-8561, Shizuoka, Japan.

*Corresponding author(s). E-mail(s): miura.kyosuke.21@shizuoka.ac.jp;

**Abstract**

Blurred endoscopic images obscure fine mucosal and vascular patterns, and the point spread function (PSF) of each frame is unknown. Diffusion models restore perceptually sharp images but need many network evaluations and may synthesise details unsupported by the observation. We propose a few-step Cold Diffusion method for blind deblurring in which every state of the reverse process is a physically blurred image. A degradation estimator predicts the blur strength and noise level of each image. The network is trained on a schedule of 50 blur strengths, but sampling starts from the time index of the predicted blur strength and visits only five evenly spaced indices of the schedule down to zero, so that each index keeps the blur strength it represented in training. In addition, the predicted noise level sets how strongly the update rule attenuates the observation noise carried over between steps. Finally, Richardson–Lucy data consistency is applied only to images predicted to be strongly blurred. As a result, on EDD2020 images with Gaussian blur and noise, the method achieved the lowest perceptual error (LPIPS, 0.101) among the compared deblurring methods, including direct regression and a conditional diffusion model, and was about six times faster than the diffusion model with 50 sampling steps. A null-space analysis of hallucinations showed that, on these images, the method recovered more of the true detail lost to blur than any compared method and, unlike the diffusion model, produced no seed-dependent content. This makes the method promising for reproducible restoration of endoscopic images.

**Keywords:** Image deblurring, Cold diffusion, Richardson–Lucy deconvolution, Data consistency, Endoscopy

# 1 Introduction

Endoscopy allows clinicians to inspect the mucosa of the gastrointestinal tract directly and is the standard examination for detecting cancers and polyps at an early stage. Endoscopic images, however, are frequently degraded by blur. Because the camera operates in a dimly lit body cavity and the scope is moved by hand, motion blur is common in endoscopic video [1, 2], and because endoscopes are equipped with a wide-angle lens that cannot be focused, out-of-focus frames are also frequent [3]. Blur obscures fine vascular patterns and mucosal textures, and a blurred frame may have to be retaken. Restoring sharp images from blurred frames after acquisition is therefore of practical value.

Blur is classically described as the convolution of the latent image with a point spread function (PSF), and deconvolution methods such as the Richardson–Lucy (RL) algorithm [4, 5] invert this model iteratively. These methods require the PSF to be known. In endoscopy, the PSF changes with the distance to the tissue and with the motion of the scope, so measuring it for every frame is impractical, and blind deconvolution, which estimates the PSF and the image jointly, is notoriously ill-posed [6]. In our previous work [7], we estimated the PSF patch by patch with a convolutional neural network and then applied RL deconvolution. This two-stage approach was effective to a certain extent, but it is limited in that it depends on the accuracy of the estimated PSF in every patch and estimation errors propagate directly into the deconvolution.

Learning-based methods avoid explicit PSF modelling. Networks that regress the sharp image directly from the blurred one, such as NAFNet [8] and Restormer [9], achieve high peak signal-to-noise ratios (PSNR). Because they are trained to minimise a pixel-wise loss, however, their outputs tend towards the conditional mean and look smoother than real images, which is the perception–distortion trade-off [10]. Diffusion models [11, 12] produce perceptually sharper results by generating the image from Gaussian noise conditioned on the observation. This generation is stochastic: different noise samples yield different images, and structures that are not supported by the observation may be synthesised, which is a concern for medical images. Diffusion models also require many network evaluations per image. Methods that enforce consistency with the observation during sampling, such as DPS [13], DDNM [14] and an RL-based correction for fluorescence microscopy [15], assume that the degradation operator, including the PSF, is known.

Cold Diffusion [16] replaces the Gaussian noise of diffusion models with a deterministic degradation. When the degradation is Gaussian blur, the forward process gradually blurs a sharp image, every intermediate state is itself a blurred image, and the reverse process removes the blur step by step without injecting noise. The blur operator of the forward process can thus serve as the observation model, so that each reverse step can be checked against a physically meaningful image. Turning this idea into a practical method for endoscopic images, however, raises four problems. (i) Sampling with few steps is needed for speed, but the time indices used at inference must correspond to the blur strengths seen during training. (ii) Real images contain observation noise, and the Cold Diffusion update rule carries the noise of the input through all reverse steps. (iii) The blur strength of a given image is unknown, whereas the reverse process must start from the time index that matches it. (iv) Data consistency

based on RL deconvolution does not always help: with a strong predictor, applying it to every image lowers the PSNR, and its benefit depends on the blur strength.

In this paper, we address these problems and propose a few-step Cold Diffusion method for blind deblurring of endoscopic images. Our contributions are as follows.

- We formulate few-step sampling that keeps the correspondence between time indices and blur strengths of the training schedule, and restore an image in five reverse steps (Section 3.2).
- We generalise the Cold Diffusion update rule with a coefficient that controls how much of the input residual, including observation noise, is carried to the next step, and set the coefficient from the estimated noise level (Section 3.3).
- We introduce a lightweight degradation estimator that predicts the blur strength and the noise level of the input, which determine the starting time index, the update coefficient and the PSF used for data consistency, so that no PSF has to be measured (Section 3.4).
- We propose a blur-adaptive RL data consistency (RL-DC) that applies a weak RL correction, with the PSF given by the estimated blur strength, only to strongly blurred images, and we show when data consistency improves or degrades the restoration (Section 3.5).
- We evaluate hallucinations with the generalised null space of the blur operator [17], comparing the proposed method with direct regression and with a conditional diffusion model that uses the same network, and we identify which part of the method changes the null-space content of the restoration (Section 5.5).

We evaluate the method with two predictor networks, an improved TransUNet and a time-conditioned NAFNet, on 386 images from the EDD2020 dataset [18] with Gaussian blur and observation noise, and compare it with direct regression (NAFNet and Restormer), a conditional denoising diffusion model and RL deconvolution with the true PSF. With the NAFNet predictor, the proposed method reaches a PSNR 0.25 dB below direct regression with the same network, while reducing the learned perceptual image patch similarity (LPIPS) [19] from 0.183 to 0.101. A conditional diffusion model with the same network and training data does not reach this LPIPS even with 50 sampling steps, at which point it is about six times slower than the proposed method and about 1 dB lower in PSNR, and its outputs vary with the random seed. On the noise set, no test image is restored below its input PSNR at any tested noise level, up to twice the maximum noise level used in training. In the null-space analysis of hallucinations, the proposed method recovers more of the true detail lost to blur than any compared method on the main test set.

# 2 Related Work

## 2.1 Classical deconvolution

Deconvolution reconstructs the sharp image by reversing the blurring process with the PSF. For a spatially invariant blur, the observation is modelled as $y = h*x+n$, where $h$ is the PSF, $x$ the sharp image and $n$ noise, and non-blind deconvolution recovers $x$ from $y$ and a known $h$. The Wiener filter [20] gives the linear estimate with the minimum

mean squared error, but it requires the power spectra of the image and the noise, and its output suffers from ringing artefacts around edges and near image borders [21]. The RL algorithm [4, 5], in contrast, updates the estimate multiplicatively (Eq. (7)), which keeps it non-negative, and is the expectation–maximisation algorithm for the maximum-likelihood estimate under Poisson noise [22]. However, because this estimate is not regularised, noise is amplified as the iterations proceed [23], and the number of iterations acts as an implicit regulariser. To address this, explicit priors have been introduced to stabilise the inversion: total variation [24] has been added to RL as a regularisation term [23], and hyper-Laplacian priors on image gradients, which follow the heavy-tailed statistics of natural images, allow fast non-blind deconvolution [25]. All of these methods, however, assume that the PSF is known.

When the PSF is unknown, blind deconvolution estimates it together with the image. Fish et al. [26] alternated RL updates of the image and the PSF and extended the algorithm to PSFs with a few unknown parameters, and Fergus et al. [27] estimated camera-shake kernels by variational inference with a heavy-tailed prior on image gradients. Levin et al. [6], on the other hand, showed that maximising the joint posterior of the image and the kernel favours the no-blur explanation, whereas estimating the kernel alone is much better constrained. Furthermore, the problem becomes much better posed when the PSF is restricted to a parametric family. For example, Zhuo and Sim [28] modelled defocus blur as a Gaussian and estimated the blur strength at edges from the ratio of the gradient magnitudes of the input and a re-blurred copy. Our degradation estimator likewise reduces the PSF to a single blur-strength parameter, but predicts it from the whole image rather than only at edges, together with the noise level, with a network.

## 2.2 Learning-based deblurring

Learning-based methods learn the mapping from blurred to sharp images without an explicit PSF. Nah et al. [29] restored dynamic scene blur end to end with a multi-scale convolutional network. Subsequently, DeblurGAN [30] trained a conditional generative adversarial network with a content loss, and multi-stage [31], Transformer (Restormer) [9] and simplified convolutional (NAFNet) [8] architectures have steadily raised the PSNR on standard benchmarks. However, networks trained with a pixel-wise loss approximate the conditional mean and therefore produce over-smoothed images. Adversarial training, on the other hand, improves the perceptual quality but at the cost of distortion, and Blau and Michaeli [10] showed that this perception–distortion trade-off is fundamental. Other networks embed the image formation model, for example by unrolling RL iterations [32] or applying a Wiener filter in feature space [33], but they still produce the output in a single forward pass and therefore cannot check the output against the observation and correct it. In this paper, we use NAFNet and Restormer as direct-regression baselines and also use NAFNet as a predictor in our sampler, which separates the effect of the sampling procedure from that of the network architecture.

## 2.3 Diffusion models for deblurring

Diffusion models have attracted attention as generative methods that avoid this over-smoothing. Denoising diffusion probabilistic models (DDPMs) [11] generate images by reversing a gradual Gaussian noising process, and DDIM [34] made the sampling deterministic and reduced the number of steps. Conditioned on a low-resolution image, diffusion models perform super-resolution by iterative refinement [12]. For deblurring, DvSR [35] refined the output of a deterministic predictor with a conditional diffusion model and generated diverse plausible restorations of the same input. Furthermore, HI-Diff [36] runs the diffusion in a compact latent space and fuses the generated prior into a regression-based deblurring network at multiple scales, and Ren et al. [37] guided a conditional diffusion model with multiscale structures of the sharp image predicted by a regression network, which improved the robustness to images from unseen domains. However, in these methods, the diffusion starts from Gaussian noise and the observation enters only as a condition of the network, so the intermediate states cannot be compared with the observation directly.

To address this, methods that start from the degraded image instead of noise have been proposed. Cold Diffusion [16] replaces the noise by an arbitrary deterministic degradation such as blur, and Hoogeboom and Salimans [38] showed that blur can be written as a Gaussian diffusion with frequency-dependent noise levels. Subsequently, InDI [39] restored an image in small steps along the linear interpolation between the degraded and the sharp image and controlled the perception–distortion trade-off by the number of steps. I2SB [40] learns a Schrödinger bridge between degraded and sharp images, and RDDM [41] separates the diffusion of the residual between them from the diffusion of noise and restores images in two to five deterministic steps. BlurDM [42] models the formation of motion blur as a dual blur–noise diffusion in a latent space. However, in InDI, I2SB and RDDM, the intermediate states are interpolations between the degraded and the sharp image, possibly with added noise, and BlurDM diffuses latent features, so a known degradation operator cannot be applied to the intermediate states. Only in Cold Diffusion and the methods built on it is every state a degraded image with a known operator. Cold Diffusion, however, accumulates errors over the steps; for sparse-view computed tomography, ReCo-Diff [43] reduced this error accumulation by feeding the residual between the current state and the re-degraded prediction back to the network. Our method keeps the Cold Diffusion formulation, in which every state is a Gaussian-blurred image, so that RL deconvolution with the corresponding PSF can be applied during sampling. Instead of conditioning the network on the residual as in ReCo-Diff, we attenuate the residual with a coefficient set from the estimated noise level. We do not compare with these diffusion-based deblurring methods directly. They were developed for colour images at much higher resolution, mostly with motion blur, and their published models cannot be applied to our $100 \times 100$ greyscale images with Gaussian blur and noise; retraining each of them in our setting would make the comparison depend on our own reimplementation and tuning. Instead, we compare with a conditional DDPM trained on the same data, which represents diffusion from Gaussian noise with the observation as a condition, and with the direct-regression networks NAFNet and Restormer (Section 4.2).

### 2.4 Data consistency with diffusion priors

When the degradation operator is known, a pretrained unconditional diffusion model can be steered towards the observation without retraining. DDRM [44] performs the sampling in the spectral domain of the operator, DPS [13] adds the gradient of the data-fidelity term at each step, and DDNM [14] replaces the range-space component of the estimate with that of the observation. However, these methods assume that the operator is known. For blind problems, in which the operator is unknown, BlindDPS [45] runs diffusion models of the image and the kernel in parallel, and GibbsDDRM [46] samples the image and the kernel parameters alternately. Both, however, run 100 or more reverse steps per image, and the restoration quality depends strongly on the accuracy of the estimated kernel. Closest to our RL-DC, Chen and Howard [15] concurrently applied RL iterations to the denoised estimate at each step of a score-based diffusion model for fluorescence microscopy images with Poisson noise. However, their method assumes a known PSF and leaves blind deconvolution to future work. In contrast, our RL-DC uses the PSF given by the estimated blur strength, runs within five deterministic steps, and is applied only to strongly blurred images.

### 2.5 Endoscopic image restoration

Blur is a common artefact in endoscopic video, along with specular reflections, saturation and bubbles, and is one of the artefact classes of the artefact detection sub-challenge (EAD2020) of EndoCV2020 [47]; the EDD2020 dataset used in this paper comes from its disease detection sub-challenge [18]. Ali et al. [2] detected six types of artefacts, scored the quality of each frame and restored mildly degraded frames with artefact-specific conditional generative adversarial networks. Their deblurring network was trained on sharp frames blurred with synthetic motion trajectories, as our networks are trained on synthetically blurred images. More recently, EndoIR [48] restored several endoscopic degradations, such as low light, smoke and bleeding, with a single diffusion model. However, blur was not among them, and to our knowledge, the reliability of diffusion-based deblurring of endoscopic images, in terms of hallucinated structures and the dependence of the output on the random seed, has not been examined; Section 5.5 addresses this question.

## 3 Method

The proposed method extends a preliminary version presented at ICPE 2026 [49]. Figure 1 gives an overview of the proposed method. Given a blurred and noisy observation $y$, a degradation estimator first predicts the blur strength and the noise level of $y$ (Section 3.4). The estimated blur strength determines the time index from which the reverse process starts and the PSF used for data consistency, and the estimated noise level determines the coefficient of the update rule (Section 3.3). The reverse process then visits five time indices of the training schedule (Section 3.2). At each step, a time-conditioned predictor network (Section 3.6) predicts the sharp image, the prediction is corrected by RL deconvolution if the image is strongly blurred (Section 3.5), and

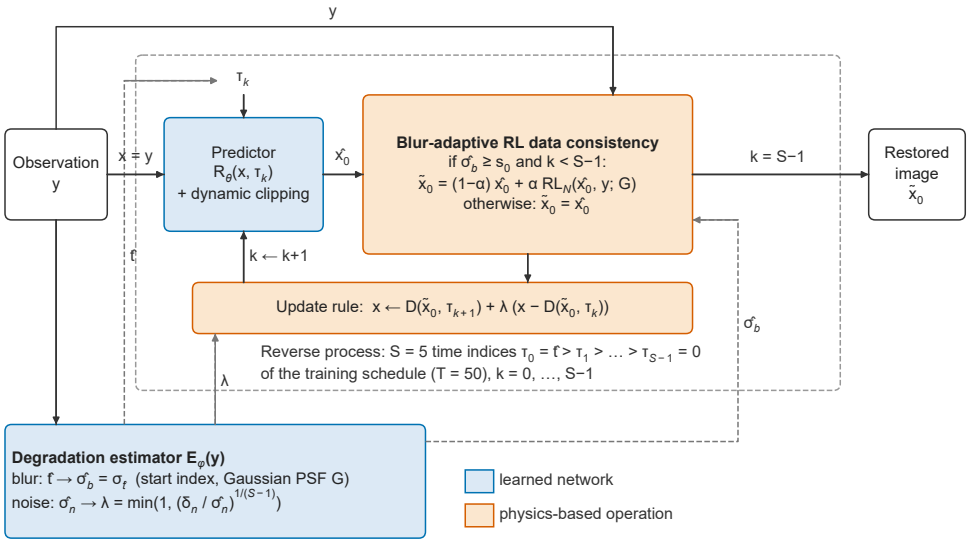



**Fig. 1** Overview of the proposed method. The degradation estimator predicts the blur strength $\hat{\sigma}_b$ and the noise level $\hat{\sigma}_n$ of the observation $y$. The reverse process starts from $x_{\tau_0} = y$ and visits five time indices of the training schedule. At each step, the predictor $R_\theta$ predicts the sharp image, RL data consistency with the Gaussian PSF $G = G_{\hat{\sigma}_b}$ is applied when $\hat{\sigma}_b \geq s_0$, and the update rule with the coefficient $\lambda(\hat{\sigma}_n)$ gives the next state

the next state is computed with the generalised update rule. Algorithm 1 summarises the procedure.

## 3.1 Blurring diffusion process and training

Following Cold Diffusion [16], we define the forward process by a deterministic degradation instead of additive Gaussian noise. The degradation operator blurs a sharp image $x_0$ with a Gaussian kernel whose width increases with the time index $t$:

$$D(x_0, t) = G_{\sigma_t} * x_0, \qquad \sigma_t = \sigma_{\min} + (\sigma_{\max} - \sigma_{\min})\frac{t}{T-1}, \qquad t = 0, 1, \ldots, T-1, \quad (1)$$

where $*$ denotes convolution and $G_\sigma$ is a two-dimensional Gaussian kernel with standard deviation $\sigma$ (in pixels). We use $T = 50$, $\sigma_{\min} = 0.02$ and $\sigma_{\max} = 2.0$, a kernel of $101 \times 101$ pixels and reflective padding at the image borders. The operator $D(\cdot, t)$ is linear in its first argument, which we use in Section 3.3. Every intermediate state of this process is a blurred image, and $D(\cdot, t)$ is a physical model of Gaussian blur with strength $\sigma_t$.

A predictor network $R_\theta(x_t, t)$ is trained to recover the sharp image from a degraded image at time index $t$. Real observations contain sensor noise, so the training input is

$$x_t = D(x_0, t) + \sigma_n \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \qquad \sigma_n \sim \mathcal{U}(0, 0.05), \quad (2)$$

where images are normalised to $[-1, 1]$, and $t$ is drawn uniformly from $\{0, \dots, T-1\}$. The noise level $\sigma_n$ is drawn independently of $t$. If the noise level increased with $t$, a network could infer the blur strength from the noise level instead of from the blur itself, and it would fail on real images, in which the two are unrelated. The network is trained with the L1 loss $\|R_\theta(x_t, t) - x_0\|_1$.

At inference, the prediction $\hat{x}_0 = R_\theta(x_t, t)$ may contain pixel values outside the training range. We apply dynamic clipping: with $s$ the 99th percentile of $|\hat{x}_0|$, bounded below by 1, the prediction is clipped to $[-s, s]$ and divided by $s$.

### 3.2 Time-consistent few-step sampling

The reverse process starts from the observation, $x_{\tau_0} = y$, at the time index $\tau_0 = \hat{t}$ that corresponds to the blur strength of $y$ (Section 3.4), and proceeds to $t = 0$. Visiting all time indices requires up to $T = 50$ network evaluations. Instead, we visit $S$ indices spaced evenly between $\hat{t}$ and 0:

$$\tau_k = \text{round}\left(\hat{t}\left(1 - \frac{k}{S-1}\right)\right), \qquad k = 0, 1, \dots, S-1, \tag{3}$$

with $S = 5$; for $\hat{t} = 49$, the indices are $(49, 37, 24, 12, 0)$. Indices that coincide for small $\hat{t}$ are visited once. At every step, the network receives the same time index $\tau_k$ as during training for the same blur strength $\sigma_{\tau_k}$, and the update rule moves the state directly from $\tau_k$ to $\tau_{k+1}$. Skipping time indices in this way reduces the computation by a factor of $S/T$ without changing the meaning of the time index, in the same way as DDIM [34] skips steps of a trained DDPM. Training a separate model with $T = 5$, or feeding the indices $0, \dots, 4$ of a five-step schedule to a network trained with $T = 50$, would instead change the blur strength that each index represents; we compare these choices in Section 5.2.

### 3.3 Generalised update rule with noise-adaptive carry-over

Bansal et al. [16] proposed two update rules for the reverse process. The naive rule re-degrades the prediction, $x_{t'} = D(\hat{x}_0, t')$, where $t'$ is the next time index. The improved rule, which is more stable for blur [16], adds the difference between the current state and the re-degraded prediction: $x_{t'} = x_t - D(\hat{x}_0, t) + D(\hat{x}_0, t')$. We generalise the two rules with a coefficient $\lambda \in [0, 1]$:

$$x_{\tau_{k+1}} = D(\tilde{x}_0, \tau_{k+1}) + \lambda\left(x_{\tau_k} - D(\tilde{x}_0, \tau_k)\right), \tag{4}$$

where $\tilde{x}_0$ is the prediction after data consistency (Section 3.5). The rule reduces to the naive rule for $\lambda = 0$ and to the improved rule for $\lambda = 1$.

To see the role of $\lambda$, write the state and the prediction as $x_t = D(x_0, t) + r_t$ and $\tilde{x}_0 = x_0 + e_t$, where $r_t$ is the residual of the state and $e_t$ is the prediction error. Substituting into Eq. (4) and using the linearity of $D$ gives

$$r_{\tau_{k+1}} = \lambda\, r_{\tau_k} + \left[D(e_{\tau_k}, \tau_{k+1}) - \lambda\, D(e_{\tau_k}, \tau_k)\right]. \tag{5}$$

The first term carries the residual over to the next step, and the second term injects the current prediction error. The initial residual is the observation noise, $r_{\tau_0} = n$ for $y = D(x_0, \hat{t}) + n$. With $\lambda = 1$, the injected error is only the difference between two blur levels of $e_t$, but the observation noise is carried through all steps and remains in the output. With $\lambda = 0$, the noise is discarded after the first step, but the full re-degraded error $D(e_t, \tau_{k+1})$ is injected at every step. For $0 < \lambda < 1$, the noise decays by a factor of $\lambda$ per step, so that $\lambda^{S-1} n$ remains after $S$ steps. With few steps, a moderate $\lambda$ already suppresses the noise; for example, $0.5^4 \approx 6\%$ remains with $S = 5$ and $\lambda = 0.5$.

Because a fixed $\lambda < 1$ weakens the error correction on noise-free images, we set $\lambda$ for each image from the estimated noise standard deviation $\hat{\sigma}_n$ as the largest value that reduces the remaining noise $\lambda^{S-1}\hat{\sigma}_n$ to a tolerance $\delta_n$:

$$\lambda = \min\left(1, \left(\frac{\delta_n}{\hat{\sigma}_n}\right)^{1/(S-1)}\right), \tag{6}$$

with $\delta_n = 0.004$ on the $[-1, 1]$ scale. Like the other inference hyperparameters, $\delta_n$ was selected on the validation set as the value with the lowest LPIPS among those whose PSNR was within 0.1 dB of the best candidate (Section 4.4). For $\hat{\sigma}_n \le \delta_n$, $\lambda = 1$ and the update rule is that of Bansal et al.

### 3.4 Degradation estimator

The blur strength of a real image is unknown, whereas the reverse process must start from the time index that matches it, and the PSF used for data consistency must match it as well. Because the time index and the blur strength correspond one-to-one through Eq. (1), estimating the time index $\hat{t}$ of the observation determines both. We train a small network that predicts $\hat{t}$ and the noise level $\hat{\sigma}_n$ from $y$ in a single evaluation. The network consists of an encoder of NAFBlocks [8] that reduces the image to 1/16 of its size (1, 1, 2 and 4 blocks per resolution, with 32 to 512 channels), global average pooling and two fully connected layers, and has 2.9 M parameters. It outputs $\hat{u}_1$ and $\hat{u}_2$, which approximate $t/(T-1)$ and $\sigma_n/0.05$, and is trained with the loss $\frac{1}{2}\big(|\hat{u}_1 - t/(T-1)| + |\hat{u}_2 - \sigma_n/0.05|\big)$, averaged over the mini-batch, on degraded images generated as in Eq. (2) and quantised to 8 bits, as the test images are. At inference, $\hat{t} = \mathrm{round}((T-1)\,\mathrm{clip}(\hat{u}_1, 0, 1))$ and $\hat{\sigma}_n = 0.05\max(\hat{u}_2, 0)$. We write $\hat{\sigma}_b = \sigma_{\hat{t}}$ for the estimated blur strength. The estimator replaces the PSF measurement required by RL deconvolution and by data consistency methods with a known operator [15].

### 3.5 Blur-adaptive Richardson–Lucy data consistency

Because every state of the reverse process is a blurred image and $D$ is a physical blur model, the prediction can be compared with the observation directly. We correct the prediction with RL deconvolution [4, 5], using the Gaussian kernel of the estimated blur strength, $h = G_{\hat{\sigma}_b}$, as the PSF. The prediction and the observation are mapped to positive values, $u^{(0)} = \max((\hat{x}_0 + 1)/2, \varepsilon)$ and $v = \max((y+1)/2, \varepsilon)$ with $\varepsilon = 10^{-6}$,

and $N$ RL iterations are applied starting from the prediction:

$$u^{(j+1)} = u^{(j)} \cdot \left[ h^{\top} * \frac{v}{\max(h * u^{(j)}, \varepsilon)} \right], \qquad j = 0, 1, \ldots, N-1, \tag{7}$$

where the multiplication, division and maximum are pixel-wise, the lower bound $\varepsilon$ prevents division by zero, and $h^{\top} = h$ for a symmetric kernel. The corrected prediction is a blend of the network prediction and the RL result:

$$\tilde{x}_0 = (1 - \alpha_k)\, \hat{x}_0 + \alpha_k \left( 2u^{(N)} - 1 \right), \tag{8}$$

with $N = 10$ and a mixing ratio $\alpha_k$ for step $k$.

Data consistency does not always help. Our predictors take the observation as input, and a single prediction already agrees with the observation within the noise level: the median of the residual $\|D(\hat{x}_0, \hat{t}) - y\|$ is 0.97 times the noise standard deviation on the validation set, compared with 1.02 for the ground truth. Further RL iterations therefore tend to fit the observation noise. For weak blur, the observation retains high frequencies in which noise dominates, and RL deconvolution amplifies them, whereas for strong blur it strengthens the contrast of structures that the network output has smoothed (Section 5.2). In our experiments, applying RL correction with $\alpha = 0.5$ at every step lowered the PSNR by 2.6–2.9 dB for weakly blurred images ($\sigma = 0.75$) and also degraded noisy images, while it improved the LPIPS of strongly blurred images (Section 5.2). We therefore apply a weak correction only to strongly blurred images and skip it at the last step:

$$\alpha_k = \begin{cases} \alpha & \text{if } \hat{\sigma}_b \geq s_0 \text{ and } k < S - 1, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

We select $\alpha$ and the threshold $s_0$ on the validation set as the pair with the lowest LPIPS among those that lower the PSNR by at most 0.1 dB from the result without data consistency, which gives $\alpha = 0.15$ and $s_0 = 1.5$ for the improved TransUNet and $\alpha = 0.1$ and $s_0 = 1.0$ for the time-conditioned NAFNet. The switch requires no additional computation, because $\hat{\sigma}_b$ is already estimated for the starting time index.

### 3.6 Predictor networks

The framework does not depend on a particular network architecture, and we use two time-conditioned predictors. Both condition on the time index through feature-wise linear modulation (FiLM) [50]: a sinusoidal embedding $\psi(t)$ of the time index is mapped to a scale $\gamma(t)$ and a shift $\beta(t)$ for each block,

$$z' = \big(1 + \gamma(t)\big) \odot \mathrm{LN}(z) + \beta(t), \qquad [\gamma(t), \beta(t)] = W\, \mathrm{SiLU}\big(\psi(t)\big), \tag{10}$$

where $z$ is a feature map, LN is layer normalisation and $\odot$ is the channel-wise product. The matrix $W$ is initialised to zero, so that each network starts training as an unconditioned network.

**Algorithm 1** Inference of the proposed method

**Require:** observation $y$, predictor $R_\theta$, estimator $E_\phi$, $S = 5$, $\delta_n$, $\alpha$, $s_0$, $N$

$(\hat{t}, \hat{\sigma}_n) \leftarrow E_\phi(y)$; $\quad \hat{\sigma}_b \leftarrow \sigma_{\hat{t}}$

$\lambda \leftarrow \min\big(1, (\delta_n/\hat{\sigma}_n)^{1/(S-1)}\big)$ ▷ Eq. (6)

choose $\tau_0 = \hat{t} > \tau_1 > \cdots > \tau_{S-1} = 0$ ▷ Eq. (3)

$x \leftarrow y$

**for** $k = 0, 1, \ldots, S-1$ **do**

  $\hat{x}_0 \leftarrow \mathrm{clip}\big(R_\theta(x, \tau_k)\big)$ ▷ dynamic clipping

  **if** $\hat{\sigma}_b \geq s_0$ **and** $k < S-1$ **then**

    $\tilde{x}_0 \leftarrow (1-\alpha)\,\hat{x}_0 + \alpha\,\mathrm{RL}_N(\hat{x}_0, y; G_{\hat{\sigma}_b})$ ▷ Eqs. (7), (8)

  **else**

    $\tilde{x}_0 \leftarrow \hat{x}_0$

  **end if**

  **if** $k = S-1$ **then return** $\tilde{x}_0$

  **end if**

  $x \leftarrow D(\tilde{x}_0, \tau_{k+1}) + \lambda\,\big(x - D(\tilde{x}_0, \tau_k)\big)$ ▷ Eq. (4)

**end for**

The *improved TransUNet* is a U-Net with residual convolution blocks and multi-head self-attention, in the spirit of TransUNet [51]. Compared with the network of our previous work, self-attention is used only at the lowest resolution ($25 \times 25$), the output is the sum of the input and a predicted residual, layer normalisation replaces group normalisation, and the time index is injected by FiLM. These changes improved the PSNR of a single prediction by 1.7 dB in a preliminary comparison and made the network five times faster; the network has 10.3 M parameters. The *time-conditioned NAFNet* adds FiLM (Eq. (10)) to every NAFBlock of NAFNet [8] in the configuration for GoPro deblurring (width 32; 1, 1, 1 and 28 encoder blocks, 1 middle block and 1 decoder block per resolution). The time embedding is mapped to 256 dimensions by two fully connected layers before it is passed to the blocks. The network has 25.8 M parameters and the same architecture as the direct-regression NAFNet baseline apart from the time conditioning, which allows the effect of the sampling procedure to be separated from that of the network (Section 5.5).

# 4 Experimental Setup

## 4.1 Datasets and degradations

### 4.1.1 Training data

We used 4,683 gastrointestinal endoscopic images: 3,683 labelled images of the HyperKvasir dataset [52] (normal caecum, pylorus and Z-line, oesophagitis and Barrett's oesophagus) and the 1,000 polyp images of the Kvasir-SEG dataset [53]. Because the endoscopic field of view is surrounded by a black background with rounded corners, each image was cropped to a rectangle containing only the gastrointestinal tract, so that the networks are trained mainly on mucosal content; black corners remain in some images (more than 10% of the pixels in about 10% of the images). The images

**Table 1** Test sets generated from EDD2020. All sets use isotropic Gaussian blur unless stated otherwise. The noise levels 0.07 and 0.1 exceed the range used in training ($\sigma_n \leq 0.05$)

| Set | Images | Blur | Noise |
|---|---|---|---|
| Main | 386 | $\sigma = 2.0$ | 0.02 |
| Noise | 97 | $\sigma = 2.0$ | 0 to 0.05 in steps of 0.01, 0.07, 0.1 |
| Blur strength | 40 | $\sigma = 0.75, 1.0, 1.25, 1.5, 1.75, 2.0$ | 0.02 |
| Unseen blur | 40 | anisotropic Gaussian $(\sigma_1, \sigma_2) = (2.0, 0.75), (1.5, 1.0)$; linear motion of 5, 9 and 13 pixels (isotropic $\sigma = 1.5$ as a reference) | 0.02 |

were converted to greyscale, resized to $100 \times 100$ pixels with bilinear interpolation and antialiasing, and normalised to $[-1, 1]$. Only the sharp images are stored; degraded inputs are generated during training as described in Section 3.1. Of these, 10% (468 images) were set aside as a validation split for hyperparameter selection, with near-duplicate images, detected with a perceptual hash, kept on the same side of the split, and all networks were trained on the remaining 4,215 images. All reported results are on a separate test dataset that was used neither for training nor for selection.

### 4.1.2 Test data

We used the 386 images of the EDD2020 dataset [18], which were acquired at different institutions and with different endoscopes from the training data. We confirmed with a perceptual hash and by visual inspection that no test image is identical or nearly identical to a training image. The test images were used as provided, without cropping, so that some of them include the black border of the field of view (more than 10% of the pixels in 89 of the 386 images); they were otherwise preprocessed in the same way as the training images, and degraded images were generated from them under the conditions in Table 1. Noise levels are standard deviations on the $[-1, 1]$ scale (0.02 corresponds to about 2.6 grey levels on the 0–255 scale), the noise was seeded by the file name so that every set can be regenerated, and all degraded images were quantised to 8 bits.

The main set contains all 386 images. The noise set consists of every fourth image (97 images), and the blur-strength and unseen-blur sets of the first 40 of these. All Gaussian-blurred sets were generated with the degradation operator $D$ of Eq. (1), including its reflective padding at the image borders. For the unseen blur, the orientation of the anisotropic Gaussian and the motion kernels was drawn at random for each image, and the kernels ($31 \times 31$ pixels) were applied with reflective padding. Neither anisotropic nor motion blur occurs in training, and these sets test how the method behaves when its assumption of isotropic Gaussian blur is violated.

## 4.2 Baselines

We compared the proposed method with the following methods.

- *RL deconvolution* with the true PSF and 20 iterations, the number of iterations being selected on the validation split. It shows what classical deconvolution achieves when the PSF is known exactly, which our method does not assume.
- *Direct regression* with NAFNet [8] (the GoPro configuration with width 32, 17.1 M parameters) and Restormer [9] (width 48 with 4, 6, 6 and 8 blocks per level, 26.1 M parameters), which map the observation to the sharp image in a single evaluation. They were trained on the same images and degradations as our predictors, including noise drawn independently of the blur strength.
- A *conditional DDPM*, which generates the sharp image from Gaussian noise conditioned on the observation [11, 12]. To isolate the effect of the diffusion process, it uses the time-conditioned NAFNet of Section 3.6 with the noisy image and the observation as a two-channel input, and it was trained on the same images and degradations as our predictors, with 1000 time steps, a linear noise schedule from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$ and the noise-prediction objective. Samples were drawn with deterministic DDIM [34] updates from random initial noise, with the predicted sharp image clipped to $[-1, 1]$ at each step. We report 5 and 50 sampling steps: more steps improve the perceptual quality and lower the PSNR, so no single number of steps is best on both criteria. The output depends on the random seed; unless stated otherwise, seed 0 was used.

The contribution of each component of the proposed method is examined with the ablated variants listed in Section 5.2, which use the same trained networks.

## 4.3 Evaluation metrics

### 4.3.1 Image quality

We report the PSNR, the structural similarity index (SSIM) and the learned perceptual image patch similarity (LPIPS) [19]. The PSNR and SSIM were computed on 8-bit images with a data range of 255, and the LPIPS with the AlexNet backbone, with the grey-level image replicated to three channels. All metrics were computed on the central $84 \times 84$ pixels, excluding an 8-pixel border. Real blur at the image borders depends on the scene outside the field of view, and none of the padding schemes used to generate synthetic blur reproduces it. Because the proposed method applies the blur operator $D$ during sampling, its PSNR over the whole image depends on whether the padding of the test data matches that of $D$: when the 97 images of the noise set ($\sigma_n = 0.02$) were regenerated with replicate instead of reflective padding, the full-image PSNR of the proposed method decreased by 0.8–1.5 dB and that of direct regression by 0.3 dB, whereas the PSNR in the central region changed by only 0.2–0.4 dB for all methods. Conversely, on our test sets, which are generated with the same padding as $D$, the proposed method would gain an unfair advantage at the borders. Excluding the border removes this dependence. For a whole-image comparison that does not favour any method, Online Resource 2 gives the metrics over the whole image on test sets generated without padding, by blurring a larger version of each image and cropping its centre. We report the mean over the test images. To test whether the differences between methods were consistent over the images, we applied the paired

two-sided Wilcoxon signed-rank test to the per-image metrics and corrected the $p$-values of all 66 comparisons with the Holm method; these tests cover the variation over test images, not over training runs.

### 4.3.2 Hallucination

Following Bhadra et al. [17], we separate the error into the components that are determined by the observation and those that are not. For a linear forward operator, the generalised measurement space is spanned by the singular vectors whose singular values exceed a tolerance, and its orthogonal complement is the generalised null space; components in the null space cannot be recovered stably from the observation and are determined by the prior learnt by the network. For Gaussian blur, the singular vectors are Fourier components and the singular values are the values of the transfer function $|H(f)| = \exp(-2\pi^2\sigma^2|f|^2)$. We set the tolerance at the frequency $f_c$ at which the signal-to-noise ratio of the observation is one, $|H(f_c)|^2 S(f_c) = N$, where $S$ is the radially averaged power spectrum of the ground-truth images and $N$ the power of the noise, including quantisation noise; frequencies above $f_c$ form the null space. To reduce boundary effects, images were mirrored to $200 \times 200$ pixels before the Fourier transform, and the 8-pixel border was excluded from the metrics. We report the root mean square (RMS, on the 0–255 scale) of the null-space hallucination map $\hat{x}_\mathrm{N} - x_\mathrm{N}$ and of the measurement-space hallucination map $\hat{x}_\mathrm{M} - H_P^{+} y$, where the subscripts denote the null-space and measurement-space components and $H_P^{+}$ is the truncated pseudoinverse [17]. Because the null-space map also counts missing detail as an error, we further decompose the null-space component of the restoration as $\hat{x}_\mathrm{N} = a\, x_\mathrm{N} + r$ with $r$ orthogonal to $x_\mathrm{N}$, and report the *added component* $\|r\|^2/\|x_\mathrm{N}\|^2$, the energy in the null space that is unrelated to the ground truth, and the *missing component* $(1-a)^2$. This decomposition is our addition to the definitions of Bhadra et al. For the conditional DDPM, we also report the RMS of the difference between the null-space components of two samples with different seeds divided by $\sqrt{2}$, which is the part of the null-space content that is determined by the random noise.

### 4.3.3 Run time

The run time per image was measured on 50 images of the main set that followed five warm-up images, processing one image at a time with GPU synchronisation. For the proposed method, it includes the degradation estimator.

## 4.4 Implementation details

All methods were implemented in PyTorch 2.7.0 and run on an NVIDIA GeForce RTX 4060 Ti GPU with 16 GB of memory. The predictors, the degradation estimator, the direct-regression NAFNet and the conditional DDPM were trained for 51 epochs with AdamW, a batch size of 8 and a learning rate of $10^{-3}$ that decayed to $10^{-7}$ with a cosine schedule ($\beta_2 = 0.9$, no weight decay). Restormer was trained for the same number of epochs with a batch size of 2 and a learning rate of $3 \times 10^{-4}$ that decayed to $10^{-6}$ ($\beta_2 = 0.999$, weight decay $10^{-4}$). For the conditional DDPM, an exponential moving average of the weights (decay 0.999) was used for sampling. All runs used a

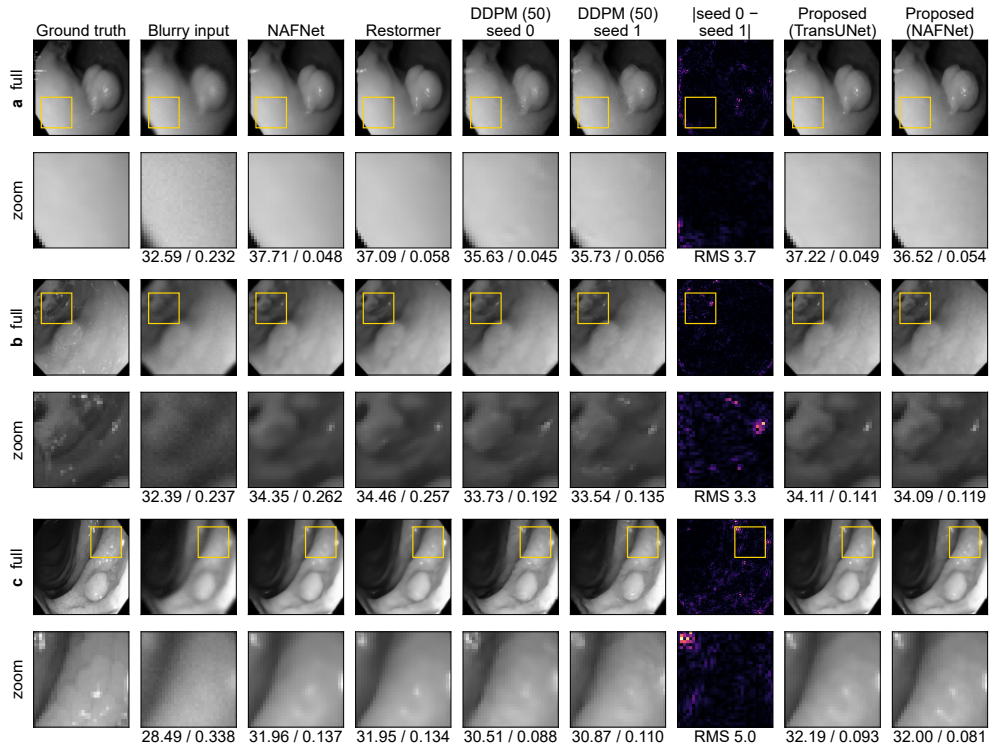



**Fig. 2** Restoration examples from the EDD2020 test set ($\sigma = 2.0$, noise 0.02). The rows (a)–(c) are the images at the 75th, 50th and 25th percentiles of the PSNR of the proposed method (NAFNet predictor) over the 386 test images; they were not selected by visual inspection. For each image, the upper row shows the whole image and the lower row the $32 \times 32$ region with the highest high-frequency energy of the ground truth (yellow box), selected automatically and shown with nearest-neighbour magnification. The conditional DDPM is shown with two random seeds, and the seventh column shows the absolute difference between them (0–64 on the 0–255 scale); this difference is zero for the deterministic methods. Numbers are PSNR (dB) / LPIPS of each image. More examples are given in Online Resource 1

fixed random seed and deterministic GPU algorithms. The number of steps was fixed to $S = 5$ as a compromise between the number of network evaluations, the perceptual quality and the fidelity: on the validation split, increasing $S$ from 3 to 10 lowered both the LPIPS and the PSNR (Online Resource 2, Table S7). The other inference hyperparameters of the proposed method ($\delta_n$, $\alpha$, $s_0$ and $N$; Section 3) were selected on the validation split with a single rule, taking the value with the lowest LPIPS among the candidates whose PSNR was within 0.1 dB of the best one, and fixed for all test sets. The results for all candidates are given in Online Resource 2 (Tables S5–S7).

**Table 2** Comparison on the main test set (386 EDD2020 images, $\sigma = 2.0$, noise 0.02). Metrics are means over the images, computed on the central $84 \times 84$ pixels. The run time is per image on an RTX 4060 Ti and includes the degradation estimator for the proposed method. The best value among the restoration methods is in bold

| Method | PSNR (dB) ↑ | SSIM ↑ | LPIPS ↓ | Time (s) ↓ |
|---|---|---|---|---|
| Blurry input | 31.22 | 0.8181 | 0.2839 | – |
| RL deconvolution (true PSF) | 32.45 | 0.8476 | 0.1799 | 0.056 |
| NAFNet (direct) | 34.31 | 0.8990 | 0.1826 | **0.034** |
| Restormer (direct) | **34.41** | **0.8999** | 0.1724 | 0.053 |
| Cond. DDPM (5 steps) | 33.97 | 0.8928 | 0.1693 | 0.218 |
| Cond. DDPM (50 steps) | 33.09 | 0.8744 | 0.1275 | 2.068 |
| Proposed (TransUNet) | 34.15 | 0.8931 | 0.1028 | 0.347 |
| Proposed (NAFNet) | 34.06 | 0.8918 | **0.1011** | 0.375 |

# 5 Results

## 5.1 Comparison with baselines

To evaluate the restoration performance of the proposed method, we compared it with the baselines on the main test set. Table 2 gives the results. Direct regression gave the highest PSNR and SSIM, and RL deconvolution with the true PSF improved the PSNR of the blurred input by only 1.2 dB. The proposed method reduced the LPIPS from 0.172–0.183 for direct regression to 0.101–0.103, the lowest among all methods. Its PSNR was 0.26–0.35 dB lower than that of Restormer and 0.25 dB lower than that of direct regression with the same NAFNet network. For the conditional DDPM, increasing the number of sampling steps from 5 to 50 reduced the LPIPS from 0.169 to 0.128 but also the PSNR from 33.97 to 33.09 dB. Compared with the conditional DDPM with 50 steps, the proposed method had a 1.0–1.1 dB higher PSNR and a lower LPIPS, and was about six times faster. Direct regression, which evaluates its network once, was faster still. These results indicate that the proposed method gives the lowest LPIPS of all methods without a large loss in PSNR. These differences were consistent over the images. With both predictors, the proposed method had a lower LPIPS than direct regression and the conditional DDPM with 5 steps on 98–99% of the images and than the conditional DDPM with 50 steps on 81–84% ($p < 10^{-33}$ for all comparisons), and its lower PSNR compared with direct regression was also significant ($p < 10^{-14}$). The difference in LPIPS between the two predictors (0.101 and 0.103) was not significant at the 5% level (corrected $p = 0.050$).

Next, to examine the visual differences between the methods, Figure 2 shows restored images at three percentiles of the PSNR. In the whole images, all learning-based methods removed most of the blur, and the differences between them appear only in fine structures such as specular highlights and mucosal texture in the magnified regions. The two samples of the conditional DDPM differ from each other by an RMS of 3.3–5.0 grey levels in these images, whereas the other methods are deterministic and give a single output for each input. These results indicate that the methods

**Table 3** Ablation of the proposed method with both predictors (PSNR in dB and LPIPS). "Mismatched time indices" gives the network the indices 0–4 of a five-step schedule instead of the corresponding indices of the training schedule. "RL-DC at all steps" applies the correction with $\alpha = 0.5$ at every step and to every image, as in our conference paper. "True blur and noise" replaces $\hat{t}$ and $\hat{\sigma}_n$ by the true values and is not compared. "Added" is the added component in the null space on the main set (Section 4.3). The best value for each predictor is in bold

| | Main | | $\sigma = 0.75$ | | Noise 0.05 | | Main |
|---|---|---|---|---|---|---|---|
| Setting | PSNR | LPIPS | PSNR | LPIPS | PSNR | LPIPS | Added |
| *TransUNet predictor* | | | | | | | |
| Single prediction | **34.55** | 0.1927 | **39.97** | 0.0528 | **33.20** | 0.2471 | 0.184 |
| 5 steps, mismatched time indices, $\lambda = 1$ | 33.29 | 0.2261 | 39.67 | 0.0488 | 32.31 | 0.2053 | – |
| 5 steps, $\lambda = 1$ | 34.19 | 0.1110 | 39.57 | 0.0270 | 32.77 | 0.1536 | 0.338 |
| + adaptive $\lambda$ | 34.23 | 0.1137 | 39.45 | **0.0237** | 32.96 | 0.1559 | 0.312 |
| + RL-DC at all steps ($\alpha = 0.5$, $\lambda = 1$) | 33.69 | **0.0898** | 36.65 | 0.1029 | 30.68 | 0.1758 | – |
| + RL-DC on all images | 34.15 | 0.1028 | 38.71 | 0.0330 | 32.52 | **0.1273** | – |
| + blur-adaptive RL-DC (proposed) | 34.15 | 0.1028 | 39.45 | **0.0237** | 32.54 | 0.1312 | 0.339 |
| Proposed with true blur and noise | 34.14 | 0.1003 | 39.78 | 0.0197 | 32.56 | 0.1275 | – |
| *NAFNet predictor* | | | | | | | |
| Single prediction | **34.33** | 0.1819 | **39.90** | 0.0546 | **33.07** | 0.2292 | 0.204 |
| 5 steps, mismatched time indices, $\lambda = 1$ | 33.36 | 0.2274 | 39.72 | 0.0548 | 32.57 | 0.2150 | – |
| 5 steps, $\lambda = 1$ | 34.10 | 0.1085 | 39.64 | 0.0309 | 32.86 | 0.1564 | 0.327 |
| + adaptive $\lambda$ | 34.10 | 0.1075 | 39.43 | **0.0241** | 32.89 | 0.1446 | 0.320 |
| + RL-DC at all steps ($\alpha = 0.5$, $\lambda = 1$) | 33.68 | **0.0902** | 37.09 | 0.0825 | 31.19 | 0.1567 | – |
| + RL-DC on all images | 34.06 | 0.1011 | 38.98 | 0.0257 | 32.66 | **0.1230** | – |
| + blur-adaptive RL-DC (proposed) | 34.06 | 0.1011 | 39.43 | **0.0241** | 32.66 | **0.1230** | 0.336 |
| Proposed with true blur and noise | 34.00 | 0.0969 | 39.69 | 0.0206 | 32.59 | 0.1207 | – |

differ mainly in the reproduction of fine structures and in the reproducibility of their outputs.

## 5.2 Ablation study

First, to examine the contribution of each component of the proposed method, we performed an ablation study on the main set, the $\sigma = 0.75$ set and the noise set. Table 3 gives the results; the results of the two predictors follow the same pattern. A single prediction gave the highest PSNR on the main set but an LPIPS of 0.18–0.19. Sampling with five steps at the corresponding time indices of the training schedule reduced the LPIPS to 0.109–0.111 at the cost of 0.23–0.36 dB in PSNR. When the network was given the indices 0–4 instead, both the PSNR and the LPIPS were worse than with a single prediction. These results indicate that the time index given to the network must correspond to the blur strength, not to the step number.

Next, we examined the role of the adaptive coefficient $\lambda$. It changed the results little at the noise levels of the main set and of the $\sigma = 0.75$ set. On the noise set with $\sigma_n = 0.1$, which is twice the maximum noise level in training, sampling with $\lambda = 1$ lowered the PSNR of the TransUNet predictor from 30.70 dB for a single prediction to 28.01 dB, because the observation noise was carried through all steps; the adaptive $\lambda$ restored it to 30.40 dB. These results indicate that the adaptive $\lambda$ is needed when the noise level is high.

Next, we examined the effect of the range over which RL-DC is applied. RL correction with $\alpha = 0.5$ at every step and for every image, as in our conference paper, gave the lowest LPIPS on the main set, but it lowered the PSNR by 0.4–0.5 dB on the

main set, by 2.6–2.9 dB on weakly blurred images ($\sigma = 0.75$), where it also increased the LPIPS from about 0.03 to 0.08–0.10, and by 1.7–2.1 dB on noisy images. With the weak correction selected on the validation set, applying it to all images still lowered the PSNR for $\sigma = 0.75$ by 0.5–0.7 dB, whereas the blur-adaptive switch left these images unchanged. On the main set, in which all images are strongly blurred, the proposed RL-DC reduced the LPIPS by 0.006–0.011 with a PSNR change of less than 0.1 dB, and on the noise set with $\sigma_n = 0.05$ it reduced the LPIPS by 0.022–0.025 with a PSNR loss of 0.2–0.4 dB. The LPIPS was reduced on 93–96% of the images of the main set and on 81–89% of those of the noise set ($p < 10^{-9}$). These results indicate that RL data consistency helps only strongly blurred images and has to be applied selectively according to the blur strength.

Next, to examine the effect of the degradation estimator, we replaced the estimated blur strength and noise level by the true values. This changed the PSNR by less than 0.1 dB and the LPIPS by less than 0.005 on the main set. The mean absolute error of the estimated blur strength was 0.07–0.10 for blur strengths of 0.75–2.0 with noise 0.02, about two to three intervals of the training schedule (0.040), and 0.19 on the noise set with $\sigma_n = 0.05$. For $\sigma = 2.0$, the largest strength of the schedule, the error can only be negative, and the mean estimates were 1.91 (noise 0.02) and 1.81 (noise 0.05). The mean absolute error of the estimated noise level was below 0.001. These results indicate that the degradation estimator is not a bottleneck under the tested conditions.

Finally, to examine what RL-DC adds to the images, Figure 3 shows two images of the noise set with $\sigma_n = 0.05$ on which RL-DC reduced the LPIPS the most. On these noisy observations, RL-DC restored the contrast of folds and grooves and the texture of the mucosa that the predictor had smoothed. On these images, the LPIPS decreased from 0.20–0.24 to 0.11–0.16, and the PSNR decreased by 0.2–0.6 dB. To examine whether the added component agrees with the ground truth, we computed the correlation between the change made by RL-DC and the remaining error of the output without RL-DC (the ground truth minus that output) in three bands, defined by differences of Gaussian filters with standard deviations of 3 and 8, 1 and 3, and 0.5 and 1 pixels, and averaged it over the images. In the coarsest band, the correlation was positive (0.14–0.18 on the noise set and 0.20–0.26 on the main set for the two predictors), and in the two finer bands, it ranged from $-0.14$ to $-0.01$. These results indicate that RL-DC strengthens coarse structures towards the ground truth and adds texture that resembles that of the ground truth but is not aligned with it pixel by pixel. This explains why RL-DC improves the LPIPS while slightly lowering the PSNR.

## 5.3 Robustness to observation noise

To examine the robustness to observation noise, we compared the methods on the noise set at different noise levels. Figure 4 shows the results against the noise level. Within the training range ($\sigma_n \leq 0.05$), the proposed method gave the lowest LPIPS at every noise level (0.068–0.131 with the two predictors), against 0.113–0.149 for the conditional DDPM with 50 steps and 0.104–0.231 for direct regression. Without noise, its PSNR (36.22–36.48 dB) was close to that of direct regression (36.23–36.77 dB), and at $\sigma_n = 0.05$ it was 0.5–0.7 dB below that of Restormer. The conditional DDPM

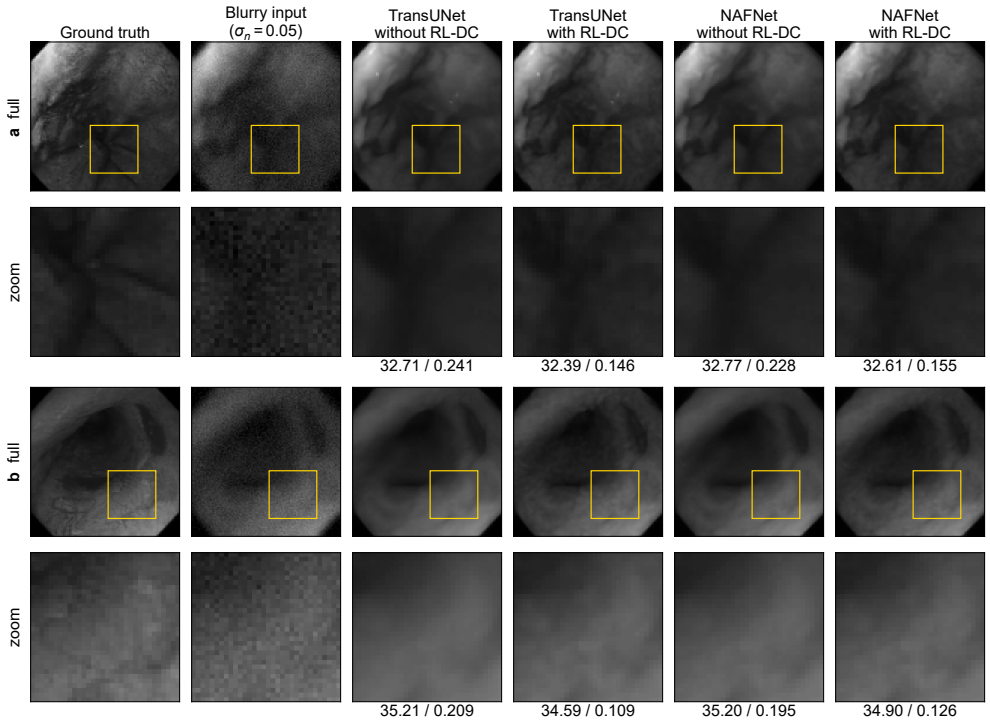



**Fig. 3** Effect of RL-DC on the noise set ($\sigma = 2.0$, $\sigma_n = 0.05$). The rows (a) and (b) are the two images with the largest LPIPS reduction by RL-DC, taking the smaller reduction of the two predictors. The results without RL-DC use the same degradation estimator and adaptive coefficient $\lambda$ as the proposed method and differ only in RL-DC. For each image, the upper row shows the whole image and the lower row the $32 \times 32$ region in the yellow box with nearest-neighbour magnification. Numbers are PSNR (dB) / LPIPS of each image

with 50 steps had the lowest PSNR among the learning-based methods at every noise level. Beyond the training range, the advantage of the proposed method in LPIPS disappeared: at $\sigma_n = 0.1$, its LPIPS (0.210–0.224) was similar to that of the conditional DDPM with 50 steps (0.213) and of NAFNet (0.204), and its PSNR was 0.7–1.2 dB below that of direct regression. At every noise level, no method restored any image of the noise set to a PSNR below that of its input. The outputs of the conditional DDPM with five steps varied with the random seed by a pixel-wise standard deviation of 0.75–1.18 grey levels over three seeds, increasing with the noise level. These results indicate that the proposed method is robust to noise within the training range but does not keep its advantage beyond it.

## 5.4 Blur strength and unseen blur

To examine the dependence on the blur strength, we compared the methods on the blur-strength set at different values of $\sigma$. Figure 5 shows the results against the blur

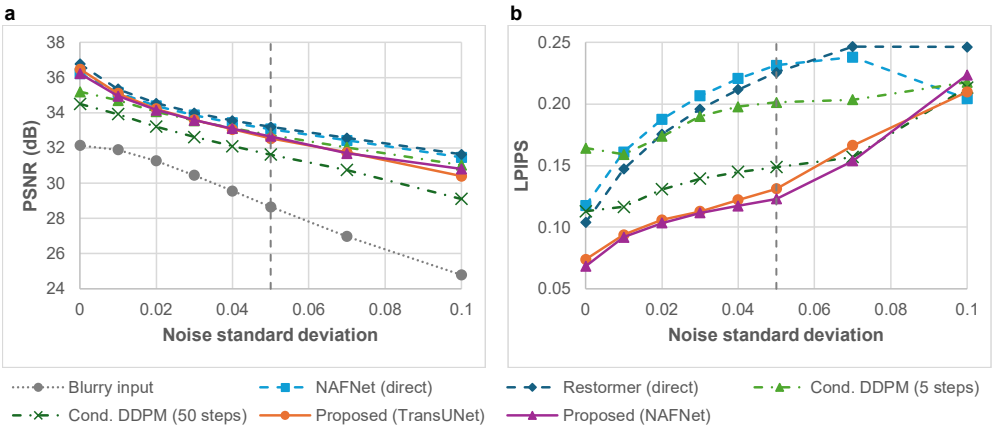



**Fig. 4** (a) PSNR and (b) LPIPS against the noise level on the noise set (97 images, $\sigma = 2.0$). Noise levels above 0.05 exceed the range used in training

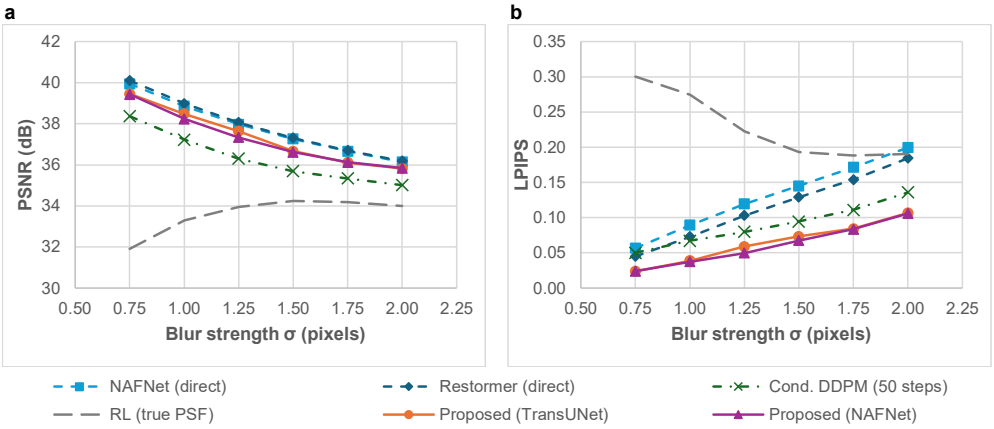



**Fig. 5** (a) PSNR and (b) LPIPS against the blur strength $\sigma$ on the blur-strength set (40 images, noise 0.02)

strength. The proposed method gave the lowest LPIPS at every blur strength (0.024–0.107), and its PSNR was 0.3–0.8 dB below that of direct regression and 0.8–1.3 dB above that of the conditional DDPM with 50 steps. RL deconvolution with the true PSF lowered the PSNR of weakly blurred images below that of the input ($\sigma \leq 1.5$), because it amplified the noise, and improved it only for $\sigma \geq 1.75$. These results indicate that the perceptual advantage of the proposed method does not depend on the blur strength within the training range.

Next, to examine the generalisation to blur that does not occur in training, we compared the methods on anisotropic Gaussian and motion blur. Table 4 gives the results. For the reference within the training distribution (isotropic, $\sigma = 1.5$), the

**Table 4** Results for blur that does not occur in training (unseen-blur set, 40 images, noise 0.02). Isotropic Gaussian blur with $\sigma = 1.5$ is shown as a reference within the training distribution. "Aniso." is an anisotropic Gaussian with standard deviations $(\sigma_1, \sigma_2)$, and "Motion" is linear motion blur of the given length; the orientations were drawn at random for each image. The best value among the restoration methods is in bold

| Method | Iso. 1.5 | Aniso. (2.0, 0.75) | Aniso. (1.5, 1.0) | Motion 5 px | Motion 9 px | Motion 13 px |
|---|---|---|---|---|---|---|
| *PSNR (dB)* ↑ | | | | | | |
| Blurry input | 34.53 | 34.69 | 35.16 | 35.46 | 33.13 | 31.56 |
| NAFNet (direct) | 37.25 | **36.54** | **37.72** | **37.04** | **33.86** | **31.98** |
| Restormer (direct) | **37.27** | 36.25 | 37.69 | 36.78 | 33.59 | 31.82 |
| Cond. DDPM (50 steps) | 35.58 | 35.17 | 35.97 | 35.78 | 33.07 | 31.72 |
| Proposed (TransUNet) | 36.71 | 35.09 | 37.13 | 35.54 | 32.56 | 31.12 |
| Proposed (NAFNet) | 36.60 | 35.25 | 36.87 | 35.59 | 32.78 | 31.39 |
| *LPIPS* ↓ | | | | | | |
| Blurry input | 0.2108 | 0.1949 | 0.1799 | 0.1487 | 0.2317 | 0.3002 |
| NAFNet (direct) | 0.1426 | 0.1563 | 0.1224 | 0.1292 | 0.1983 | 0.2665 |
| Restormer (direct) | 0.1266 | 0.1504 | 0.1074 | 0.1317 | 0.2035 | 0.2669 |
| Cond. DDPM (50 steps) | 0.0958 | 0.1221 | 0.0903 | 0.1077 | 0.1661 | **0.2135** |
| Proposed (TransUNet) | 0.0709 | 0.1279 | 0.0693 | 0.1147 | 0.1832 | 0.2346 |
| Proposed (NAFNet) | **0.0649** | **0.1071** | **0.0584** | **0.1019** | **0.1618** | 0.2158 |

PSNR of the proposed method was 0.6–0.7 dB below that of direct regression. For anisotropic Gaussian and motion blur, this gap grew to up to 1.5 dB, and for motion blur of 13 pixels the PSNR of the proposed method fell below that of the input, whereas direct regression and the conditional DDPM remained slightly above it. The LPIPS of the proposed method with the NAFNet predictor was the lowest for five of the six kernels and close to the lowest for motion blur of 13 pixels. These results indicate that the proposed method remains perceptually competitive under unseen blur but loses more fidelity to the observation than direct regression does.

## 5.5 Hallucination analysis

To examine whether the detail added by the proposed method is based on the ground truth, we compared the added and missing null-space content and the error in the measurement space. Table 5 gives the hallucination metrics. On the main set, direct regression added the least null-space content among the deterministic methods (0.199–0.242) but also missed most of the null-space content of the ground truth (missing component 0.561–0.609), consistent with its smooth outputs. The conditional DDPM with five steps added even less and missed more. With 50 steps, its added component increased to 0.365, and its null-space content depended on the random seed by an RMS of 2.66 grey levels. The proposed method added 0.336–0.339, slightly less than the conditional DDPM with 50 steps, missed the least of the ground-truth content (0.531–0.555), and had the smallest error in the measurement space (2.87–2.92 grey levels, against 3.28–3.75 for the other methods). Its output does not depend on a random seed. The same ordering held on the noise set with $\sigma_n = 0.05$. Figure 6 shows examples of the null-space hallucination maps. These results indicate that the proposed method reproduces the most true detail but adds almost as much content unrelated to the ground truth as the conditional DDPM with 50 steps.

**Table 5** Hallucination metrics after Bhadra et al. [17] (Section 4.3). "Added" and "Missing" are the energies of the null-space component that is unrelated to the ground truth and that is missing, relative to the null-space energy of the ground truth. The hallucination maps (HM) are given as RMS on the 0–255 scale. "Seed-dependent" is the RMS of the null-space content that changes with the random seed. The truncated pseudoinverse, which uses no prior, is shown for reference

| Method | Added ↓ | Missing ↓ | Null-space HM | Measurement HM | Seed-dependent |
|---|---|---|---|---|---|
| *Main set ($\sigma_n = 0.02$)* | | | | | |
| NAFNet (direct) | 0.199 | 0.609 | 4.65 | 3.40 | 0.00 |
| Restormer (direct) | 0.242 | 0.561 | **4.62** | 3.28 | 0.00 |
| Cond. DDPM (5 steps) | **0.154** | 0.685 | 4.73 | 3.52 | 1.06 |
| Cond. DDPM (50 steps) | 0.365 | 0.691 | 5.31 | 3.75 | 2.66 |
| Proposed (TransUNet) | 0.339 | **0.531** | 4.74 | **2.87** | 0.00 |
| Proposed (NAFNet) | 0.336 | 0.555 | 4.83 | 2.92 | 0.00 |
| Truncated pseudoinverse | 0.000 | 1.000 | 0.00 | 0.00 | 0.00 |
| *Noise set, $\sigma_n = 0.05$* | | | | | |
| NAFNet (direct) | 0.177 | 0.602 | 5.26 | 3.74 | 0.00 |
| Restormer (direct) | 0.199 | 0.565 | **5.19** | 3.69 | 0.00 |
| Cond. DDPM (5 steps) | **0.155** | 0.675 | 5.41 | 3.78 | 1.18 |
| Cond. DDPM (50 steps) | 0.419 | 0.676 | 6.17 | 4.06 | 3.21 |
| Proposed (TransUNet) | 0.407 | **0.522** | 5.58 | **2.88** | 0.00 |
| Proposed (NAFNet) | 0.362 | 0.537 | 5.55 | 3.00 | 0.00 |
| Truncated pseudoinverse | 0.000 | 1.000 | 0.00 | 0.00 | 0.00 |

Next, to examine where the added content comes from, Figure 7 relates the added component to the LPIPS. As the number of sampling steps of the conditional DDPM increased from 5 to 50, its LPIPS decreased from 0.169 to 0.128 while the added component increased from 0.154 to 0.365 on the main set. The proposed method reached a lower LPIPS with an added component slightly below that of the conditional DDPM with 50 steps. In the stages of the ablation, a single prediction added 0.184–0.204, and the five-step sampling raised it to 0.327–0.338 while reducing the LPIPS from 0.18–0.19 to 0.11. The adaptive $\lambda$ reduced it slightly, and the RL-DC raised it again on the noise set ($\sigma_n = 0.05$: from 0.276–0.286 to 0.362–0.407), where RL iterations amplify part of the observation noise. These results indicate that most of the perceptual improvement of the proposed method, and most of its added null-space content, come from the iterative sampling itself rather than from the data consistency.

# 6 Discussion and Limitations

## 6.1 Position on the perception–distortion trade-off

As the results in Section 5.1 show, direct regression and the proposed method occupy different points of the perception–distortion trade-off [10]. With the same NAFNet architecture and training data, direct regression gave a 0.25 dB higher PSNR, whereas the proposed sampler reduced the LPIPS by 45%. In contrast, the conditional DDPM moves along this trade-off with the number of sampling steps, but even with 50 steps it did not reach the LPIPS of the proposed method, and it lost about 1 dB of PSNR in doing so. We attribute this difference to the way in which the perceptual quality is obtained. The proposed method obtains it in five deterministic steps from the observation instead of many stochastic steps from noise. Therefore, it is about six times

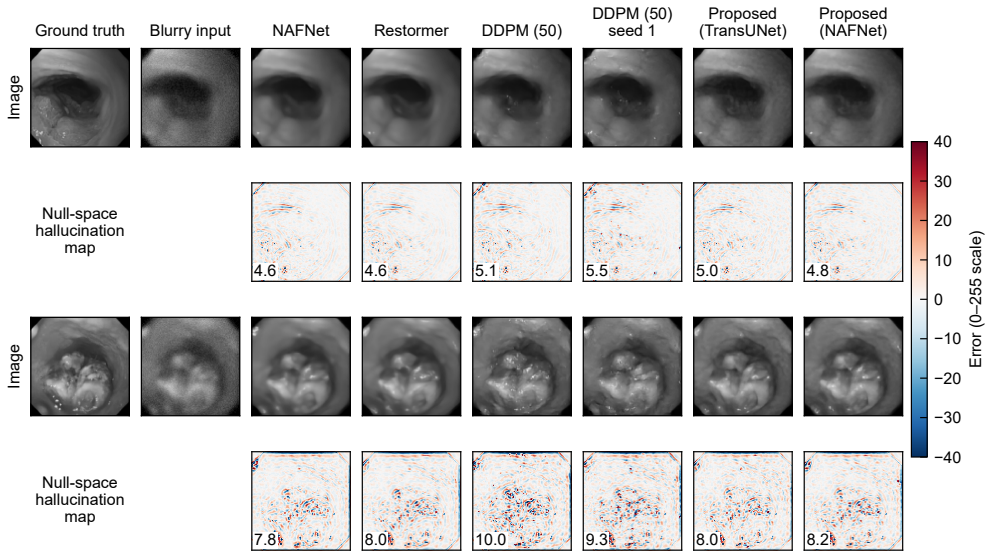



**Fig. 6** Restored images and their null-space hallucination maps $\hat{x}_{\mathrm{N}} - x_{\mathrm{N}}$ for the first two images of the noise set ($\sigma_n = 0.05$), which were not selected by inspection. The conditional DDPM (50 steps) is shown for two random seeds

faster than the conditional DDPM with 50 steps, and its output is reproducible. However, it remains about ten times slower than direct regression with the same network, which may matter for real-time use in video endoscopy.

## 6.2 Role of the RL data consistency

As the results in Section 5.2 show, the effect of the RL-DC depends on the observation noise. On the main set, it reduced the LPIPS by 0.006–0.011 with a PSNR change of less than 0.1 dB, so it does not cost fidelity there. Furthermore, on the noise set with $\sigma_n = 0.05$, it reduced the LPIPS by 0.022–0.025, more than twice as much as on the main set (Table 3). We attribute this difference to the behaviour of the predictor. With noisy observations, the predictor smooths the mucosa together with the noise, and RL deconvolution with the observation restores the contrast of folds and grooves (Fig. 3). Because endoscopic frames are acquired with sensor noise, this is the condition in which the data consistency is most useful. However, its cost is a PSNR loss of 0.2–0.4 dB on the noise set and an increase of the added null-space content (Section 6.3). Therefore, the correction is kept weak and applied only to strongly blurred images.

## 6.3 Reliability of the restored detail

The hallucination analysis in Section 5.5 shows that the perceptual improvement is not free. Compared with a single prediction, the five-step sampling increased the null-space content that is unrelated to the ground truth from about 0.2 to 0.33–0.34 of the null-space energy of the ground truth, a level close to that of the conditional DDPM with

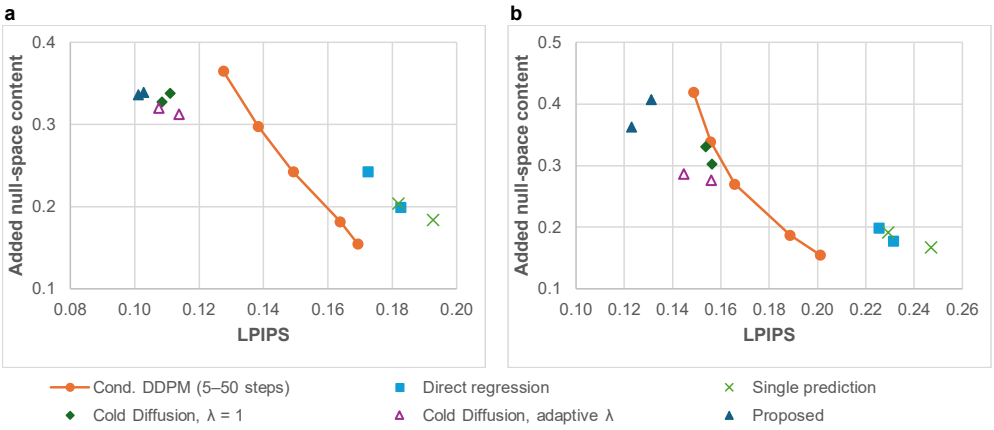



**Fig. 7** Added null-space component against LPIPS on (a) the main set and (b) the noise set with $\sigma_n = 0.05$. The conditional DDPM is shown for 5, 10, 20, 30 and 50 sampling steps (connected by a line), direct regression for NAFNet and Restormer, and the proposed method for each stage of the ablation with the two predictors. Single prediction is one evaluation of the time-conditioned predictor

50 steps. At the same time, on noisy observations within the noise range of training (the main set and the noise set with $\sigma_n = 0.05$; Table 5), the proposed method missed less of the true null-space content than any other method and kept the measurement-space error lowest, and it produced no seed-dependent content. Because endoscopic frames are acquired with sensor noise, this robustness to observation noise is relevant to the intended application. Furthermore, its added content is a deterministic function of the observation, and the same input always gives the same output, which allows the restoration to be checked and reproduced.

However, structures in the null space are not verified by the observation, whichever method produces them, and fine details in the restored images should not be used for diagnosis without such caution. In addition, the RL-DC, which we introduced to make the restoration consistent with the observation, did reduce the LPIPS of strongly blurred images, but it also increased the added component under noise. We attribute this to the fact that a single prediction of our predictors already agrees with the observation within the noise level (Section 3.5). Therefore, data consistency in the measurement space cannot constrain the null-space content further. To resolve this issue, a prior on the null space itself is required, for example a penalty on the added component during sampling, which we leave to future work.

## 6.4 Limitations

This study has several limitations. First, the degradation estimator and the RL-DC assume spatially invariant isotropic Gaussian blur. As Table 4 shows, for anisotropic Gaussian and motion blur, the proposed method lost more PSNR than direct regression, and for strong motion blur its PSNR fell below that of the input. Real endoscopic blur includes motion blur and defocus that varies across the image. To relax this

assumption, the forward process and the estimator need to be extended to parametric families of anisotropic and motion kernels, or the blur needs to be estimated locally.

Second, all degradations were synthetic, because blurred endoscopic frames with sharp ground truth are not available. Therefore, evaluation on real blurred frames, with no-reference metrics and assessment by clinicians, is required before clinical use. Third, the images were greyscale and $100 \times 100$ pixels, much smaller than endoscopic video frames. This size was set by the memory of the single GPU with 16 GB used to train the predictors, the baselines and the conditional DDPM. Furthermore, the self-attention of the improved TransUNet scales poorly with resolution. Colour images at the native resolution remain to be tested.

Fourth, as Fig. 4 shows, the advantage of the proposed method decreased beyond the noise levels used in training. Therefore, the training range should cover the noise of the target endoscope. Fifth, the hallucination metrics rely on the linear model of Gaussian blur and can be computed only for synthetic degradations. Finally, each network was trained once with a fixed seed. The statistical tests account for the variation over test images but not over training runs. Hence, differences of a few hundredths of a decibel or of 0.001 in LPIPS should not be overinterpreted.

# 7 Conclusion

We proposed a few-step Cold Diffusion method for blind deblurring of endoscopic images that combines time-consistent sampling in five steps, an update rule whose carry-over of observation noise adapts to the estimated noise level, a degradation estimator that replaces the measurement of the PSF, and RL data consistency applied only to strongly blurred images. On 386 EDD2020 images with Gaussian blur and noise, the method gave the lowest LPIPS among direct regression, a conditional DDPM and RL deconvolution with the true PSF (0.101 with the NAFNet predictor), at a PSNR 1.0 dB above that of the conditional DDPM with 50 steps, in 0.38 s per image. Furthermore, its outputs are deterministic, and on the main set it missed less of the true detail and had a smaller measurement-space error than the other methods. However, its added null-space content was close to that of the conditional DDPM with 50 steps. In addition, the ablation showed that, for few-step sampling, the time index given to the network must correspond to the blur strength, that the noise-adaptive update rule prevents the failure of Cold Diffusion at high noise levels, and that data consistency should be applied according to the blur strength. These results indicate that the perceptual quality of a generative model and a deterministic, reproducible restoration based on the observation can be achieved together in a few steps. On the other hand, our evaluation was limited to synthetic isotropic Gaussian blur and small greyscale images. Future work will therefore address blur models beyond isotropic Gaussian blur, constraints on the null-space content during sampling, and evaluation on real blurred endoscopic frames.

**Supplementary information. Online Resource 1** (ESM_1.pdf): Additional restoration examples from the EDD2020 test set ($\sigma = 2.0$, noise 0.02), arranged as in Fig. 2, for the images at the 90th, 60th, 40th and 10th percentiles of the PSNR of the proposed method (NAFNet predictor).

**Online Resource 2** (ESM_2.pdf): Image-quality metrics computed over the whole image instead of the central region (Tables S1–S4, corresponding to Table 2, Table 3, Fig. 5 and Table 4, respectively), on test sets generated without padding by blurring a larger version of each image and cropping its centre; and the validation results for all candidate values of the inference hyperparameters $\delta_n$, $\alpha$, $s_0$, $S$ and $N$ (Tables S5–S7).

## Declarations

- Funding: No funding was received for conducting this study.
- Competing interests: The authors declare that they have no competing interests.
- Ethics approval and consent to participate: Not applicable. This study used only publicly available, de-identified endoscopic image datasets (HyperKvasir [52], Kvasir-SEG [53] and EDD2020 [18, 47]).
- Consent for publication: Not applicable.
- Data availability: The HyperKvasir dataset is available at https://doi.org/10.17605/OSF.IO/MH9SJ and the Kvasir-SEG dataset at https://datasets.simula.no/kvasir-seg/; both are released for research and educational purposes. The EDD2020 dataset is available at https://doi.org/10.21227/f8xg-wb80 under the CC BY-NC-SA 4.0 licence. The lists of training, validation and test images and the code to regenerate all degraded test images are provided with the code (see Code availability).
- Materials availability: Not applicable.
- Code availability: The code used in this study is available from the corresponding author on reasonable request.
- Author contribution: K.M.: conceptualisation, methodology, software, investigation, formal analysis, visualisation, writing – original draft. S.U.: supervision, writing – review and editing. All authors read and approved the final manuscript.
- Use of generative AI: The authors used Claude (Anthropic), a large language model, as an assistant to refactor and extend the experimental code, to run and summarise experiments designed by the authors, to check the literature and bibliographic data, and to draft and edit parts of the manuscript text. The authors reviewed and verified all code, results, references and text, and take full responsibility for the content of this article.

## References


[1] Peng L, Liu S, Xie D, Zhu S, Zeng B. Endoscopic video deblurring via synthesis. In: Proceedings of the IEEE Visual Communications and Image Processing (VCIP); 2017. p. 1–4. https://doi.org/10.1109/VCIP.2017.8305021.

[2] Ali S, Zhou F, Bailey A, Braden B, East JE, Lu X, et al. A deep learning framework for quality assessment and restoration in video endoscopy. Medical Image Analysis. 2021;68:101900. https://doi.org/10.1016/j.media.2020.101900.

[3] Oh J, Hwang S, Lee J, Tavanapong W, Wong J, de Groen PC. Informative frame classification for endoscopy video. Medical Image Analysis. 2007;11(2):110–127.

https://doi.org/10.1016/j.media.2006.10.003.

[4] Richardson WH. Bayesian-based iterative method of image restoration. Journal of the Optical Society of America. 1972;62(1):55–59. https://doi.org/10.1364/JOSA.62.000055.

[5] Lucy LB. An iterative technique for the rectification of observed distributions. The Astronomical Journal. 1974;79(6):745–754. https://doi.org/10.1086/111605.

[6] Levin A, Weiss Y, Durand F, Freeman WT. Understanding and evaluating blind deconvolution algorithms. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2009. p. 1964–1971. https://doi.org/10.1109/CVPR.2009.5206815.

[7] Miura K, Usuki S, Sekine T, Miura KT. Image quality enhancement of microscope images assisted by deep learning. In: Proceedings of JSPE Semestrial Meeting, 2025 JSPE Spring Conference; 2025. p. 122–123. In Japanese. https://doi.org/10.11522/pscjspe.2025S.0_122.

[8] Chen L, Chu X, Zhang X, Sun J. Simple baselines for image restoration. In: Proceedings of the European Conference on Computer Vision (ECCV); 2022. p. 17–33. https://doi.org/10.1007/978-3-031-20071-7_2.

[9] Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang MH. Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022. p. 5718–5729. https://doi.org/10.1109/CVPR52688.2022.00564.

[10] Blau Y, Michaeli T. The perception-distortion tradeoff. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018. p. 6228–6237. https://doi.org/10.1109/CVPR.2018.00652.

[11] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33; 2020. p. 6840–6851.

[12] Saharia C, Ho J, Chan W, Salimans T, Fleet DJ, Norouzi M. Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023;45(4):4713–4726. https://doi.org/10.1109/TPAMI.2022.3204461.

[13] Chung H, Kim J, McCann MT, Klasky ML, Ye JC. Diffusion posterior sampling for general noisy inverse problems. In: International Conference on Learning Representations; 2023. .

[14] Wang Y, Yu J, Zhang J. Zero-shot image restoration using denoising diffusion null-space model. In: International Conference on Learning Representations; 2023. .

[15] Chen H, Howard SS. Improving Richardson–Lucy deconvolution with diffusion priors for fluorescence microscopy. arXiv preprint arXiv:260625924. 2026;https://doi.org/10.48550/arXiv.2606.25924.

[16] Bansal A, Borgnia E, Chu HM, Li J, Kazemi H, Huang F, et al. Cold diffusion: Inverting arbitrary image transforms without noise. In: Advances in Neural Information Processing Systems. vol. 36; 2023. p. 41259–41282. https://doi.org/10.52202/075280-1789.

[17] Bhadra S, Kelkar VA, Brooks FJ, Anastasio MA. On hallucinations in tomographic image reconstruction. IEEE Transactions on Medical Imaging. 2021;40(11):3249–3260. https://doi.org/10.1109/TMI.2021.3077857.

[18] Ali S, Ghatwary N, Braden B, Lamarque D, Bailey A, Realdon S, et al. Endoscopy disease detection challenge 2020. arXiv preprint arXiv:200303376. 2020;https://doi.org/10.48550/arXiv.2003.03376.

[19] Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018. p. 586–595. https://doi.org/10.1109/CVPR.2018.00068.

[20] Wiener N. Extrapolation, Interpolation, and Smoothing of Stationary Time Series. Cambridge, MA: MIT Press; 1949. https://doi.org/10.7551/mitpress/2946.001.0001.

[21] Šroubek F, Kerepecký T, Kamenický J. Iterative Wiener filtering for deconvolution with ringing artifact suppression. In: Proceedings of the 27th European Signal Processing Conference (EUSIPCO); 2019. p. 1–5. https://doi.org/10.23919/EUSIPCO.2019.8903114.

[22] Shepp LA, Vardi Y. Maximum likelihood reconstruction for emission tomography. IEEE Transactions on Medical Imaging. 1982;1(2):113–122. https://doi.org/10.1109/TMI.1982.4307558.

[23] Dey N, Blanc-Féraud L, Zimmer C, Roux P, Kam Z, Olivo-Marin JC, et al. Richardson–Lucy algorithm with total variation regularization for 3D confocal microscope deconvolution. Microscopy Research and Technique. 2006;69(4):260–266. https://doi.org/10.1002/jemt.20294.

[24] Rudin LI, Osher S, Fatemi E. Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena. 1992;60(1–4):259–268. https://doi.org/10.1016/0167-2789(92)90242-F.

[25] Krishnan D, Fergus R. Fast image deconvolution using hyper-Laplacian priors. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 22; 2009. p. 1033–1041.

[26] Fish DA, Brinicombe AM, Pike ER, Walker JG. Blind deconvolution by means of the Richardson–Lucy algorithm. Journal of the Optical Society of America A. 1995;12(1):58–65. https://doi.org/10.1364/JOSAA.12.000058.

[27] Fergus R, Singh B, Hertzmann A, Roweis ST, Freeman WT. Removing camera shake from a single photograph. ACM Transactions on Graphics. 2006;25(3):787–794. https://doi.org/10.1145/1141911.1141956.

[28] Zhuo S, Sim T. Defocus map estimation from a single image. Pattern Recognition. 2011;44(9):1852–1858. https://doi.org/10.1016/j.patcog.2011.03.009.

[29] Nah S, Kim TH, Lee KM. Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017. p. 257–265. https://doi.org/10.1109/CVPR.2017.35.

[30] Kupyn O, Budzan V, Mykhailych M, Mishkin D, Matas J. DeblurGAN: Blind motion deblurring using conditional adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2018. p. 8183–8192. https://doi.org/10.1109/CVPR.2018.00854.

[31] Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang MH, et al. Multi-stage progressive image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021. p. 14816–14826. https://doi.org/10.1109/CVPR46437.2021.01458.

[32] Li Y, Su Y, Guo M, Han X, Liu J, Vishwasrao HD, et al. Incorporating the image formation process into deep learning improves network performance. Nature Methods. 2022;19(11):1427–1437. https://doi.org/10.1038/s41592-022-01652-7.

[33] Dong J, Roth S, Schiele B. DWDN: Deep Wiener deconvolution network for non-blind image deblurring. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022;44(12):9960–9976. https://doi.org/10.1109/TPAMI.2021.3138787.

[34] Song J, Meng C, Ermon S. Denoising diffusion implicit models. In: International Conference on Learning Representations (ICLR); 2021. .

[35] Whang J, Delbracio M, Talebi H, Saharia C, Dimakis AG, Milanfar P. Deblurring via stochastic refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022. p. 16272–16282. https://doi.org/10.1109/CVPR52688.2022.01581.

[36] Chen Z, Zhang Y, Liu D, Xia B, Gu J, Kong L, et al. Hierarchical integration diffusion model for realistic image deblurring. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36; 2023. p. 29114–29125. https://doi.org/10.52202/075280-1267.

[37] Ren M, Delbracio M, Talebi H, Gerig G, Milanfar P. Multiscale structure guided diffusion for image deblurring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2023. p. 10687–10699. https://doi.org/10.1109/ICCV51070.2023.00984.

[38] Hoogeboom E, Salimans T. Blurring diffusion models. In: International Conference on Learning Representations (ICLR); 2023. .

[39] Delbracio M, Milanfar P. Inversion by direct iteration: An alternative to denoising diffusion for image restoration. Transactions on Machine Learning Research. 2023;.

[40] Liu GH, Vahdat A, Huang DA, Theodorou EA, Nie W, Anandkumar A. $I^2$SB: Image-to-image Schrödinger bridge. In: Proceedings of the 40th International Conference on Machine Learning (ICML). vol. 202 of Proceedings of Machine Learning Research; 2023. p. 22042–22062.

[41] Liu J, Wang Q, Fan H, Wang Y, Tang Y, Qu L. Residual denoising diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024. p. 2773–2783. https://doi.org/10.1109/CVPR52733.2024.00268.

[42] He JT, Tsai FJ, Peng YT, Chen MH, Lin CW, Lin YY. BlurDM: A blur diffusion model for image deblurring. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 38; 2025. p. 58117–58148. https://doi.org/10.52202/085713-1740.

[43] Choi Y, Park HS, Jeon K, Park HC, Kang SH. ReCo-Diff: Residual-conditioned deterministic sampling for cold diffusion in sparse-view CT. Physics in Medicine & Biology. 2026;71(20):205027. https://doi.org/10.1088/1361-6560/aeb34a.

[44] Kawar B, Elad M, Ermon S, Song J. Denoising diffusion restoration models. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 35; 2022. p. 23593–23606. https://doi.org/10.52202/068431-1714.

[45] Chung H, Kim J, Kim S, Ye JC. Parallel diffusion models of operator and image for blind inverse problems. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023. p. 6059–6069. https://doi.org/10.1109/CVPR52729.2023.00587.

[46] Murata N, Saito K, Lai CH, Takida Y, Uesaka T, Mitsufuji Y, et al. GibbsDDRM: A partially collapsed Gibbs sampler for solving blind inverse problems with denoising diffusion restoration. In: Proceedings of the 40th International Conference on Machine Learning (ICML). vol. 202 of Proceedings of Machine Learning Research; 2023. p. 25501–25522.

[47] Ali S, Dmitrieva M, Ghatwary N, Bano S, Polat G, Temizel A, et al. Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. Medical Image Analysis. 2021;70:102002. https://doi.org/10.1016/j.media.2021.102002.

[48] Chen T, Ma X, Bai L, Wang W, Sun Y, Zhou L. EndoIR: Degradation-agnostic all-in-one endoscopic image restoration via noise-aware routing diffusion. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40; 2026. p. 3047–3055. https://doi.org/10.1609/aaai.v40i4.37297.

[49] Miura K, Usuki S. Deep learning assisted image deblurring for endoscopy. In: Proceedings of the International Conference on Precision Engineering (ICPE 2026); 2026. Accepted for publication.

[50] Perez E, Strub F, de Vries H, Dumoulin V, Courville A. FiLM: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32; 2018. p. 3942–3951. https://doi.org/10.1609/aaai.v32i1.11671.

[51] Chen J, Mei J, Li X, Lu Y, Yu Q, Wei Q, et al. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis. 2024;97:103280. https://doi.org/10.1016/j.media.2024.103280.

[52] Borgli H, Thambawita V, Smedsrud PH, Hicks S, Jha D, Eskeland SL, et al. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific Data. 2020;7(1):283. https://doi.org/10.1038/s41597-020-00622-y.

[53] Jha D, Smedsrud PH, Riegler MA, Halvorsen P, de Lange T, Johansen D, et al. Kvasir-SEG: A segmented polyp dataset. In: MultiMedia Modeling (MMM 2020). vol. 11962 of Lecture Notes in Computer Science; 2020. p. 451–462. https://doi.org/10.1007/978-3-030-37734-2_37.