The Perception-Robustness Tradeoff in Deterministic Image Restoration

Guy Ohayon [1] **Tomer Michaeli** [2] **Michael Elad** [1]

Abstract

We study the behavior of deterministic methods for solving inverse problems in imaging. These methods are commonly designed to achieve two goals: (1) attaining high perceptual quality, and (2) generating reconstructions that are consistent with the measurements. We provide a rigorous proof that the better a predictor satisfies these two requirements, the larger its Lipschitz constant must be, regardless of the nature of the degradation involved. In particular, to approach perfect perceptual quality and perfect consistency, the Lipschitz constant of the model must grow to infinity. This implies that such methods are necessarily more susceptible to adversarial attacks. We demonstrate our theory on single image super-resolution algorithms, addressing both noisy and noiseless settings. We also show how this undesired behavior can be leveraged to explore the posterior distribution, thereby allowing the deterministic model to imitate stochastic methods.

1. Introduction

Inverse problems arise in numerous imaging applications. Indeed, the images we use on our phones, in medical diagnosis devices, and in remote sensing systems, are all created by algorithms that reconstruct and enhance raw, low-quality observations. Examples include image denoising, super-resolution, magnetic resonance imaging (MRI) reconstruction, and any image-to-image translation task.

Denoting a natural image as a random vector X, the degraded measurement as Y, and their joint distribution as $p_{X,Y}$, a deterministic image restoration algorithm[1] $\hat{X}$ (an estimator) often strives to generate outputs of high perceptual quality that are also consistent with the input measurements. Perceptual quality is commonly measured by the extent to which human observers cannot distinguish between the estimator's outputs and natural images (Isola et al., 2017; Dahl et al., 2017; Denton et al., 2015; Guadarrama et al., 2017; Iizuka et al., 2016; Salimans et al., 2016; Zhang et al., 2016; 2017). This can be quantified by the statistical distance between the distributions $p_{\hat{X}}$ and p_X (Blau & Michaeli, 2018). Of course, perceptual quality by itself does not necessarily indicate the usefulness of an algorithm. For instance, an estimator that completely ignores the input and outputs samples from p_X attains perfect perceptual quality ($p_{\hat{X}} = p_X$), but is obviously useless. Indeed, a good estimator is one whose outputs are also consistent with the measurements.

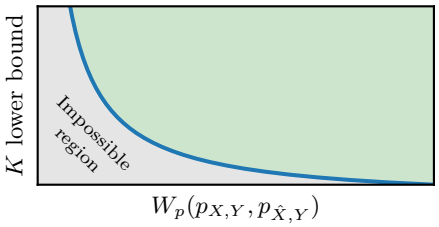

Figure 1: Qualitative illustration of Theorem 4.1. In any restoration task with a non-invertible degradation (Definition 3.1), the Lipschitz constant K of a deterministic estimator $\hat{X} = f(Y)$ is lower bounded by a function that grows to infinity as the Wasserstein distance between $p_{\hat{X},Y}$ and $p_{X,Y}$ decreases to zero.

Consistency with the measurements is best understood in the case where the degradation is deterministic, such that $Y = D(X)$. In this case, $\hat{X}$ is perfectly consistent if $D(\hat{X}) = Y$, and as a measure, consistency is typically quantified by the discrepancy between $D(\hat{X})$ and Y (Lugmayr et al., 2021; 2022). As with perceptual quality, consistency by itself is not a sufficient indication that an algorithm is useful. Indeed, the fact that the outputs are consistent with the measurements does not guarantee that they look natural. Moreover, consistency becomes a questionable concept in the setting of stochastic degradations: What can be regarded as a consistent reconstruction when many different measurements can emerge from the same source image?

[1]Faculty of Computer Science, Technion, Haifa, Israel [2]Faculty of Electrical and Computer Engineering, Technion, Haifa, Israel. Correspondence to: Guy Ohayon <ohayonguy@campus.technion.ac.il>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1]An algorithm that always produces the same output per input.

To circumvent these complexities and address all types of degradations, we integrate perceptual quality and consistency into a single concept. Specifically, we refer to the *joint perceptual quality* of an estimator as the extent to which human observers cannot distinguish between pairs of restored outputs and degraded inputs, and pairs of natural images and degraded measurements. This can be quantified by the statistical distance between $p_{\hat{X},Y}$ and $p_{X,Y}$. In Section 3.3 we explain why attaining perfect perceptual quality and perfect consistency is equivalent to attaining perfect joint perceptual quality, making it a useful stand-alone measure.

In this paper we rigorously prove that, whenever the degradation is not invertible[2], the Lipschitz constant of a deterministic estimator is bounded from below by a factor inversely proportional to the Wasserstein distance between $p_{\hat{X},Y}$ and $p_{X,Y}$ (see Figure 1 for a qualitative illustration). This means that the Lipschitz constant must grow to infinity as the Wasserstein distance between $p_{\hat{X},Y}$ and $p_{X,Y}$ approaches zero. An immediate practical implication of this result is that the higher the joint perceptual quality of an estimator (i.e., the better its perceptual quality and consistency), the more susceptible it is to input adversarial attacks, which is a clear disadvantage. On the positive side, we demonstrate that this behavior can be leveraged to explore the posterior distribution by slightly perturbing its input.

It is important to note that studying deterministic estimators of this type is not a pursuit of mere theoretical interest. In fact, estimators that satisfy $p_{\hat{X},Y} \approx p_{X,Y}$[3] appear frequently in the literature and are being used across many types of inverse problems, e.g., in tasks where the estimator is a Conditional Generative Adversarial Network (CGAN) (Mirza & Osindero, 2014; Isola et al., 2017; Park et al., 2019; Shaham et al., 2021; Wang et al., 2018a; Ma et al., 2018; Ohayon et al., 2021; Man et al., 2023; Yi & Babyn, 2018; Kim & Lee, 2020; Sun et al., 2023; Agustsson et al., 2023; Islam et al., 2020).

Our work is closely related to a previous study by Ohayon et al. (2023). Here, we generalize and extend the results in (Ohayon et al., 2023) by considering any type of degradation, by replacing intuitive claims with a formal proof, and by demonstrating the predicted behaviors on a slew of estimators (see Section 2.1 for a thorough discussion). To conclude, our main contributions are the following:

1. We rigorously prove that any deterministic estimator $\hat{X}$ that satisfies $p_{\hat{X},Y} \approx p_{X,Y}$ must possess a high Lipschitz constant. Specifically, we show that the Lipschitz constant of $\hat{X}$ is bounded from below by a factor inversely proportional to the Wasserstein distance between $p_{\hat{X},Y}$ and $p_{X,Y}$ (Theorem 4.1).

2. Through experiments on popular deterministic image super-resolution methods (Section 4.2), we demonstrate that a lower statistical distance between $p_{\hat{X},Y}$ and $p_{X,Y}$ indicates *worse* average robustness to adversarial attacks. Our measure of average robustness is effectively a *lower bound* of the Lipschitz constant of an estimator. These experiments thereby validate Theorem 4.1.

3. To demonstrate the practical relevance of Theorem 4.1, we show in Section 4.3 how an adversary can exploit the instability of a deterministic method (with high joint perceptual quality) to alter the results of an image decision-making pipeline. In Section 4.4 we also demonstrate how the erratic behavior of this estimator can practically be leveraged to explore the posterior distribution.

The paper is organized as follows. In Section 2 we provide background and motivation, explore the requirement $p_{\hat{X},Y} \approx p_{X,Y}$, and draw its connection to the literature. In Section 3 we formulate the image restoration task and set mathematical notations. In Section 4 we present our main result (Theorem 4.1) and illustrate its practical implications. In Section 5 we summarize our paper. Finally, in Section 6 we discuss the limitations of this work and propose ideas for future research.

2. Motivation and Related Works

2.1. Deterministic Degradations

In the setting of deterministic degradations, a fundamental property desired from an image restoration algorithm $\hat{X}$ is that its outputs be perfectly consistent with its inputs, i.e., $D(\hat{X}) = Y$. As Ohayon et al. (2023) explain, this condition is equivalent to $p_{Y|\hat{X}} = p_{Y|X}$. Such a property stems from the definition of solving an inverse problem (Kirsch, 2011) and is quite intuitive, as it implies that the estimated solution $\hat{X}$ is a valid explanation for the measurements. Another highly desired property of image restoration algorithms is to attain perfect perceptual quality, so that humans could not distinguish between the restored outputs and natural images. This property is equivalent to attaining $p_{\hat{X}} = p_X$ (Blau & Michaeli, 2018).

Can a deterministic estimator satisfy $p_{Y|\hat{X}} = p_{Y|X}$ (perfect consistency) and $p_{\hat{X}} = p_X$ (perfect perceptual quality) simultaneously? By making a simple use Bayes' rule, Ohayon et al. (2023) showed that the answer is negative. An estimator $\hat{X}$ satisfies these two conditions if and only if it is a posterior sampler, one that satisfies $p_{\hat{X}|Y} = p_{X|Y}$ (equivalently, $p_{\hat{X},Y} = p_{X,Y}$), and thus, a deterministic estimator cannot concurrently attain both of these properties (only a stochastic sampler from the posterior can). Since both of these properties are desired, it is therefore important to ask

[2] A degradation is invertible when X can be reconstructed from Y with zero error. See Definition 3.1.

[3] By $p_{\hat{X},Y} \approx p_{X,Y}$ we mean that the statistical distance (e.g., Wasserstein) between $p_{\hat{X},Y}$ and $p_{X,Y}$ is small.

what would be the consequences if a deterministic estimator still tries to satisfy them. As Ohayon et al. (2023) empirically demonstrate, a deterministic estimator $\hat{X}$ that satisfies $p_{\hat{X}} \approx p_X$ (high perceptual quality) and $p_{Y|\hat{X}} = p_{Y|X}$ should be erratic[4] and sensitive to input adversarial attacks.

While considering estimators that produce perfectly consistent outputs is important, it is often practically sufficient that the distance between $D(\hat{X})$ and Y would just be very small (Lugmayr et al., 2021; 2022), in which case $p_{Y|X}$ is not precisely equal to $p_{Y|\hat{X}}$. Indeed, image restoration algorithms with only near-perfect consistency are much more prevalent in practice (e.g., (Ledig et al., 2017; Wang et al., 2018b; Lugmayr et al., 2021; 2022)). Does the aforementioned sensitivity issue holds for deterministic restoration algorithms that strive for high perceptual quality and only *near-perfect* consistency? Instead of breaking down this question into separate discussions on perceptual quality and consistency, we can more generally explore the behavior of deterministic estimators that satisfy $p_{\hat{X},Y} \approx p_{X,Y}$. This includes estimators with high perceptual quality producing outputs with either perfect or near-perfect consistency. As we show in Theorem 4.1 below, estimators that satisfy $p_{\hat{X},Y} \approx p_{X,Y}$ must indeed behave erratically.

2.2. Stochastic Degradations

While deterministic degradations are prevalent in academic discussions, real-world degradations often involve a stochastic component (noise, random motion blur, etc.). Our Theorem 4.1 holds for any form of $p_{Y|X}$ (including stochastic measurements). Therefore, not only do we confirm the result of Ohayon et al. (2023) with a rigorous proof, but we also extend it to any type of degradation.

One may wonder why we are interested in deterministic estimators that satisfy $p_{\hat{X},Y} \approx p_{X,Y}$ in the case of stochastic degradations. Our motivation to explore such estimators stems from a practical observation: Aiming for this objective, whether implicitly or explicitly, is a recurring trend in the literature when using deterministic estimators in stochastic degradation settings, as we describe next.

2.3. CGAN Image Restoration Algorithms

CGANs have become highly popular for solving image inverse problems, such as image-to-image translation tasks (Isola et al., 2017; Wang et al., 2018a; Park et al., 2019; Shaham et al., 2021), image denoising (Yi & Babyn, 2018; Ma et al., 2018; Kim & Lee, 2020; Ohayon et al., 2021; Wang et al., 2021a; Chen et al., 2021; Sun et al., 2023), motion deblurring (Kupyn et al., 2018; 2019; Zhang et al., 2020), JPEG decoding (Man et al., 2023; Agustsson et al., 2023), and more. Their popularity can be partly attributed to the fact that they do not require a problem-specific loss, a model of the prior distribution, nor any assumptions about the relation between X and Y. To train a CGAN, one only needs a dataset of paired i.i.d. samples from $p_{X,Y}$.

In the image restoration literature, the term CGAN is commonly used to describe models that are trained with two different types of optimization frameworks. Some models use this term because the discriminator receives either pairs of natural images and their degraded measurements ("real" examples), or pairs of estimated outputs and the corresponding degraded inputs ("fake" examples) (Isola et al., 2017; Park et al., 2019; Shaham et al., 2021; Wang et al., 2018a; Ma et al., 2018; Ohayon et al., 2021; Man et al., 2023; Yi & Babyn, 2018; Kim & Lee, 2020; Sun et al., 2023; Agustsson et al., 2023; Islam et al., 2020). This approach follows from the original definition of a CGAN (Mirza & Osindero, 2014). Theoretically, the optimal solution of a parametric model trained solely with such an adversarial discriminator is a stochastic estimator $\hat{X}$ which satisfies $p_{\hat{X},Y} = p_{X,Y}$.

Other models use the term CGAN simply because the estimator is regarded as a conditional "generator" (conditioned on the degraded measurement), trained together with an adversarial discriminator which does not receive the degraded measurements as an additional input (Kupyn et al., 2018; 2019; Wang et al., 2018c; Wang et al.; Zhang et al., 2020; Zhu et al., 2017; Liang et al., 2021; Chen et al., 2022; Muhammad Umer et al., 2020; Zhang et al., 2021). Although these models do not explicitly aim to attain $p_{\hat{X},Y} = p_{X,Y}$, in Section 4.2 we demonstrate that they still attain a relatively small statistical distance between $p_{\hat{X},Y}$ and $p_{X,Y}$. Therefore, according to our Theorem 4.1 they must possess a high Lipschitz constant, which we indeed demonstrate through experiments. It is important to note that the means by which a small distance between $p_{\hat{X},Y}$ and $p_{X,Y}$ is obtained is irrelevant. Any deterministic estimator with high joint perceptual quality must be erratic.

Lastly, it is important to note that many CGAN-based image restoration algorithms make use of a deterministic estimator (even though the optimal solution is a stochastic one). One reason is that, in practice, the estimator often becomes nearly deterministic anyway, mapping each input to (almost) a single output (Mathieu et al., 2016; Isola et al., 2017; Ohayon et al., 2021; Park et al., 2019; Yang et al., 2019). Thus, many implementations simply omit the use of input random seed, anticipating that the estimator will likely ignore it. Another reason to omit the use of such a random seed is that in some applications it is not desirable to obtain a stochastic estimator, but rather to provide a point estimate which simply "corresponds" to the input measurement. See Appendix A for the motivation to use deterministic estimators.

We further address other related works in Appendix B, e.g., studies that discuss a tradeoff between robustness and dis-

[4] By erratic we mean that the Lipschitz constant of $\hat{X}$ is high.

tortion for deterministic restoration algorithms.

3. Problem Formulation

We adopt the Bayesian perspective for solving inverse problems (Davison, 2003; Kaipio & Somersalo, 2005; Blau & Michaeli, 2018), where a natural image $x \in \mathbb{R}^{n_x}$ is regarded as a realization of a random vector X with probability density function p_X. The degraded measurement $y \in \mathbb{R}^{n_y}$ is related to x via the conditional density $p_{Y|X}$. We are dealing with non-invertible degradations, where x cannot be retrieved from y with zero-error. Formally:

Definition 3.1. A degradation is invertible if $p_{X|Y}(\cdot|y)$ is a Dirac delta function for almost every $y \in \operatorname{supp} p_Y$.

The task of an estimator $\hat{X}$ is to estimate X from Y, i.e., $X \to Y \to \hat{X}$ is a Markov chain (X and $\hat{X}$ are statistically independent given Y). We say that $\hat{X}$ is a deterministic estimator if it can be written as a function of Y, i.e. $\hat{X} = f(Y)$. This means that $p_{\hat{X}|Y}(\cdot|y)$ is a Dirac delta function for any y. If this does not hold, then the algorithm is stochastic.

We say that the degradation is deterministic if $p_{Y|X}(\cdot|x)$ is a Dirac delta function for almost any x, and otherwise it is a stochastic degradation. In this paper we consider any form of $p_{Y|X}$, including both deterministic degradations (e.g., JPEG compression, down-sampling) and stochastic degradations (e.g., additive noise, random blur, color-jitter).

3.1. Lipschitz Continuity

We define the robustness of a deterministic estimator $\hat{X} = f(Y)$ using the well known notion of Lipschitz continuity. Formally, $\hat{X}$ is said to be Lipschitz continuous (or K-Lipschitz) if there exists $K \geq 0$ such that

$$\|f(y_1) - f(y_2)\| \leq K \|y_1 - y_2\| \tag{1}$$

for every $y_1, y_2 \in \operatorname{supp} p_Y$, and we denote $\operatorname{Lip}(\hat{X}) \coloneqq K$. When saying that a deterministic estimator $\hat{X}$ is robust, it is meant that $\operatorname{Lip}(\hat{X})$ is small. Similarly, when a deterministic estimator is said to be erratic, it means that $\operatorname{Lip}(\hat{X})$ is very large (it may be ∞).

3.2. Joint Perceptual Quality

We focus on the case where statistical distances between distributions are measured via the Wasserstein distance defined on a norm-induced metric space. The Wasserstein distance between $p_{\hat{X},Y}$ and $p_{X,Y}$ is defined by

$$W_p(p_{X,Y}, p_{\hat{X},Y}) = \inf_{p_{X_1,Y_1,X_2,Y_2} \in \mathcal{V}} \left\{ \left(\mathbb{E}\left[\left\| \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix} - \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix} \right\|^p \right] \right)^{\frac{1}{p}} \right\}, \tag{2}$$

where $\mathcal{V} = \{p_{X_1,Y_1,X_2,Y_2} : p_{X_1,Y_1} = p_{X,Y}, p_{X_2,Y_2} = p_{\hat{X},Y}\}$ and the notation $\left(\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right)$ refers to concatenation of the vectors z_1, z_2. We refer to the statistical distance between $p_{\hat{X},Y}$ and $p_{X,Y}$ as the *joint perceptual index* of $\hat{X}$. When this index is small, we say that the estimator attains high joint perceptual quality.

3.3. Consistency

Ohayon et al. (2023) considered only the case of deterministic degradations, where one can write $Y = D(X)$, i.e., $p_{Y|X}(y|x) = \delta(y - D(x))$ for all x, y. In this case, $\hat{X}$ is said to produce perfectly consistent outputs if $D(\hat{X}) = Y$. This condition holds if and only if $p_{Y|\hat{X}}(\cdot|x) = p_{Y|X}(\cdot|x)$ for almost every $x \in \operatorname{supp} p_{\hat{X}}$, which we write as $p_{Y|\hat{X}} = p_{Y|X}$ in short. Since here we consider any type of degradation, we generalize the notion of perfect consistency and say that $\hat{X}$ produces perfectly consistent outputs if $p_{Y|\hat{X}} = p_{Y|X}$, regardless of the form of $p_{Y|X}$. In Appendix D.2.2 we discuss the meaning of this definition of consistency for degradations of the form $Y = X + N$, where X and N are statistically independent.

While Ohayon et al. (2023) discussed estimators that satisfy $p_{Y|\hat{X}} = p_{Y|X}$ (where the degradation is deterministic), here we also consider estimators that do not necessarily satisfy this equality. Yet, we avoid referring to consistency as a stand-alone property of an estimator, but rather consider perceptual quality and consistency jointly via the joint perceptual index. When the joint perceptual index is zero ($p_{\hat{X},Y} = p_{X,Y}$), from Bayes' rule we have that $\hat{X}$ attains perfect perceptual quality and perfect consistency according to the previous notations, $p_{\hat{X}} = p_X$ and $p_{Y|\hat{X}} = p_{Y|X}$. In that sense, minimizing the joint perceptual index means approaching perfect perceptual quality and perfect consistency.

4. The Perception-Robustness Tradeoff

We move on to provide the main result of our paper.

Theorem 4.1. *Consider any joint probability density function $p_{X,Y}$ of the random variables X and Y, such that the degradation is not invertible (according to Definition 3.1). For every $\gamma > 0$, there exist constants $m_1, m_2 > 0$ such that*

$$\mathit{Lip}(\hat{X}) \geq \frac{m_1}{\sqrt{W_p(p_{X,Y}, p_{\hat{X},Y})}} - m_2, \tag{3}$$

for any deterministic estimator $\hat{X} = f(Y)$ of X from Y with joint perceptual index $W_p(p_{X,Y}, p_{\hat{X},Y}) \leq \gamma$.

Proof. See Appendix C. □

Note that when the degradation is not invertible, a deter-

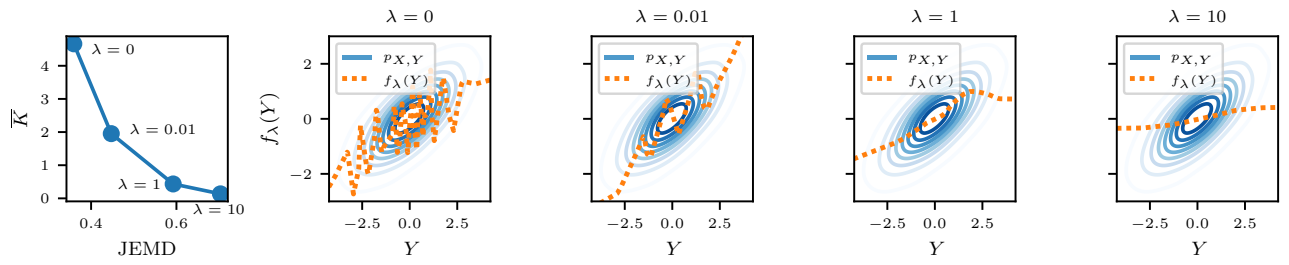

Figure 2: An illustration of Theorem 4.1 on the toy example from Section 4.1. **On the left**, we plot the Lipschitz constant lower bound $\overline{K}$ versus the JEMD (joint perceptual index) of $\hat{X}_\lambda = f_\lambda(Y)$, for several values of λ (the coefficient of the robustness loss). **On the right**, we present contour plots of the density $p_{X,Y}$ (blue concentric ellipses) and outputs from $\hat{X}_\lambda$ as a function of Y, for several values of λ. We clearly see that $\hat{X}_\lambda$ is more erratic for smaller values of λ, as anticipated by Theorem 4.1. Refer to Section 4.1 for more details.

ministic estimator $\hat{X}$ cannot satisfy $p_{\hat{X},Y} = p_{X,Y}$ (Ohayon et al., 2023), so we always have $W_p(p_{X,Y}, p_{\hat{X},Y}) > 0$. Theorem 4.1 shows that the Lipschitz constant of $\hat{X}$ is bounded from below by a function that strictly increases when $W_p(p_{X,Y}, p_{\hat{X},Y})$ decreases. Thus, as long as the joint perceptual index $W_p(p_{X,Y}, p_{\hat{X},Y})$ is small, $\hat{X}$ is not robust, and this is regardless of the nature of the degradation, the model architecture, the distribution of the images, etc. We next demonstrate Theorem 4.1 on a toy example and on widely recognized image super-resolution models.

4.1. Toy Example Illustration

Consider a simple example where $X \sim \mathcal{N}(0,1)$, $N \sim \mathcal{N}(0,1)$, X and N are statistically independent, and $Y = X + N$. Our goal is to predict X from Y. To do so, we train a deterministic denoiser $\hat{X}_\lambda = f_\lambda(Y)$ (a neural network) by optimizing the objective

$$\mathcal{L}_{CGAN} + \lambda \mathcal{L}_R. \tag{4}$$

Here, $\mathcal{L}_{CGAN}$ is a CGAN loss (Mirza & Osindero, 2014), where the discriminator is conditioned on both X and Y, so $\mathcal{L}_{CGAN}$ is theoretically minimized only when $p_{\hat{X}_\lambda,Y} = p_{X,Y}$ (such an objective promotes low joint perceptual index). The loss $\mathcal{L}_R$ is defined by

$$\mathcal{L}_R = \mathbb{E}\left[\|f_\lambda(Y) - f_\lambda(Y+Z)\|_2^2\right], \tag{5}$$

where $Z \sim \mathcal{N}(0, 0.2)$ is statistically independent of Y. This loss drives the outputs originating from Y and from (randomly chosen) inputs close to Y to be roughly equal, i.e., such an objective promotes robustness, and the level of robustness is controlled with the coefficient $\lambda \geq 0$. See Appendix D for more training details.

Lipschitz Constant Lower Bound. We compute the average of $K(y, y_{adv}) = \frac{\|f_\lambda(y) - f_\lambda(y_{adv})\|_2}{\|y - y_{adv}\|_2}$ over the entire validation set, where y is the original input, $y_{adv} = y + z$, and z is independently sampled from p_Z. The result is denoted by $\overline{K}$, and serves as a lower bound on the Lipschitz constant of the estimator.

Joint Perceptual Index. We measure the Earth Mover's Distance (EMD) between $p_{\hat{X}_\lambda,Y}$ and $p_{X,Y}$, which is the Wasserstein 1-distance between these distributions, and denote the result by Joint-EMD (JEMD). To compute this distance, we sample 10,000 points from $p_{\hat{X}_\lambda,Y}$ and $p_{X,Y}$, compute the pairwise L_1 distances between each pair of points, and then use the `ot.emd2` function from (Flamary et al., 2021) to compute the JEMD. Each sample from $p_{X,Y}$ is obtained by independently sampling a point x from p_X and a point n from p_N, and concatenating x with $y = x + n$. The samples from $p_{\hat{X}_\lambda,Y}$ are obtained by concatenating $f_\lambda(y)$ to each y.

In Figure 2 we plot $\overline{K}$ versus the JEMD for several choices of λ. As anticipated by Theorem 4.1, we clearly see that the lower the value of JEMD, the higher the value of $\overline{K}$. We also present the outputs of $\hat{X}_\lambda$, and indeed see that $\hat{X}_\lambda$ is more erratic for smaller values of λ. We further discuss this toy example in Appendix D, revealing additional observations and ideas for future research.

Note that, while $\overline{K}$ is indeed a lower bound of $\text{Lip}(\hat{X}_\lambda)$, it is not necessarily equal to the lower bound provided in Theorem 4.1. However, we demonstrate that $\overline{K}$ still exhibits the behavior anticipated by the theorem: It grows inversely proportional to the joint perceptual index. This inverse proportionality between $\overline{K}$ and the joint perceptual index offers a valuable insight. A large value of $\overline{K}$ suggests that $\hat{X}_\lambda$ is unstable near *many* different inputs, rather than just near a single point (as a large value of $\text{Lip}(\hat{X}_\lambda)$ implies). Thus, our experiments suggest that deterministic estimators achieving a low joint perceptual index are likely to be unstable near many inputs.

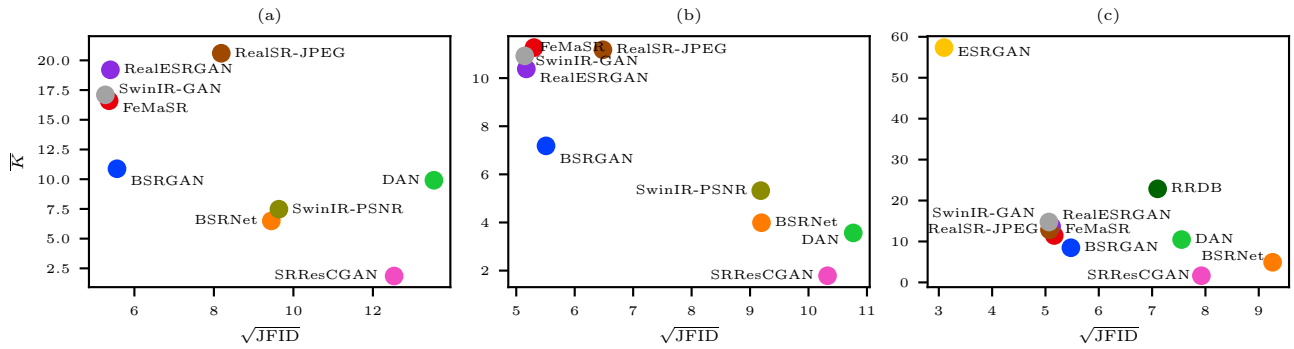

Figure 3: Quantitative demonstration of Theorem 4.1. We plot $\overline{K}$ versus $\sqrt{\text{JFID}}$ of several image super-resolution algorithms evaluated on several degradations. (a) Results on the Track2 challenge degradation by Lugmayr et al. (2019). (b) Results on the Track1 challenge degradation by Lugmayr et al. (2020). (c) Results on the standard bicubic $\times 4$ down-sampling degradation. As anticipated by Theorem 4.1, we see a tradeoff between $\overline{K}$ and $\sqrt{\text{JFID}}$ for all three degradations, i.e., the Lipschitz constant is lower bounded by a function that increases as the joint perceptual index decreases (See Section 4.2).

4.2. Quantitative Demonstration

We support Theorem 4.1 via a quantitative evaluation on the task of image super-resolution, by showing that algorithms that attain a lower joint perceptual index tend to behave more erratically. The algorithms we evaluate are RealESRGAN (Wang et al.), BSRGAN and BSRNet (Zhang et al., 2021), FeMaSR (Chen et al., 2022), DAN (setting 2) (Luo et al., 2020), ESRGAN and RRDB (Wang et al., 2018c), RealSR-JPEG and RealSR-DPED (Ji et al., 2020), FSSR-JPEG and FSSR-DPED (Fritsche et al., 2019), SRResCGAN (Muhammad Umer et al., 2020), SwinIR-GAN and SwinIR-PSNR (Liang et al., 2021), all of which perform $4\times$ up-scaling. We use the DIV2K (Agustsson & Timofte, 2017; Timofte et al., 2017) test set with the degradations from the Track1 and Track2 challenges in (Lugmayr et al., 2019), as well as the common $4\times$ bicubic down-sampling.

Lipschitz Constant Lower Bound. We conduct the ℓ_∞ I-FGSM basic attack (Choi et al., 2019; Kurakin et al., 2017) on each algorithm and each input y using $\alpha = 1/255$ and compute the ratio $K(y, y_{adv}) = \frac{\|f(y)-f(y_{adv})\|_2}{\|y-y_{adv}\|_2}$, where $f(\cdot)$ is the model and y_{adv} is the attacked version of y. We specifically choose a small value of α in order to assess the rate of change in each algorithm's output while using inputs that remain within $\operatorname{supp} p_Y$ (this constraints the attacked outputs to be in $\operatorname{supp} p_{\hat{X}}$). Indeed, at $\alpha = 1/255$, the difference between each pixel in y and y_{adv} is at most 1 gray level, i.e., the PSNR between y and y_{adv} is at least 48.13dB, making this a minor change for images. We compute the average of $K(y, y_{adv})$ over all pairs (y, y_{adv}) and denote the result by $\overline{K}$, which serves as a lower bound on the Lipschitz constant. Besides, by taking the average, we can better determine if an algorithm exhibits erratic behavior across many inputs rather than being influenced by a single outlier. Note that $\overline{K}$ has an intuitive scale: Changing the input pixels by an average of 1 gray levels leads to an average change of $\overline{K}$ gray levels in the output pixels.

Joint Perceptual Index. We approximate the joint perceptual index with a slightly tweaked version of the Fréchet Inception Distance (FID) (Heusel et al., 2017) for joint distributions. Given pairs (x, y) of source images and their degraded versions, we construct pairs $(\hat{x}, y)$ using each algorithm $f(\cdot)$, where $\hat{x} = f(y)$. We extract all patches of size 32×32 with stride 16 from each y, and all the corresponding patches of size 128×128 with stride 64 from x and $\hat{x}$. From the 100 images of the DIV2K test set, this leads to a set of $N = 46,500$ pairs of patches $\{(x^{(i)}, y^{(i)})\}_{i=1}^N$, and N pairs of patches $\{(\hat{x}^{(i)}, y^{(i)})\}_{i=1}^N$. We then feed each $x^{(i)}$, $y^{(i)}$ and $\hat{x}^{(i)}$ into the Inception-V3 network (Szegedy et al., 2016) and obtain their features $x_F^{(i)}, y_F^{(i)}$, and $\hat{x}_F^{(i)}$ from the last pooling layer, each of which is a vector of size 2048. Lastly, we compute

$$\text{FID}\left(\left\{\begin{pmatrix} x_F^{(i)} \\ y_F^{(i)} \end{pmatrix}\right\}_{i=1}^N, \left\{\begin{pmatrix} \hat{x}_F^{(i)} \\ y_F^{(i)} \end{pmatrix}\right\}_{i=1}^N\right) \tag{6}$$

and denote the result by Joint-FID (JFID).

In Figure 3 we present the plot of $\overline{K}$ versus $\sqrt{\text{JFID}}$ for each degradation. We take the square-root of JFID so that both axes have the same scale. The results clearly show that smaller values of $\sqrt{\text{JFID}}$ correspond to larger values of $\overline{K}$. For a better view of the tradeoff between $\overline{K}$ and $\sqrt{\text{JFID}}$, notice that in each plot we omit the algorithms which are too far-off from the tradeoff. In Appendix E we provide complementary figures that include all the evaluated algorithms, as well as distortion-perception plots of RMSE versus $\sqrt{\text{FID}}$, where FID is computed between the

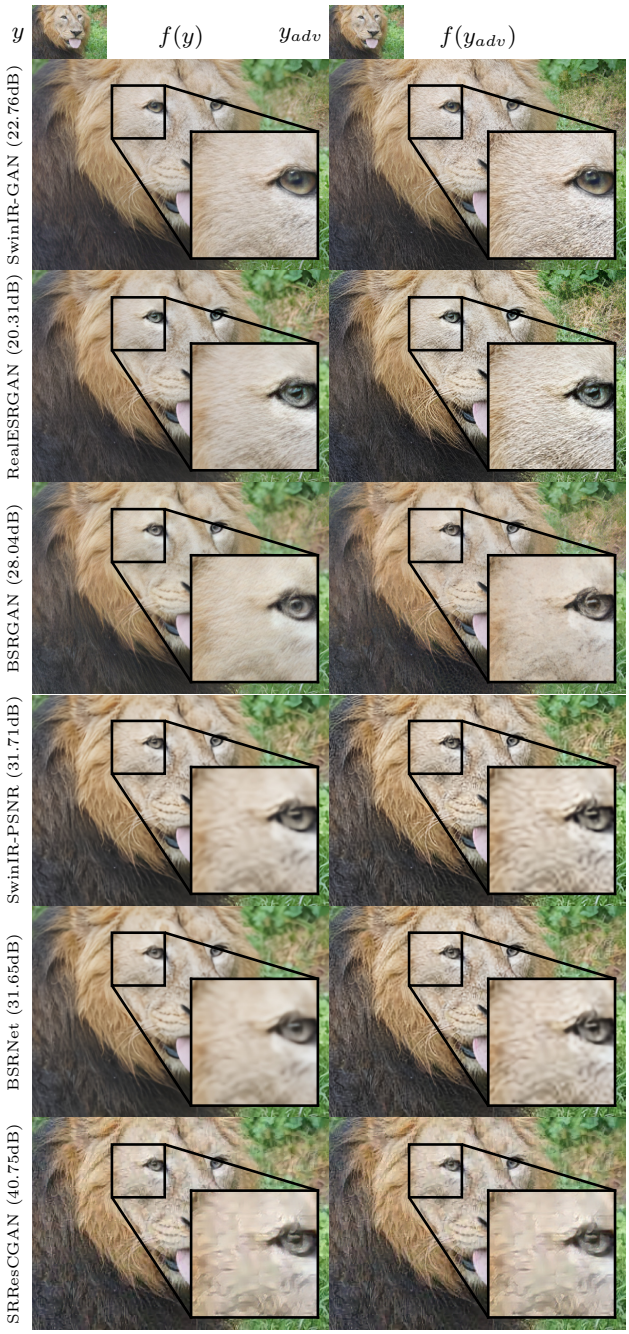

Figure 4: Visual comparison of some of the super-resolution algorithms evaluated in Section 4.2 on the Track2 challenge degradation by Lugmayr et al. (2019), sorted from top to bottom by their JFID (increasing). The original and the attacked outputs are denoted by $f(y)$ and $f(y_{adv})$, respectively. The PSNR between y and y_{adv} is at least 48.13dB (obtained with $\alpha = 1/255$ in I-FGSM), so the difference is visually negligible (see the attacked input y_{adv} of SwinIR-GAN at the top). The PSNR between $f(y)$ and $f(y_{adv})$ is reported next to the name of each algorithm. As can be seen, the better the joint perceptual quality, the higher the sensitivity to adversarial attacks.

sets $\{x_F^{(i)}\}_{i=1}^N$ and $\{\hat{x}_F^{(i)}\}_{i=1}^N$ (i.e., it is a perceptual quality evaluation of each algorithm). In Appendix E we present additional plots to support Theorem 4.1, by replacing the JFID with alternative measures related to Kernel Inception Distance (KID) (Binkowski et al., 2018) and Precision and Recall (Kynkäänniemi et al., 2019). In Figure 4 we present visual results of selected methods that appear in favorable locations in Figure 3.

4.3. Adversarial Attacks

While slightly perturbing the original input y of a deterministic estimator with high joint perceptual quality may lead to a large change in its output, the new output may still look "valid", i.e., it may still be in $\operatorname{supp} p_{X|Y}(\cdot|y)$ (see, e.g., the outputs of SwinIR-GAN in Figure 4). In what sense then can this new output be considered as "adversarial"? More generally, what is the practical relevance of Theorem 4.1 when the outputs originating from any slightly perturbed input y_{adv} may remain valid, even when the algorithm is highly erratic? Here we demonstrate a possible drawback of such erratic deterministic estimators that arises when using them in a decision-making pipeline.

Consider an experiment where the goal is to predict the gender in low-resolution noisy face images, by performing gender classification after enhancing the images' quality with a super-resolution algorithm. We take the first 1000 face images from the CelebA-HQ (Karras et al., 2018) data set and degrade each image by resizing it to 32×32 pixels using bilinear interpolation, and then adding isotropic Gaussian noise with zero mean and standard deviation $10/255$. The super-resolution algorithm we use is GFPGAN (Wang et al., 2021b), a deterministic face image restoration model with high perceptual quality, which we expect to also attain high joint perceptual quality (we use the v1.3 checkpoint and official code provided by the authors). This algorithm performs $\times 16$ up-scaling, resulting in an output image of size 512×512. We attack each degraded image using a tweaked version of the I-FGSM basic attack with $\alpha = 16/255$ and $T = 100$. Instead of using the L_2 loss in I-FGSM like in (Choi et al., 2019), we forward each attacked output through a vision transformer model (Dosovitskiy et al., 2021) trained to classify the gender in face images (Dwikifirdaus, 2022), and then try to maximize the log-probability of the class "female". Put simply, we forward each low-quality image through GFPGAN and then feed the result to the face gender classification model. We then use the I-FGSM update rule (Choi et al., 2019) to maximize the softmax probability of the category "female".

Observe in Figure 5 that there is barely any visual difference between the original and the attacked inputs. Yet, compared to the original outputs, the ones resulting from the attacked inputs exhibit newly generated features, such as makeup

Figure 5: Adversarial attacks on low-resolution face images intended to alter the outputs of GFPGAN and RRDB to produce a face which is classified as "female" rather than "male". From left to right: Original input y, attacked input y_{adv}, original output $f(y)$, output obtained from the attacked input $f(y_{adv})$. The perturbed inputs y_{adv} are obtained with $\alpha = 16/255$ in I-FGSM. With such a value of α we barely see any visual difference between y and y_{adv}. As anticipated, y_{adv} indeed leads to outputs with newly generated features when using GFPGAN (e.g., makeup), yet we barely see any significant change for RRDB. An image gender classifier associates these features with the "female" category, as the predicted class in the GFPGAN outputs switch from "male" when the input is y, to "female" when the input is y_{adv}. Refer to Section 4.3 for more details.

and jewelry. As for a quantitative analysis, we use (Serengil & Ozpinar, 2021) to perform face gender classification on all the source images, the outputs resulting from the original inputs, and the outputs resulting from the attacked inputs. Note that we use a different gender classification model for

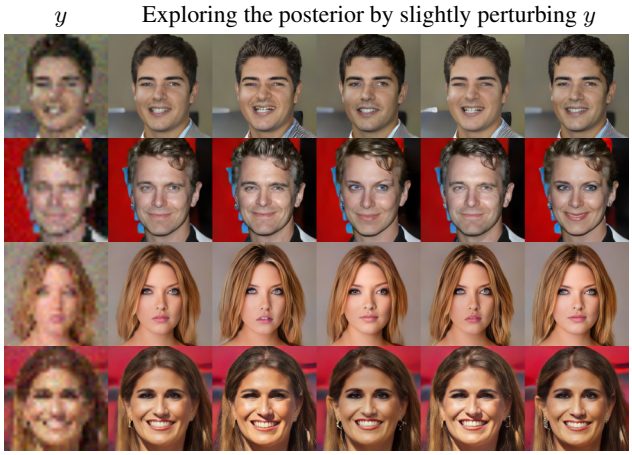

Figure 6: Exploring the posterior distribution with GFPGAN (Wang et al., 2021b). In each row we show the original input y and outputs generated from "adversarial" inputs, each of which is a slight perturbation of y (with a minimum of 30dB PSNR). By slightly perturbing y, we observe meaningful semantic variation in the outputs, such as change of gender and hair textures, all of which are expected to be possible varying features in the posterior.

evaluation to alleviate the possibility of over-fitting. 61.2% of the source images and 63.3% of the original reconstructed images are classified as "female", so GFPGAN is quite accurate in terms of reconstructing the true gender population in the data. Yet, the percentage of "female"-classified attacked images is increased to 72.2%.

Lastly, to support our claim that such a vulnerability is associated with the joint perceptual quality of the algorithm, we also train the RRDB model (Wang et al., 2018c) solely with the L_1 reconstruction loss, using the exact same training setup as that of GFPGAN (same training data, same degradation, same input normalization, etc.). Additional details are disclosed in Appendix F. This model is expected to produce blurry images with low perceptual quality (Blau & Michaeli, 2018), so it is also expected to produce lower joint perceptual quality compared to GFPGAN (see Figure 5 for a qualitative comparison). We evaluate the outputs of the trained model and observe that, unlike GFPGAN, this model is robust to such attacks, as 59.1% and 56.9% of its original and attacked outputs, respectively, are classified as "female". Interestingly, even though we attack the model to produce "female"-like outputs, the fraction of "female"-classified images is *reduced* for the attacked outputs. We discuss this phenomenon and provide more visual results in Appendix F.1. In Appendix F.2 we present similar experiments where we attack the models to produce faces of older people, and find that, compared to RRDB, GFPGAN is more vulnerable to such attacks as well.

4.4. Exploring the Posterior Distribution

Exploring the posterior distribution in imaging inverse problems may hold a practical significance in many scenarios (please refer to Appendix G for a discussion). Here, we demonstrate that, as an implication of Theorem 4.1, doing so is possible with a *deterministic* estimator. Specifically, when $\hat{X}$ is a deterministic estimator and $W_p(p_{X,Y}, p_{\hat{X},Y})$ is small, we can intuitively say that the outputs produced by $\hat{X}$ given the inputs in a small ball around y should approximately constitute samples from the posterior distribution (this idea inspired our proof of Theorem 4.1). Let us leverage this insight.

To explore the posterior, for each input $y^{(1)}$ we search for a new input $y^{(2)}$ close to $y^{(1)}$, such that $f(y^{(2)})$ is far from $f(y^{(1)})$. Then, to obtain another sample, we seek another input $y^{(3)}$ such that $f(y^{(3)})$ is far from both $f(y^{(1)})$ and $f(y^{(2)})$. We continue in this fashion to obtain more samples. This is a Farthest Point Sampling (FPS) (Eldar et al., 1994) approach to generate an arbitrary number of outputs that correspond to (almost) the same input (refer to Appendix G for a formal description). We use this algorithm only for demonstration purposes and do not claim any optimality.

Figure 6 presents such an experiment. We take several images from the CelebA-HQ data set and degrade each image using the same procedure as described in Section 4.3. To explore the posterior, we execute the FPS algorithm on GFPGAN. In Figure 6 we present several examples of degraded inputs and corresponding sequential samples that result from the FPS algorithm. We indeed see that by slightly perturbing the original input, we obtain diverse outputs that all seem to correspond to the measurements.

5. Summary

We prove that the better the perceptual quality and consistency of a deterministic image restoration algorithm, the higher its Lipschitz constant must be. We confirm that widely used image super-resolution algorithms indeed adhere to such a tradeoff, and perform experiments that showcase the practical consequences (both positive and negative) of this result.

6. Limitations

While Theorem 4.1 holds true for any signal restoration task involving a non-invertible degradation, in this work we limit its demonstration to the image super-resolution task. It would be interesting to explore the discovered tradeoff in alternative tasks, such as MRI reconstruction, image-to-image translation, or other signal restoration tasks such as video, audio, or even text restoration.

In Section 4.2, we evaluate deterministic algorithms with high joint perceptual quality, all of which are GAN-based, as currently such methods are the state-of-the-art. However, other types of deterministic methods that achieve high joint perceptual quality may emerge in the future. If so, it would be interesting to assess where those methods lie on the perception-robustness plane compared to GAN-based methods.

In Section 4.4, we illustrate the feasibility of exploring the posterior distribution using a deterministic estimator with high joint perceptual quality. However, we provide only qualitative results and do not showcase the practical applications of this phenomenon. For instance, it could prove beneficial in cases where practitioners employ a deterministic algorithm but still desire to perform uncertainty quantification of the restoration task at hand (e.g., to observe the range of possible pixel values in a specific region in the image).

Impact Statement

Our work uncovers a fundamental tradeoff between the stability of deterministic restoration algorithms and their performance, as measured by their joint perceptual index. This tradeoff presents a vulnerability that could be exploited by malicious users to introduce bias and manipulate the outcomes of imaging systems (e.g., through adversarial attacks). We encourage practitioners to address this inherent tradeoff by trading performance in order to safeguard against such misuse.

Acknowledgments

We thank Yonatan Shadmi for reviewing the proof of our theorem. This research was partially supported by the Israel Science Foundation (ISF) under Grant 2318/22 and by the Council For Higher Education - Planning & Budgeting Committee.

References

Agustsson, E. and Timofte, R. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

Agustsson, E., Minnen, D., Toderici, G., and Mentzer, F. Multi-realism image compression with a conditional generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22324–22333, June 2023.

Belhasin, O., Romano, Y., Freedman, D., Rivlin, E., and Elad, M. Principal uncertainty quantification with spatial correlation for image restoration problems. *arXiv*, 2024.

Binkowski, M., Sutherland, D. J., Arbel, M., and Gretton,

A. Demystifying MMD gans. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=r1lUOzWCW.

Blau, Y. and Michaeli, T. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Chen, C., Shi, X., Qin, Y., Li, X., Han, X., Yang, T., and Guo, S. Real-world blind super-resolution via feature matching with implicit high-resolution priors. 2022.

Chen, J., Zhang, C., Traverso, A., Zhovannik, I., Dekker, A., Wee, L., and Bermejo, I. Generative models improve radiomics reproducibility in low dose cts: a simulation study. *Physics in Medicine & Biology*, 66(16):165002, 2021.

Choi, J.-H., Zhang, H., Kim, J.-H., Hsieh, C.-J., and Lee, J.-S. Evaluating robustness of deep image super-resolution against adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Dahl, R., Norouzi, M., and Shlens, J. Pixel recursive super resolution. 2017. URL https://arxiv.org/abs/1702.00783.

Davison, A. C. *Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2003. doi: 10.1017/CBO9780511815850.

Denton, E. L., Chintala, S., szlam, a., and Fergus, R. Deep generative image models using a laplacian pyramid of adversarial networks. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/aa169b49b583a2b5af89203c2b78c67c-Paper.pdf.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Dwikifirdaus, R. Face gender classification with a vision transformer. https://huggingface.co/rizvandwiki/gender-classification-2, 2022. Accessed: 2023-10-23.

Eldar, Y., Lindenbaum, M., Porat, M., and Zeevi, Y. The farthest point strategy for progressive image sampling. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 2 - Conference B: Computer Vision & Image Processing. (Cat. No.94CH3440-5)*, pp. 93–97 vol.3, 1994. doi: 10.1109/ICPR.1994.577129.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Freirich, D., Michaeli, T., and Meir, R. A theory of the distortion-perception tradeoff in wasserstein space. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 25661–25672. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/d77e68596c15c53c2a33ad143739902d-Paper.pdf.

Fritsche, M., Gu, S., and Timofte, R. Frequency separation for real-world super-resolution. In *IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019.

Gandikota, K. V., Chandramouli, P., Dröge, H., and Moeller, M. Evaluating adversarial robustness of low dose CT recovery. In *Medical Imaging with Deep Learning*, 2023. URL https://openreview.net/forum?id=L-N1uAxfQk1.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Gottschling, N. M., Antun, V., Hansen, A. C., and Adcock, B. The troublesome kernel – on hallucinations, no free lunches and the accuracy-stability trade-off in inverse problems. *arXiv*, 2023.

Guadarrama, S., Dahl, R., Bieber, D., Norouzi, M., Shlens, J., and Murphy, K. Pixcolor: Pixel recursive colorization.

Proceedings of the 28th British Machine Vision Conference (BMVC), 2017. URL https://arxiv.org/abs/1705.07208.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Iizuka, S., Simo-Serra, E., and Ishikawa, H. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Trans. Graph.*, 35(4), jul 2016. ISSN 0730-0301. doi: 10.1145/2897824.2925974. URL https://doi.org/10.1145/2897824.2925974.

Islam, M. J., Xia, Y., and Sattar, J. Fast underwater image enhancement for improved visual perception. *IEEE Robotics and Automation Letters (RA-L)*, 5(2):3227–3234, 2020.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

Ji, X., Cao, Y., Tai, Y., Wang, C., Li, J., and Huang, F. Real-world super-resolution via kernel estimation and noise injection. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

Kaipio, J. and Somersalo, E. *Statistical and Computational Inverse Problems*. Springer, Dordrecht, 2005. doi: 10.1007/b138659. URL https://cds.cern.ch/record/1338003.

Karkkainen, K. and Joo, J. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1548–1558, 2021.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hk99zCeAb.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, 2019. doi: 10.1109/CVPR.2019.00453.

Kim, H.-J. and Lee, D. Image denoising with conditional generative adversarial networks (cgan) in low dose chest images. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 954:161914, 2020. ISSN 0168-9002. doi: https://doi.org/10.1016/j.nima.2019.02.041. URL https://www.sciencedirect.com/science/article/pii/S0168900219302293. Symposium on Radiation Measurements and Applications XVII.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. doi: 10.48550/ARXIV.1412.6980. URL https://arxiv.org/abs/1412.6980.

Kirsch, A. *An Introduction to the Mathematical Theory of Inverse Problems*, volume 120. 01 2011. ISBN 978-1-4419-8473-9. doi: 10.1007/978-1-4419-8474-6.

Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., and Matas, J. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Kupyn, O., Martyniuk, T., Wu, J., and Wang, Z. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Kurakin, A., Goodfellow, I. J., and Bengio, S. Adversarial machine learning at scale. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=BJm4T4Kgx.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. Improved precision and recall metric for assessing generative models. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., and Timofte, R. Swinir: Image restoration using swin transformer. *arXiv preprint arXiv:2108.10257*, 2021.

Lugmayr, A., Danelljan, M., Timofte, R., Fritsche, M., Gu, S., Purohit, K., Kandula, P., Suin, M., Rajagoapalan, A., Joon, N. H., et al. Aim 2019 challenge on real-world image super-resolution: Methods and results. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3575–3583. IEEE, 2019.

Lugmayr, A., Danelljan, M., and Timofte, R. Ntire 2020 challenge on real-world image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

Lugmayr, A., Danelljan, M., and Timofte, R. Ntire 2021 learning the super-resolution space challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 596–612, June 2021.

Lugmayr, A., Danelljan, M., Timofte, R., Kim, K.-w., Kim, Y., Lee, J.-y., Li, Z., Pan, J., Shim, D., Song, K.-U., Tang, J., Wang, C., and Zhao, Z. Ntire 2022 challenge on learning the super-resolution space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 786–797, June 2022.

Luo, Z., Huang, Y., Li, S., Wang, L., and Tan, T. Unfolding the alternating optimization for blind super resolution. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020.

Ma, Y., Chen, X., Zhu, W., Cheng, X., Xiang, D., and Shi, F. Speckle noise reduction in optical coherence tomography images based on edge-sensitive cgan. *Biomedical optics express*, 9(11):5129–5146, 2018.

Man, S., Ohayon, G., Adrai, T., and Elad, M. High-perceptual quality jpeg decoding via posterior sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1272–1282, June 2023.

Mathieu, M., Couprie, C., and LeCun, Y. Deep multi-scale video prediction beyond mean square error. In Bengio, Y. and LeCun, Y. (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.05440.

Mescheder, L., Nowozin, S., and Geiger, A. Which training methods for gans do actually converge? In *International Conference on Machine Learning (ICML)*, 2018.

Mirza, M. and Osindero, S. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. URL http://arxiv.org/abs/1411.1784.

Muhammad Umer, R., Luca Foresti, G., and Micheloni, C. Deep generative adversarial residual convolutional networks for real-world super-resolution. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

Mukherjee, S., Hauptmann, A., Oktem, O., Pereyra, M., and Schonlieb, C. Learned reconstruction methods with convergence guarantees: A survey of concepts and applications. *IEEE Signal Processing Magazine*, 40 (1):164–182, January 2023. ISSN 1053-5888. doi: 10.1109/MSP.2022.3207451. Publisher Copyright: © 1991-2012 IEEE.

Nate Raw. vit-age-classifier (revision 461a4c4). 2023. doi: 10.57967/hf/1259. URL https://huggingface.co/nateraw/vit-age-classifier.

Ohayon, G., Adrai, T., Vaksman, G., Elad, M., and Milanfar, P. High perceptual quality image denoising with a posterior sampling cgan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1805–1813, 2021.

Ohayon, G., Adrai, T. J., Elad, M., and Michaeli, T. Reasons for the superiority of stochastic estimators over deterministic ones: Robustness, consistency and perceptual quality. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26474–26494. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/ohayon23a.html.

Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., and Chen, X. Improved techniques for training gans. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf.

Serengil, S. I. and Ozpinar, A. Hyperextended lightface: A facial attribute analysis framework. In *2021 Interna-*

tional Conference on Engineering and Emerging Technologies (ICEET), pp. 1–4. IEEE, 2021. doi: 10.1109/ICEET53442.2021.9659697. URL https://doi.org/10.1109/ICEET53442.2021.9659697.

Shaham, T. R., Gharbi, M., Zhang, R., Shechtman, E., and Michaeli, T. Spatially-adaptive pixelwise networks for fast image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14882–14891, June 2021.

Sun, J., Jiang, H., Du, Y., Li, C.-Y., Wu, T.-H., Liu, Y.-H., Yang, B.-H., and Mok, G. S. Deep learning-based denoising in projection-domain and reconstruction-domain for low-dose myocardial perfusion spect. *Journal of Nuclear Cardiology*, 30(3):970–985, 2023.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016. doi: 10.1109/CVPR.2016.308.

Timofte, R., Agustsson, E., Van Gool, L., Yang, M.-H., Zhang, L., Lim, B., et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

Wang, M., Zhu, W., Yu, K., Chen, Z., Shi, F., Zhou, Y., Ma, Y., Peng, Y., Bao, D., Feng, S., Ye, L., Xiang, D., and Chen, X. Semi-supervised capsule cgan for speckle noise reduction in retinal oct images. *IEEE Transactions on Medical Imaging*, PP:1–1, 01 2021a. doi: 10.1109/TMI.2020.3048975.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018a.

Wang, X., Xie, L., Dong, C., and Shan, Y. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops (ICCVW)*.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., and Change Loy, C. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018b.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., and Loy, C. C. Esrgan: Enhanced super-resolution generative adversarial networks. In *The European Conference on Computer Vision Workshops (ECCVW)*, September 2018c.

Wang, X., Li, Y., Zhang, H., and Shan, Y. Towards real-world blind face restoration with generative facial prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021b.

Yang, D., Hong, S., Jang, Y., Zhao, T., and Lee, H. Diversity-sensitive conditional generative adversarial networks. In *Proceedings of the International Conference on Learning Representations*, 2019.

Yi, X. and Babyn, P. Sharpness-aware low-dose ct denoising using conditional generative adversarial network. *Journal of digital imaging*, 31:655–669, 2018.

Zhang, K., Luo, W., Zhong, Y., Ma, L., Stenger, B., Liu, W., and Li, H. Deblurring by realistic blurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Zhang, K., Liang, J., Van Gool, L., and Timofte, R. Designing a practical degradation model for deep blind image super-resolution. In *IEEE International Conference on Computer Vision*, pp. 4791–4800, 2021.

Zhang, R., Isola, P., and Efros, A. A. Colorful image colorization. In *ECCV*, 2016.

Zhang, R., Zhu, J.-Y., Isola, P., Geng, X., Lin, A. S., Yu, T., and Efros, A. A. Real-time user-guided image colorization with learned deep priors. *ACM Transactions on Graphics (TOG)*, 9(4), 2017.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

A. On the Desire to Use Deterministic Image Restoration Methods

In many applications users desire a deterministic outcome every time an input image is processed. This allows practitioners to trust the results and make informed decisions based on them. For instance, a deterministic image restoration algorithm ensures that doctors and radiologists do not get different interpretations from the same input. They would typically desire a single "best" estimate to diagnose a condition, rather than multiple images from which they may not draw a definite conclusion. While in those cases some may suggest to somehow select only one of the outputs produced by a stochastic estimator, why go through the burden of obtaining a sampler if we do not intend to leverage its full potential anyways (e.g., uncertainty evaluation)?

Another reason that deterministic image restoration algorithms are often desirable is that they are simpler to obtain, especially in high dimensional distributions such as that of natural images. In the era of deep learning, designing a well-performing model is as straightforward as designing a loss function that targets our specific desired objective. With such a loss function, we can optimize a neural network to obtain the objective. Hence, it is appealing (and typical) to address image inverse problems simply by training a neural network (deterministic estimator) to meet certain objectives. Perhaps the most common objectives are high perceptual quality (e.g., obtaining a low Wasserstein distance between $p_{\hat{X}}$ and p_X) and low average distortion (e.g., obtaining a low MSE), which may be at odds with each other (Blau & Michaeli, 2018).

While there might be situations where leveraging the diverse outcomes from stochastic methods could be beneficial, especially when drawing insights about uncertainties or making probabilistic decisions, these come with their own set of challenges. Processing multiple outcomes to derive a singular, actionable insight often involves additional computational steps. This not only increases the processing time but can also introduce complexities in decision-making. On the other hand, a deterministic estimator streamlines this process by directly predicting the desired result. It offers a straightforward approach that eliminates the need for sifting through multiple potential outcomes and consolidates the decision-making process. In many scenarios, especially where timely and clear responses are essential (e.g., self driving cars), the efficiency and directness of deterministic algorithms make them a favorable choice. Perhaps for the reasons above, among potentially others, we see that deterministic image restoration algorithms are much more commonly used in practice (compared to stochastic ones).

B. Additional Related Works

Previous works have discussed that image restoration algorithms with extremely low values of distortion tend to be unstable (Gottschling et al., 2023; Mukherjee et al., 2023). Namely, as Gottschling et al. (2023) prove, a restoration algorithm must be unstable (with a high Lipschitz constant) if it produces accurate estimates for two *distinct* source images that correspond to almost identical degraded measurements. This phenomenon stems directly from the definition of Lipschitz continuity. Indeed, an algorithm that generates two distinct outputs from (almost) indistinguishable inputs, is, by definition, unstable. Importantly, such a tradeoff is fundamentally different than the tradeoff we reveal in our paper: We are discussing perceptual quality & consistency, not distortion. Moreover, the tradeoff we reveal cannot be alleviated. It has nothing to do with issues such as over-fitting, "over-performance", or "inconsistent-performance", as discussed in (Gottschling et al., 2023).

As we show in Section 4.3, one can imitate stochastic sampling by slightly perturbing the input of an unstable deterministic restoration algorithm. This approach allows to explore the posterior distribution if the algorithm attains high joint perceptual quality. The idea to imitate stochastic sampling by perturbing the input of a deterministic algorithm has been proposed before (Gandikota et al., 2023), but not fully demonstrated with a concrete algorithm such as Algorithm 1.

C. Proof of Theorem 4.1

Here we prove Theorem 4.1. We start with a useful definition and several lemmas.

Definition C.1. A degradation is called β-ill-posed if there exist $\beta > 0$ and $\mathcal{S}_y$ with $P(Y \in \mathcal{S}_y) > 0$, such that for all $y \in \mathcal{S}_y$ there exist sets $Q_1(y)$ and $Q_2(y)$, where $P(X \in Q_i(y)|Y = y) > 0$ and $\left\|x^{(1)} - x^{(2)}\right\| \geq \beta$ for every $(x^{(1)}, x^{(2)}) \in Q_1(y) \times Q_2(y)$.

We use Definition C.1 since it allows us to link β to the Lipschitz constant of a deterministic estimator. Importantly, we show that such a definition does not pose any special requirements on the degradation, i.e., any non-invertible degradation is also β-ill-posed for some $\beta > 0$. The lower bound of the Lipschitz constant K will then depend on β.

Lemma C.2. *Any non-invertible degradation is β-ill-posed for some $\beta > 0$.*

Proof. The degradation is not invertible, so there exists $\mathcal{S}_y$ with $P(Y \in \mathcal{S}_y) > 0$, such that $p_{X|Y}(\cdot|y)$ is not a delta for every $y \in \mathcal{S}_y$. From the definition of a probability density function, for any $y \in \mathcal{S}_y$ there exists a point $x_1 \in \operatorname{supp} p_{X|Y}(\cdot|y)$ such that for every $\lambda > 0$ we have

$$P(X \in B(x_1, \lambda)|Y = y) > 0, \tag{7}$$

where

$$B(x, \lambda) = \{x' \,:\, \|x - x'\| < \lambda\}. \tag{8}$$

This is true since otherwise $p_{X|Y}(\cdot|y)$ integrates to zero. It is well known that

$$\{x_1\} = \cap_{n\geq 1} B(x_1, \frac{1}{n}),\ n \in \mathbb{N}. \tag{9}$$

Suppose by contradiction that for every $n \in \mathbb{N}$ it holds that

$$P(X \in B(x_1, \frac{1}{n})|Y = y) = 1, \tag{10}$$

so

$$P(X \notin B(x_1, \frac{1}{n})|Y = y) = P(X \in B^c(x_1, \frac{1}{n})|Y = y) = 0. \tag{11}$$

Thus, it holds that

$$\begin{aligned} P(X \in \{x_1\}|Y = y) &= P(X \in \cap_{n\geq 1} B(x_1, \frac{1}{n})|Y = y) && (12)\\ &= 1 - P(X \in \cup_{n\geq 1} B^c(x_1, \frac{1}{n})|Y = y) && (13)\\ &\geq 1 - \sum_{n\geq 1} P(X \in B^c(x_1, \frac{1}{n})|Y = y) = 1, && (14) \end{aligned}$$

so $P(X \in \{x_1\}|Y = y) = 1$, which is a contradiction to the fact that $p_{X|Y}(\cdot|y)$ is not a delta. Hence, there exists $n^* \in \mathbb{N}$ such that

$$P(X \in B(x_1, \frac{1}{n^*})|Y = y) < 1, \tag{15}$$

and it holds that

$$P(X \in B^c(x_1, \frac{1}{n^*})|Y = y) > 0. \tag{16}$$

Again, from the definition of a density, there exists $x_2 \in B^c(x_1, \frac{1}{n^*})$, $x_2 \neq x_1$, such that

$$P(X \in B(x_2, \lambda)|Y = y) > 0, \tag{17}$$

and this is true for every $\lambda > 0$. In particular, for $\lambda^* = \|x_1 - x_2\| > 0$ we have

$$\begin{aligned} P(X \in B(x_1, \frac{\lambda^*}{4})|Y = y) &> 0, && (18)\\ P(X \in B(x_2, \frac{\lambda^*}{4})|Y = y) &> 0, && (19) \end{aligned}$$

and thus, for every $(x^{(1)}, x^{(2)}) \in B(x_1, \frac{\lambda^*}{4}) \times B(x_2, \frac{\lambda^*}{4})$ we have $\left\|x^{(1)} - x^{(2)}\right\| \geq \frac{\lambda^*}{2}$.

To summarize the proof thus far, for every $y \in \mathcal{S}_y$ we have found two sets $Q_1(y) = B(x_1, \frac{\lambda^*}{4})$ and $Q_2(y) = B(x_2, \frac{\lambda^*}{4})$ (notice that λ^* depends on y), such that for every $(x^{(1)}, x^{(2)}) \in Q_1(y) \times Q_2(y)$ we have $\left\|x^{(1)} - x^{(2)}\right\| \geq \frac{\lambda^*}{2}$. Yet, we are not finished with the proof since we need to find a "shared" $\beta > 0$ over a set of y's with positive probability.

Let

$$\mathcal{S}_y(k) = \left\{ y \in \mathcal{S}_y \ : \ \frac{\lambda^*}{2} \geq \frac{1}{k} \right\}, \ k \in \mathbb{N}. \tag{20}$$

For every $y \in \mathcal{S}_y$ it holds that $\lambda^* > 0$, so it is clear that there exists $k \geq 1$ such that $\frac{\lambda^*}{2} \geq \frac{1}{k}$. Hence, we have

$$\mathcal{S}_y = \cup_{k \geq 1} \mathcal{S}_y(k). \tag{21}$$

Now, there exists $k^* \geq 1$ such that $P(Y \in \mathcal{S}_y(k^*)) > 0$, since otherwise $\mathcal{S}_y$ is a countable union of sets with zero probability, which is a contradiction to the fact that $P(Y \in \mathcal{S}_y) > 0$. Hence, for every $y \in \mathcal{S}_y(k^*)$ there exists sets $Q_1(y), Q_2(y)$, such that for every $(x^{(1)}, x^{(2)}) \in Q_1(y) \times Q_2(y)$ we have $\left\|x^{(1)} - x^{(2)}\right\| \geq \frac{\lambda^*}{2} \geq \frac{1}{k^*}$. Choosing $\beta = \frac{1}{k^*}$, the set $\mathcal{S}_y(k^*)$, and the sets $Q_1(y), Q_2(y)$ (for every $y \in \mathcal{S}_y(k^*)$) concludes the proof. □

Lemma C.3. *Assume a β-ill-posed degradation with the sets $\mathcal{S}_y, Q_1(y), Q_2(y)$. There exist $k > 0$ and a subset $\tilde{S}_y \subseteq \mathcal{S}_y$ with $P(Y \in \tilde{S}_y) > 0$, such that for every $y \in \tilde{S}_y$ it holds that $P(X \in Q_i(y)|Y = y) \geq k$.*

Proof. Let

$$\mathcal{S}_y(n) = \left\{ y \in \mathcal{S}_y \ : \ P(X \in Q_1(y)|Y = y) \geq \frac{1}{n}, \ P(X \in Q_2(y)|Y = y) \geq \frac{1}{n} \right\}, \tag{22}$$

and notice that

$$\mathcal{S}_y = \cup_{n \geq 1} \mathcal{S}_y(n). \tag{23}$$

There exists $n^* \geq 1$ such that

$$P(Y \in \mathcal{S}_y(n^*)) > 0, \tag{24}$$

since otherwise $\mathcal{S}_y$ is a countable union of sets with zero probability, implying that $P(Y \in \mathcal{S}_y) = 0$. By the definition of $\mathcal{S}_y(n^*)$, for every $y \in \mathcal{S}_y(n^*)$ we have

$$P(X \in Q_1(y)|Y = y) \geq \frac{1}{n^*}, \tag{25}$$

$$P(X \in Q_2(y)|Y = y) \geq \frac{1}{n^*}. \tag{26}$$

Choosing $k = \frac{1}{n^*}$ and $\tilde{\mathcal{S}}_y = \mathcal{S}_y(n^*)$ concludes the proof. □

Lemma C.4. *There exists $C > 0$ such that, for any vectors x_1, y_1, x_2, y_2 in finite dimensional space, if*

$$\left\| \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} - \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \right\| \leq \epsilon, \tag{27}$$

then

$$\|x_1 - x_2\|_2 \leq C\epsilon, \tag{28}$$

$$\|y_1 - y_2\|_2 \leq C\epsilon, \tag{29}$$

where $\|\cdot\|_2$ is the standard Euclidean norm.

Proof. From the equivalence of norms, for any norm $\|\cdot\|$ there exists $C > 0$ such that for all z in finite dimensional space we have $C\|z\| \geq \|z\|_2$. Hence, we have

$$C\epsilon \geq C\left\|\begin{pmatrix}x_1\\y_1\end{pmatrix} - \begin{pmatrix}x_2\\y_2\end{pmatrix}\right\| \geq \left\|\begin{pmatrix}x_1\\y_1\end{pmatrix} - \begin{pmatrix}x_2\\y_2\end{pmatrix}\right\|_2. \tag{30}$$

By the definition of the L_2 norm we have

$$\left\|\begin{pmatrix}x_1\\y_1\end{pmatrix} - \begin{pmatrix}x_2\\y_2\end{pmatrix}\right\|_2^2 = \|x_1 - x_2\|_2^2 + \|y_1 - y_2\|_2^2 \leq C^2\epsilon^2, \tag{31}$$

so

$$\|x_1 - x_2\|_2^2 \leq C^2\epsilon^2, \tag{32}$$
$$\|y_1 - y_2\|_2^2 \leq C^2\epsilon^2. \tag{33}$$

Taking square root on both sides concludes the proof. □

Theorem 4.1. *Consider any joint probability density function $p_{X,Y}$ of the random variables X and Y, such that the degradation is not invertible (according to Definition 3.1). For every $\gamma > 0$, there exist constants $m_1, m_2 > 0$ such that*

$$Lip(\hat{X}) \geq \frac{m_1}{\sqrt{W_p(p_{X,Y}, p_{\hat{X},Y})}} - m_2, \tag{3}$$

for any deterministic estimator $\hat{X} = f(Y)$ of X from Y with joint perceptual index $W_p(p_{X,Y}, p_{\hat{X},Y}) \leq \gamma$.

Proof. Let $\gamma \geq \epsilon$, suppose that $\hat{X}$ satisfies

$$W_p(p_{X,Y}, p_{\hat{X},Y}) = \inf_{p_{X_1,Y_1,X_2,Y_2}} \left\{ \left(\mathbb{E}\left[\left\|\begin{pmatrix}X_1\\Y_1\end{pmatrix} - \begin{pmatrix}X_2\\Y_2\end{pmatrix}\right\|^p\right]\right)^{\frac{1}{p}} : p_{X_1,Y_1} = p_{X,Y},\ p_{X_2,Y_2} = p_{\hat{X},Y} \right\} = \epsilon, \tag{34}$$

and let $p_{X_1^*,Y_1^*,X_2^*,Y_2^*}$ be the solution attaining the infimum in Equation (34). Denote

$$D = \left\|\begin{pmatrix}X_1^*\\Y_1^*\end{pmatrix} - \begin{pmatrix}X_2^*\\Y_2^*\end{pmatrix}\right\|^p, \tag{35}$$

so by the definition of D we have

$$\mathbb{E}[D] = \epsilon^p. \tag{36}$$

Since $D \geq 0$ almost surely, from Markov's inequality it holds that, for any $a > 0$,

$$P(D \geq a) \leq \frac{\mathbb{E}[D]}{a} = \frac{\epsilon^p}{a}. \tag{37}$$

Let us choose $a = t^p\epsilon^{0.5p}$ for some $t > 0$, so we have

$$P(D \geq t^p\epsilon^{0.5p}) \leq \frac{\epsilon^p}{t^p\epsilon^{0.5p}} = \frac{1}{t^p}\epsilon^{0.5p}. \tag{38}$$

The degradation is not invertible, so from Lemma C.2 the degradation is β-ill-posed for some $\beta > 0$, and from Lemma C.3 there exist $k > 0$ and a set $\mathcal{S}_y$ with $P(Y \in \mathcal{S}_y) > 0$ (denoted as $\tilde{\mathcal{S}}_y$ in Lemma C.3), where for every $y \in \mathcal{S}_y$ there exist sets $Q_1(y), Q_2(y)$ such that

$$P(X \in Q_1(y)|Y = y) \geq k, \tag{39}$$
$$P(X \in Q_2(y)|Y = y) \geq k, \tag{40}$$

and for every $(x^{(1)}, x^{(2)}) \in Q_1(y) \times Q_2(y)$ we have $\left\|x^{(1)} - x^{(2)}\right\| \geq \beta$.

Let

$$A_i = \left\{(x_1, y_1, x_2, y_2) \ : \ \left\|\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} - \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}\right\|^p < t^p \epsilon^{0.5p}, \ y_1 \in \mathcal{S}_y, \ x_1 \in Q_i(y_1), (x_2, y_2) \in \operatorname{supp} p_{\hat{X},Y}\right\}, \tag{41}$$

$$B_i = \{y_1 \ : \ (_, y_1, _, _) \in A_i\}. \tag{42}$$

We will show that for a large enough $t > 0$ we have

$$P(Y \in B_1 \cap B_2) > 0, \tag{43}$$

and this will hold true as long as $\epsilon \leq \gamma$.

By the definition of A_i and B_i we have

$$\begin{aligned}
P(X \in Q_i(Y), Y \in B_i) &= P(X_1^* \in Q_i(Y_1^*), Y_1^* \in B_i) && (44)\\
&\geq P((X_1^*, Y_1^*, X_2^*, Y_2^*) \in A_i) && (45)\\
&= P(D < t^p \epsilon^{0.5p}, X_1^* \in Q_i(Y_1^*), Y_1^* \in \mathcal{S}_y) && (46)\\
&= P(X_1^* \in Q_i(Y_1^*), Y_1^* \in \mathcal{S}_y) - P(D \geq t^p \epsilon^{0.5p}, X_1^* \in Q_i(Y_1^*), Y_1^* \in \mathcal{S}_y) && (47)\\
&= P(X \in Q_i(Y), Y \in \mathcal{S}_y) - P(D \geq t^p \epsilon^{0.5p}, X_1^* \in Q_i(Y_1^*), Y_1^* \in \mathcal{S}_y) && (48)\\
&\geq P(X \in Q_i(Y), Y \in \mathcal{S}_y) - P(D \geq t^p \epsilon^{0.5p}) && (49)\\
&\geq P(X \in Q_i(Y), Y \in \mathcal{S}_y) - \frac{1}{t^p}\epsilon^{0.5p}, && (50)
\end{aligned}$$

so $P(X \in Q_i(Y), Y \in B_i) \geq P(X \in Q_i(Y), Y \in \mathcal{S}_y) - \frac{1}{t^p}\epsilon^{0.5p}$. Since $B_i \subseteq \mathcal{S}_y$, by rearranging the inequality we have

$$\frac{1}{t^p}\epsilon^{0.5p} \geq P(X \in Q_i(Y), Y \in \mathcal{S}_y \setminus B_i), \tag{51}$$

so it holds that

$$\begin{aligned}
\frac{1}{t^p}\epsilon^{0.5p} &\geq P(X \in Q_i(Y), Y \in \mathcal{S}_y \setminus B_i) && (52)\\
&= \int_{\mathcal{S}_y \setminus B_i} p_Y(y) \int_{Q_i(y)} p_{X|Y}(x|y) dx dy && (53)\\
&= \int_{\mathcal{S}_y \setminus B_i} p_Y(y) P(X \in Q_i(y) | Y = y) dy && (54)\\
&\geq k \int_{\mathcal{S}_y \setminus B_i} p_Y(y) dy && (55)\\
&= k P(Y \in \mathcal{S}_y \setminus B_i) && (56)\\
\Rightarrow \frac{1}{kt^p}\gamma^{0.5p} &\geq \frac{1}{kt^p}\epsilon^{0.5p} \geq P(Y \in \mathcal{S}_y \setminus B_i). && (57)
\end{aligned}$$

Since $P(Y \in \mathcal{S}_y) > 0$, for a large enough $t > 0$ it holds that $\frac{1}{kt^p}\gamma^{0.5p} < \frac{1}{2}P(Y \in \mathcal{S}_y)$ (this choice does not depend on ϵ), and so

$$P(Y \in \mathcal{S}_y \setminus B_i) < \frac{1}{2}P(Y \in \mathcal{S}_y). \tag{58}$$

Therefore, we have

$$\begin{aligned}
P(Y \in B_i) &= P(Y \in \mathcal{S}_y) - P(Y \in \mathcal{S}_y \setminus B_i) && (59)\\
&> P(Y \in \mathcal{S}_y) - \frac{1}{2}P(Y \in \mathcal{S}_y) && (60)\\
&= \frac{1}{2}P(Y \in \mathcal{S}_y), && (61)
\end{aligned}$$

and finally we get

$$P(Y \in B_1 \cap B_2) = P(Y \in B_1) + P(Y \in B_2) - P(Y \in B_1 \cup B_2) \tag{62}$$
$$> 2 \cdot \frac{1}{2} P(Y \in \mathcal{S}_y) - P(Y \in \mathcal{S}_y) = 0, \tag{63}$$

so indeed $P(Y \in B_1 \cap B_2) > 0$.

Let us now construct a lower bound on the Lipschitz constant of $\hat{X}$. From the definition of B_i, for any $y_1 \in B_i$ there exists $x_1^{(i)} \in Q_i(y_1)$ and $(x_2^{(i)}, y_2^{(i)}) \in \operatorname{supp} p_{\hat{X},Y}$ such that $(x_1^{(i)}, y_1, x_2^{(i)}, y_2^{(i)}) \in A_i$. In particular, since $P(Y \in B_1 \cap B_2) > 0$ and $B_1 \cap B_2 \subseteq B_i$, we can take $y_1 \in B_1 \cap B_2$ so that $(x_1^{(i)}, y_1, x_2^{(i)}, y_2^{(i)}) \in A_i$ are both with the same y_1 for $i = 1, 2$. From the definition of A_i we have

$$\left\| \begin{pmatrix} x_1^{(i)} \\ y_1 \end{pmatrix} - \begin{pmatrix} x_2^{(i)} \\ y_2^{(i)} \end{pmatrix} \right\| \leq t\sqrt{\epsilon}, \tag{64}$$

and thus from Lemma C.4 we have

$$\left\| x_1^{(i)} - x_2^{(i)} \right\|_2 \leq Ct\sqrt{\epsilon}, \tag{65}$$
$$\left\| y_1 - y_2^{(i)} \right\|_2 \leq Ct\sqrt{\epsilon}. \tag{66}$$

Hence, from the triangle inequality we have

$$\left\| y_2^{(1)} - y_2^{(2)} \right\|_2 \leq \left\| y_2^{(1)} - y_1 \right\|_2 + \left\| y_1 - y_2^{(2)} \right\|_2 \tag{67}$$
$$\leq 2Ct\sqrt{\epsilon}, \tag{68}$$

and

$$\left\| x_1^{(1)} - x_1^{(2)} \right\|_2 \leq \left\| x_1^{(1)} - x_2^{(1)} \right\|_2 + \left\| x_2^{(1)} - x_1^{(2)} \right\|_2 \tag{69}$$
$$\leq \left\| x_1^{(1)} - x_2^{(1)} \right\|_2 + \left\| x_2^{(1)} - x_2^{(2)} \right\|_2 + \left\| x_2^{(2)} - x_1^{(2)} \right\|_2. \tag{70}$$

We are interested in $\left\| x_2^{(1)} - x_2^{(2)} \right\|_2$. Since $(x_1^{(1)}, x_1^{(2)}) \in Q_1(y_1) \times Q_2(y_1)$ we have

$$\left\| x_1^{(1)} - x_1^{(2)} \right\|_2 \geq \beta. \tag{71}$$

Thus, by rearranging Equation (70) we get

$$\left\| x_2^{(1)} - x_2^{(2)} \right\|_2 \geq \left\| x_1^{(1)} - x_1^{(2)} \right\|_2 - \left\| x_2^{(2)} - x_1^{(2)} \right\|_2 - \left\| x_1^{(1)} - x_2^{(1)} \right\|_2 \tag{72}$$
$$\geq \beta - 2Ct\sqrt{\epsilon}, \tag{73}$$

so

$$\frac{\left\| x_2^{(1)} - x_2^{(2)} \right\|_2}{\left\| y_2^{(1)} - y_2^{(2)} \right\|_2} \geq \frac{\beta - 2Ct\sqrt{\epsilon}}{2Ct\sqrt{\epsilon}}. \tag{74}$$

From the equivalence of norms in finite dimensional space, there exist constants $c_1, c_2 > 0$, such that for any

$x_2^{(1)}, x_2^{(2)}, y_2^{(1)}, y_2^{(2)}$ it holds that $\left\|x_2^{(1)} - x_2^{(2)}\right\| \geq c_1 \left\|x_2^{(1)} - x_2^{(2)}\right\|_2$ and $\left\|y_2^{(1)} - y_2^{(2)}\right\| \leq c_2 \left\|x_2^{(1)} - x_2^{(2)}\right\|_2$. Hence,

$$\frac{\left\|x_2^{(1)} - x_2^{(2)}\right\|}{\left\|y_2^{(1)} - y_2^{(2)}\right\|} \geq \frac{c_1 \left\|x_2^{(1)} - x_2^{(2)}\right\|_2}{c_2 \left\|y_2^{(1)} - y_2^{(2)}\right\|_2} \tag{75}$$

$$\geq \frac{c_1}{c_2} \frac{\beta - 2Ct\sqrt{\epsilon}}{2Ct\sqrt{\epsilon}} \tag{76}$$

$$\geq \frac{c_1}{c_2} \frac{\beta}{2Ct\sqrt{\epsilon}} - \frac{c_1}{c_2} \tag{77}$$

$$= \frac{m_1}{\sqrt{\epsilon}} - m_2, \tag{78}$$

where $m_1 = \frac{c_1}{c_2}\frac{\beta}{2Ct}$ and $m_2 = \frac{c_1}{c_2}$. To connect this result with the Lipschitz constant of $\hat{X}$, notice that from the definition of A_i we have $(x_2^{(i)}, y_2^{(i)}) \in \operatorname{supp} p_{\hat{X},Y}$, so $x_2^{(i)} = f(y_2^{(i)})$. Thus, from Equation (78) we have

$$\frac{\left\|f\left(y_2^{(1)}\right) - f\left(y_2^{(2)}\right)\right\|}{\left\|y_2^{(1)} - y_2^{(2)}\right\|} \geq \frac{m_1}{\sqrt{\epsilon}} - m_2. \tag{79}$$

To conclude, for every $0 < \epsilon \leq \gamma$, we have found two inputs $y_2^{(1)}, y_2^{(2)} \in \operatorname{supp} p_Y$ for which Equation (79) holds, i.e.,

$$\operatorname{Lip}(\hat{X}) \geq \frac{m_1}{\sqrt{\epsilon}} - m_2. \tag{80}$$

Note that $C > 0$ depends only on the norm $\|\cdot\|$, and is constant for any ϵ. Lastly, while this result holds for any $0 < \epsilon \leq \gamma$, one can always choose an arbitrarily large value of γ (and accordingly, an appropriate value of t) to obtain a lower bound of K for any range of values of $0 < \epsilon \leq \gamma$, but this lower bound becomes less tight for larger values of γ. □

D. Additional Details of Section 4.1 (Toy Example Illustration)

D.1. Full Training Details of the CGAN-Based Denoisers

Both the denoiser $\hat{X}_\lambda$ and the discriminator consist of 7 fully-connected layers with ReLU activations. For both the denoiser and the discriminator, the ReLU activation is applied on all layers except for the last one. All the layers have an input and output size of 16, except for the first and last layers. The input size of the last layer is 16 and the output size is 1 for both the denoiser and the discriminator. The input size of the first layer is 1 for the generator and 2 for the discriminator, and the output size is 16 for both. To clarify, as in (Mirza & Osindero, 2014), the discriminator is fed one time with (X, Y) ("real" samples) and one time with $(\hat{X}, Y)$ ("fake" samples) to compute the loss. The training set consists of $100,000$ independently drawn samples from $p_{X,Y}$, which is a Gaussian distribution with mean 0 and covariance $\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$. We train the networks with a non-saturating generative adversarial training loss (Goodfellow et al., 2014), and the discriminator is regularized using the R_1 gradient penalty (Mescheder et al., 2018) with coefficient 1.0. We optimize both of the networks using the Adam optimizer (Kingma & Ba, 2014) with $\beta_1 = 0.5$, $\beta_2 = 0.9$, a learning rate of 10^{-4}, a batch size of 128, and for a total of $100,000$ training steps for each network (we perform one training step at a time for each network). We multiply the learning rate by 0.5 every 5,000 training steps, starting after 50,000 steps, i.e., we perform multi-step learning rate scheduling.

D.2. Discussion and Ideas for Future Work

Consider the same toy example as in Section 4.1, with $N \sim \mathcal{N}(0, \sigma_N^2)$. It is well known that

$$\hat{X}_{\text{MMSE}} = \frac{1}{1 + \sigma_N^2} Y, \tag{81}$$

$$\hat{X}_p = \frac{1}{1 + \sigma_N^2} Y + W, \tag{82}$$

where $\hat{X}_{\text{MMSE}}$ attains the lowest possible Mean-Squared-Error (MSE), $\hat{X}_p$ is a sampler from the posterior distribution $p_{X|Y}$, and $W \sim \mathcal{N}(0, 1 - \frac{1}{1+\sigma_N^2})$ is independent of X, Y and N. Moreover, consider the estimator which attains the lowest possible MSE among all estimators with perfect perceptual quality, $\hat{X}_{D_{max}}$. In our specific example X and Y are jointly Gaussian, so from (Freirich et al., 2021; Blau & Michaeli, 2018) we have

$$\hat{X}_{D_{max}} = \frac{1}{\sqrt{1+\sigma_N^2}} Y. \tag{83}$$

Now, let us compare $\hat{X}_p$ and $\hat{X}_{D_{max}}$ with $\hat{X}_0$, the estimator $\hat{X}_\lambda$ from Section 4.1 with $\lambda = 0$, i.e., an estimator that is trained solely with a CGAN loss. In Figure 7 we show a contour plot of the density $p_{X,Y}$ ($p_{\hat{X}_p,Y} = p_{X,Y}$), as well as samples from $p_{\hat{X}_0,Y}$ and $p_{\hat{X}_{D_{max}},Y}$. The samples from $p_{\hat{X}_0,Y}$ and $p_{\hat{X}_{D_{max}},Y}$ are obtained by independently drawing samples $y \in p_Y$ and passing each y through $\hat{X}_0$ (the trained neural network) and $\hat{X}_{D_{max}}$ (Equation (83)). Observe that $\hat{X}_0$, a deterministic estimator that seeks to satisfy $p_{\hat{X},Y} \approx p_{X,Y}$, indeed behaves erratically, as anticipated by Theorem 4.1.

It appears, though, that $\hat{X}_{D_{max}}$ is a robust estimator. So why would one choose to use $\hat{X}_0$ over an estimator such as $\hat{X}_{D_{max}}$, which, amongst all the estimators with perfect perceptual quality, attains the lowest possible MSE? Namely, instead of using an erratic estimator which attains $p_{\hat{X},Y} \approx p_{X,Y}$, why not choose a robust estimator which attains very high perceptual quality and produces outputs that are sufficiently close to the source image? If such an estimator exists, we describe next two potential issues.

D.2.1. Conditional MSE

Let us evaluate the conditional MSE, defined by $\mathbb{E}[(X - \hat{X})^2|Y = y]$. It holds that

$$\mathbb{E}[(X - \hat{X}_{\text{MMSE}})^2|Y = y] = 1 - \frac{1}{1+\sigma_N^2}, \tag{84}$$

$$\mathbb{E}[(X - \hat{X}_p)^2|Y = y] = 2\left(1 - \frac{1}{1+\sigma_N^2}\right), \tag{85}$$

$$\mathbb{E}[(X - \hat{X}_{D_{max}})^2|Y = y] = 1 - \frac{1}{1+\sigma_N^2} + \left(\frac{\sqrt{1+\sigma_N^2} - 1}{1+\sigma_N^2}\right)^2 y^2, \tag{86}$$

(see visual illustration in Figure 8 and proof below). Notice that both the conditional MSE of $\hat{X}_{\text{MMSE}}$ and $\hat{X}_p$ are constant in y, while that of $\hat{X}_{D_{max}}$ is proportional to y^2 and bounded from below by the conditional MSE of $\hat{X}_{\text{MMSE}}$. We find it interesting that even though $\hat{X}_{D_{max}}$ attains a lower MSE compared to $\hat{X}_p$ when averaging over all y's, it produces worse conditional MSE when the absolute value of y is large. Hence, $\hat{X}_{D_{max}}$ does not simply attain better MSE "for free", as it sacrifices the conditional MSE originating from less likely values of y to serve better more likely values of y. Perhaps this phenomenon can be extended to more general settings, but we leave this for future work.

To see why the expressions above hold, first note that

$$\mathbb{E}[(X - \hat{X}_{\text{MMSE}})^2|Y = y] = Var(X|Y = y) = 1 - \frac{1}{1+\sigma_N^2} \tag{87}$$

is a well known result. Now, for any estimator $\hat{X}$ we have

$$\mathbb{E}[(X - \hat{X})^2|Y = y] = \mathbb{E}[X^2|Y = y] + \mathbb{E}[\hat{X}^2|Y = y] - 2\mathbb{E}[X\hat{X}|Y = y] \tag{88}$$

$$= \mathbb{E}[X^2|Y = y] + \mathbb{E}[\hat{X}^2|Y = y] - 2\mathbb{E}[X|Y = y]\mathbb{E}[\hat{X}|Y = y], \tag{89}$$

where Equation (89) is true since $\hat{X}$ and X are independent given Y. Hence, we have

$$\mathbb{E}[(X - \hat{X}_p)^2|Y = y] = \mathbb{E}[X^2|Y = y] + \mathbb{E}[\hat{X}_p^2|Y = y] - 2\mathbb{E}[X|Y = y]\mathbb{E}[\hat{X}_p|Y = y] \tag{90}$$

$$= 2\mathbb{E}[X^2|Y = y] - 2\left(\mathbb{E}[X|Y = y]\right)^2 \tag{91}$$

$$= 2Var(X|Y = y) \tag{92}$$

$$= 2\left(1 - \frac{1}{1+\sigma_N^2}\right). \tag{93}$$

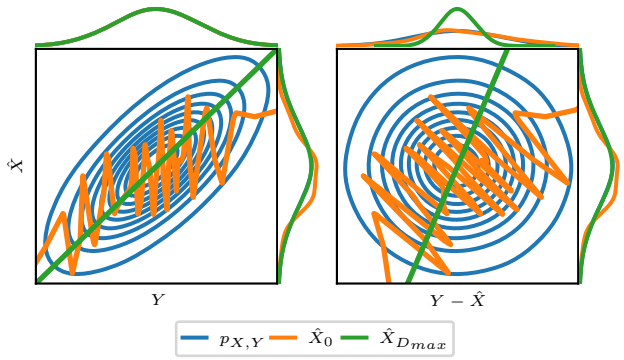

Figure 7: A visual illustration of $\hat{X}_0$ and $\hat{X}_{D_{max}}$ from the toy example in Appendix D.2 (with $\sigma_N = 1$). **Left**: Contour plot of the density $p_{X,Y}$ (blue concentric ellipses), outputs from $\hat{X}_0$ (orange "zigzag" line), and outputs from $\hat{X}_{D_{max}}$ (green straight line). **Right**: the same as the left plot, but instead of Y we consider $Y - \hat{X}$. We also present the marginal density functions of each axis, i.e., the densities $p_{\hat{X}}, p_Y$ and $p_{Y-\hat{X}}$. These marginal densities are computed using KDE.

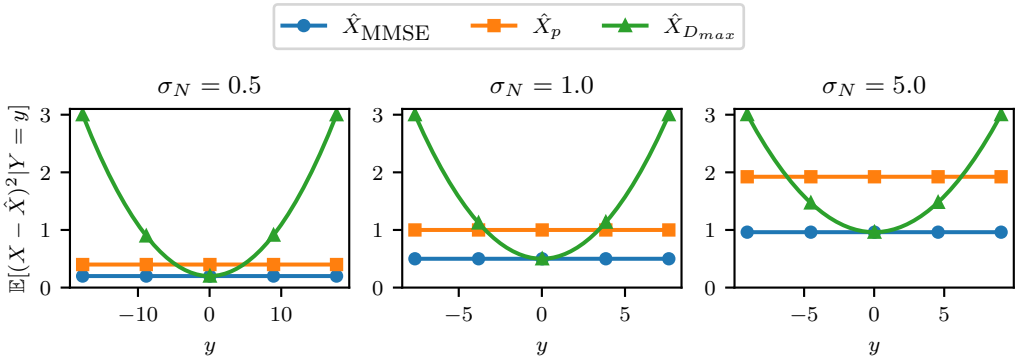

Figure 8: The conditional MSE $\mathbb{E}[(X - \hat{X})|Y = y]$ versus y obtained by each of the estimators $\hat{X}_{\text{MMSE}}$, $\hat{X}_p$ and $\hat{X}_{D_{max}}$ which are described in Appendix D.2. Each subplot corresponds to a different value of σ_N (mentioned in the title). The y values in each subplot are taken such that the conditional MSE is below or equal to 3.

We are left with obtaining $\mathbb{E}[(X-\hat{X}_{D_{max}})^2|Y=y]$. It holds that

$$\mathbb{E}[X^2|Y=y] = Var(X|Y=y) + (\mathbb{E}[X|Y=y])^2 = 1 - \frac{1}{1+\sigma_N^2} + \frac{1}{(1+\sigma_N^2)^2}y^2. \tag{94}$$

Moreover, for $\hat{X}_{D_{max}}$ we have

$$\mathbb{E}[\hat{X}_{D_{max}}|Y=y] = \mathbb{E}\left[\frac{1}{\sqrt{1+\sigma_N^2}}Y\middle|Y=y\right] = \frac{1}{\sqrt{1+\sigma_N^2}}y, \tag{95}$$

$$\mathbb{E}[\hat{X}^2_{D_{max}}|Y=y] = \mathbb{E}\left[\frac{1}{1+\sigma_N^2}Y^2\middle|Y=y\right] = \frac{1}{1+\sigma_N^2}y^2, \tag{96}$$

and thus,

$$\mathbb{E}[(X-\hat{X}_{D_{max}})^2|Y=y] = 1 - \frac{1}{1+\sigma_N^2} + \frac{1}{(1+\sigma_N^2)^2}y^2 + \frac{1}{1+\sigma_N^2}y^2 - \frac{2}{(1+\sigma_N^2)\sqrt{1+\sigma_N^2}}y^2 \tag{97}$$

$$= 1 - \frac{1}{1+\sigma_N^2} + \left(\frac{1+(1+\sigma_N^2)-2(1+\sigma_N^2)^{0.5}}{(1+\sigma_N^2)^2}\right)y^2 \tag{98}$$

$$= 1 - \frac{1}{1+\sigma_N^2} + \left(\frac{\sqrt{1+\sigma_N^2}-1}{1+\sigma_N^2}\right)^2 y^2. \tag{99}$$

D.2.2. Consistency in the additive noise setting

Figure 7 provides a visual intuition for another apparent issue with $\hat{X}_{D_{max}}$, which clarifies the distinction between perceptual quality and joint perceptual quality. While $\hat{X}_{D_{max}}$ attains perfect perceptual quality, observe that it produces each source $x \in \operatorname{supp} p_X$ for only one particular measurement $y \in \operatorname{supp} p_Y$, i.e., it is an invertible function. Is it practically acceptable that an estimator could produce a given natural image only for one particular noisy measurement? The issue with this phenomenon is clearer when looking at the right plot of $\hat{X}$ versus $Y-\hat{X}$. There, we see that $\hat{X}_{D_{max}}$ associates each output $\hat{X}$ with one possible residual noise $\hat{N}_{D_{max}} = Y - \hat{X}_{D_{max}}$, so $\hat{N}_{D_{max}}$ and $\hat{X}_{D_{max}}$ are statistically dependent, whereas N and X are not. Moreover, from the Kernel Density Estimation (KDE) of the marginal distributions in Figure 7, we also see that the distribution of $\hat{N}_{D_{max}}$ is quite different from that N. These phenomena reveal that $\hat{X}_{D_{max}}$ does not fully "separate" X from N, leaving irrelevant portions from the noise in the estimated output.

Let us elaborate on this phenomenon for general distributions of X and N. Suppose that $Y = X + N$, where N is some random noise independent of X. Let $\hat{X}$ be an image denoiser, and let $\hat{N} = Y - \hat{X}$ be the residual noise. A common, generally desired property of $\hat{X}$ is that $\hat{N}$ would follow the same distribution as N (e.g., isotropic Gaussian) and be statistically uncorrelated with $\hat{X}$. It turns out that $p_{\hat{X},Y} = p_{X,Y}$ if and only if $p_{\hat{X}} = p_X$ and $\forall(x,n)\ :\ p_{\hat{N}|\hat{X}}(n|x) = p_N(n)$ (which we write as $p_{\hat{N}|\hat{X}} = p_N$ in short). Namely, $\hat{X}$ is a posterior sampler if and only if it attains perfect perceptual quality, and $\hat{N}$ is statistically *independent* of $\hat{X}$ and follows the same distribution as that of N. To show that this is true, we simply show that $p_{Y|\hat{X}} = p_{Y|X}$ if and only if $p_{\hat{N}|\hat{X}} = p_N$. Assume first that $p_{Y|\hat{X}} = p_{Y|X}$. We have

$$p_{\hat{N}|\hat{X}}(n|x) = \int p_{\hat{N},Y|\hat{X}}(n,y|x)dy \tag{100}$$

$$= \int p_{\hat{N}|Y,\hat{X}}(n|y,x)p_{Y|\hat{X}}(y|x)dy \tag{101}$$

$$= \int \delta(n-(y-x))p_{Y|\hat{X}}(y|x)dy \tag{102}$$

$$= \int \delta(n-(y-x))p_{Y|X}(y|x)dy \tag{103}$$

$$= p_{Y|X}(n+x|x) \tag{104}$$

$$= p_N(n). \tag{105}$$

From the other direction, assume that $p_{\hat{N}|\hat{X}} = p_N$. In that case we have

$$p_{Y|\hat{X}}(y|x) = \int p_{Y,\hat{N}|\hat{X}}(y,n|x)dn \tag{106}$$

$$= \int p_{Y|\hat{X},\hat{N}}(y|x,n)p_{\hat{N}|\hat{X}}(n|x)dn \tag{107}$$

$$= \int \delta(y-(x+n))p_N(n)dn \tag{108}$$

$$= p_N(y-x) \tag{109}$$

$$= p_{Y|X}(y|x). \tag{110}$$

This is quite an interesting result. First, it gives a comprehensible meaning to the condition $p_{Y|\hat{X}} = p_{Y|X}$ in the case of additive noise degradations, a meaning which is already discussed in the literature. Second, when $\hat{X}$ attains high perceptual quality, and if one desires that the estimator would (almost) perfectly separate the signal from the noise in the sense that $\hat{N}$ is independent of $\hat{X}$ and $p_{\hat{N}} = p_N$, then $\hat{X}$ must be close to be a posterior sampler. Hence, when $\hat{X}$ is a deterministic estimator, to approach perfect perceptual quality and perfect noise separation, $\hat{X}$ must possess a high Lipschitz constant (Theorem 4.1). This also means that, whenever $\hat{X}_{D_{max}}$ is a robust deterministic estimator, it cannot possibly lead to near-perfect noise separation, and therefore cannot lead to near-perfect consistency; such an estimator cannot satisfy $p_{\hat{X}_{D_{max}},Y} \approx p_{X,Y}$.

E. Additional Details of Section 4.2 (Quantitative Demonstration)

All our experiments rely on the official codes and checkpoints of the evaluated methods.

E.1. Complementary Plots of $\overline{K}$ Versus $\sqrt{\text{JFID}}$ with All the Evaluated Methods

In Figure 9 we provide the same plots as in Figure 3 ($\overline{K}$ versus $\sqrt{\text{JFID}}$) but this time include all the algorithms that are evaluated in Section 4.2 which we omitted from Figure 3.

E.2. Plots of Additional Joint Perceptual Quality Measures

In Figures 11, 13 and 15 we provide the same plots as in Figure 9, but instead of using FID to compute statistical discrepancy, we use KID (Binkowski et al., 2018), Precision (Kynkäänniemi et al., 2019), and Recall (Kynkäänniemi et al., 2019). Similarly to JFID, we denote by Joint-KID (JKID), Joint-Precision (JP), and Joint-Recall (JR), the KID, precision, and recall (respectively) between the sets

$$\left\{ \begin{pmatrix} x_F^{(i)} \\ y_F^{(i)} \end{pmatrix} \right\}_{i=1}^N \text{ and } \left\{ \begin{pmatrix} \hat{x}_F^{(i)} \\ y_F^{(i)} \end{pmatrix} \right\}_{i=1}^N . \tag{111}$$

Moreover, similar to the computation of the perceptual quality via FID, we compute the KID, Precision (P), and Recall (R) between the sets $\{x_F^{(i)}\}_{i=1}^N$ and $\{\hat{x}_F^{(i)}\}_{i=1}^N$. Interestingly, we observe an empirical tradeoff between $\overline{K}$ and each of these additional joint perceptual quality measures as well, i.e., larger values of $\overline{K}$ correspond to lower values of $\sqrt{\text{JKID}}$ (lower is better), higher values of JP (higher is better), and higher values of JR (higher is better). These results show that the tradeoff between $\overline{K}$ and $\sqrt{\text{JFID}}$ is not merely a byproduct of using FID as a statistical discrepancy measure, since we observe the tradeoff when using other types of statistical discrepancy measures as well.

E.3. Assessing the Effectiveness of the Joint Perceptual Quality Measures

We question whether our approach to measure joint perceptual quality is predominantly affected by perceptual quality or is it also influenced by consistency. In Figures 10, 12, 14 and 16 we plot $\sqrt{\text{FID}}$, $\sqrt{\text{KID}}$, P and R versus $\sqrt{\text{JFID-FID}}$, $\sqrt{\text{KID-JKID}}$, JP $-$ P and JR $-$ R, respectively. Observe that a particular value of $\sqrt{\text{FID}}$ may correspond to different values of $\sqrt{\text{JFID-FID}}$. This means that JFID is not simply a scaled version of the FID, i.e., two different algorithms with the same FID may attain a different value of JFID. This shows that JFID is affected by consistency and not only by perceptual quality.

The same result holds for the JP and JR as well, but it does not hold for the JKID. Indeed, we observe a linear relationship between $\sqrt{\text{KID}}$ and $\sqrt{\text{KID-JKID}}$, i.e., the observed empirical tradeoff between $\overline{K}$ and $\sqrt{\text{JKID}}$ is actually a tradeoff between

$\overline{K}$ and $\sqrt{\text{KID}}$. Thorough analysis of these results and the search for better ways to measure joint perceptual quality are left for future work.

E.4. Perception-Distortion Tradeoff

As a complementary, we provide in Figures 9, 11, 13 and 15 plots of the perception-distortion tradeoff (Blau & Michaeli, 2018) (e.g., $\sqrt{\text{FID}}$ or $\sqrt{\text{KID}}$ versus the RMSE), where each statistical distance (FID, KID, etc.) is computed between the sets $\{x_F^{(i)}\}_{i=1}^N$ and $\{\hat{x}_F^{(i)}\}_{i=1}^N$. Namely, each statistical distance measures the perceptual quality of each algorithm. Furthermore, in Figures 17 to 20 we also plot the *joint* perceptual quality (computed via JFID, JKID, etc.) versus the RMSE. Due to the perception-distortion tradeoff, we anticipate a tradeoff between the joint perceptual quality and distortion as well, which we can indeed observe in these figures.

F. Additional Details of Section 4.3 (Adversarial Attacks)

We train the RRDB model (Wang et al., 2018c) using the official model's code provided by the authors, deployed into the official training code of GFPGAN (Wang et al., 2021b). We omit all the losses of GFPGAN except for the L_1 reconstruction loss, and use the same training data (FFHQ (Karras et al., 2019) face images resized to 512×512) and degradation model as that of GFPGAN v1. To traing the model, we use the Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.99$ and a learning rate of $2 \cdot 10^{-4}$, and perform the same learning rate scheduling as in the official training configuration of RRDB. The model is trained with a batch side of 16 for 1,000,000 iterations. We record a model checkpoint every 5000 steps, and for evaluation we use the checkpoint that attains the highest PSNR (which is 24.7dB).

F.1. Adversarial Attacks to Produce "Female"-Classified Faces

In Figure 21 and Figure 22 we present more output examples of the gender adversarial attacks performed on GFPGAN and the trained RRDB model, respectively. Note that the outputs of the RRDB model are blurry, as expected from a restoration model with high PSNR, so one might question whether the face gender prediction models we use in Section 4.3 (for attacking the model and for evaluation) can even properly handle such images. First, notice that the probability of the "female" class is almost identical for the source images (61.2%) and the images reconstructed by the RRDB model (59.1%). To verify that this is not a coincidence, we compute the fraction of restored images whose class, as determined by the gender prediction model, is the same class as that of the corresponding source images. We discover that the result is above 90% for both of the face gender prediction models we use. Hence, according to the face gender classifiers, the RRDB model almost always preserves the true gender of the source image.

Moreover, note that the fraction of the "female" class for the attacked RRDB output images is *lower* than that of the original restored images. This means that some of the output images produced by RRDB are originally classified as "female", but their class is switched to "male" after performing the adversarial attack, even though the attack is intended to make the opposite effect. In Figure 22 we show several examples where the original class is "female" and the class after the attack is "male", and vice versa. Subjectively, we do not see any bold qualitative reason for such an unexpected outcome. It is perhaps related to the sensitivity of classification models on examples where the class probabilities are almost identical, e.g., when the probabilities of both classes (in a binary classification problem) are around 0.5. We leave the study of this phenomenon for future work.

F.2. Adversarial Attacks to Produce "Old"-Classified Faces

Like in the gender adversarial attacks, we perform similar attacks to alter the outputs of GFPGAN and RRDB to produce faces of older people. To do so, we use the same adversarial attack procedure as described in Section 4.3, but instead of maximizing the log-probability of the class "female", we maximize the log-probability of the class that corresponds to the oldest age category in a facial age classification model. Specifically, we maximize the log-probability of the 70+ age category as predicted by the model in (Nate Raw, 2023) (this model was trained on the FairFace data set (Karkkainen & Joo, 2021)). Like in Section 4.3, we use here $\alpha = 16/255$ and $T = 100$ in the I-FGSM update rule as well. We use the age prediction model in (Serengil & Ozpinar, 2021) to evaluate the both GFPGAN and RRDB, which predicts the age as a scalar.

We present visual results of both GFPGAN and RRDB in Figure 23. The input and output images we choose to present for each model correspond to the 12 attacked output images $f(y_{adv})$ with the oldest predicted age according to the evaluation model. The attacked outputs of GFPGAN contains features that may be associated with older people (e.g., wrinkles). It

seems that RRDB attempts to produce the same effect but is unsuccessful due to its lower perceptual quality, and perhaps its lower sensitivity to such adversarial attacks. In Table 1 we confirm these results quantitatively.

	Source images	GFPGAN	Attacked GFPGAN	RRDB	Attacked RRDB
Predicted age average	32.7	32.3	38.8	29.98	33.87
Predicted age standard deviation	6.6	5.9	8.9	4	5.4
Maximum predicted age	60	62	66	48	53

Table 1: Quantitative analysis of the adversarial attacks described in Appendix F.2, intended to alter GFPGAN and RRDB to produce faces of older people. The average, standard deviation, and maximum are taken over all the 1000 test images chosen from CelebA-HQ. The columns denoted by GFPGAN and Attacked GFPGAN correspond to the original outputs of GFPGAN (originating from the original degraded inputs), and the attacked outputs of GFPGAN (originating from the attacked inputs), respectively. The same notation holds for RRDB and Attacked RRDB.

G. Additional Details of Section 4.4 (Exploring the Posterior Distribution)

G.1. Why Should We Explore the Posterior Distribution?

Exploring the posterior distribution may hold a practical significance in several scenarios. First and foremost, it enables taking into account the uncertainty inherent to the restoration task at hand. For example, in CT-reconstruction from partial measurements, radiologists may make better-informed decisions if they have a clear understanding of the uncertainty of the images obtained. Having access to the posterior distribution, one can either quantify this uncertainty or enable visualizing it by traversing through different samples from the posterior. This would allow medical practitioners to assess the reliability of their diagnoses and potentially avoid misinterpretations or misdiagnoses. See (Belhasin et al., 2024) for more details about such possibilities. More broadly, when there is a variety of possible solutions to our inverse problem, exploring the posterior gives the ability to the user to view the possible outcomes and choose the more favored solution.

G.2. Supplementary Details on the Experiments

Algorithm 1 is the formal description of the FPS approach we use to explore the posterior distribution. We perform $T = 150$ I-FGSM update steps to obtain each sample, and choose $\alpha = 8/255$ to be the scale of the adversarial attack. With such a value of α, the visual difference between y and each attacked input is negligible (see, e.g., Figure 5, where we use $\alpha = 16/255$). The images we present in Figure 6 are the $S = 5$ samples resulting from the algorithm. Notice that the first sample resulting from the algorithm is simply the output originating from the original input y, without performing any attack.

Algorithm 1 Farthest Point Sampling approach to explore the posterior distribution with a deterministic estimator $\hat{X} = f(Y)$ that attains high joint perceptual quality.

Input: Degraded measurement y, number of output samples S, adversarial attack scale α, number of adversarial attack update steps T.
Output: List of samples $\mathcal{X} = [\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_S]$.
$\hat{x} \leftarrow f(y)$ {First output of the estimator.}
$\mathcal{X} \leftarrow [\hat{x}]$
for $i = 1$ **to** S **do**
  $y_{adv} \leftarrow y$
  for $t = 1$ **to** T **do**
    $\hat{x}_t \leftarrow f(y_{adv})$
    $\mathcal{L} = \frac{1}{i} \sum_{s=1}^{i} \|\mathcal{X}[s] - \hat{x}_t\|_2^2$
    $y_{adv} \leftarrow$ I-FGSM$(\mathcal{L}, \alpha, y)$ {The basic I-FGSM update rule in (Choi et al., 2019), but with our loss $\mathcal{L}$.}
  end for
end for
$\mathcal{X} \leftarrow \mathcal{X} \cup \{f(y_{adv})\}$ {Append the new output to the list of output samples.}

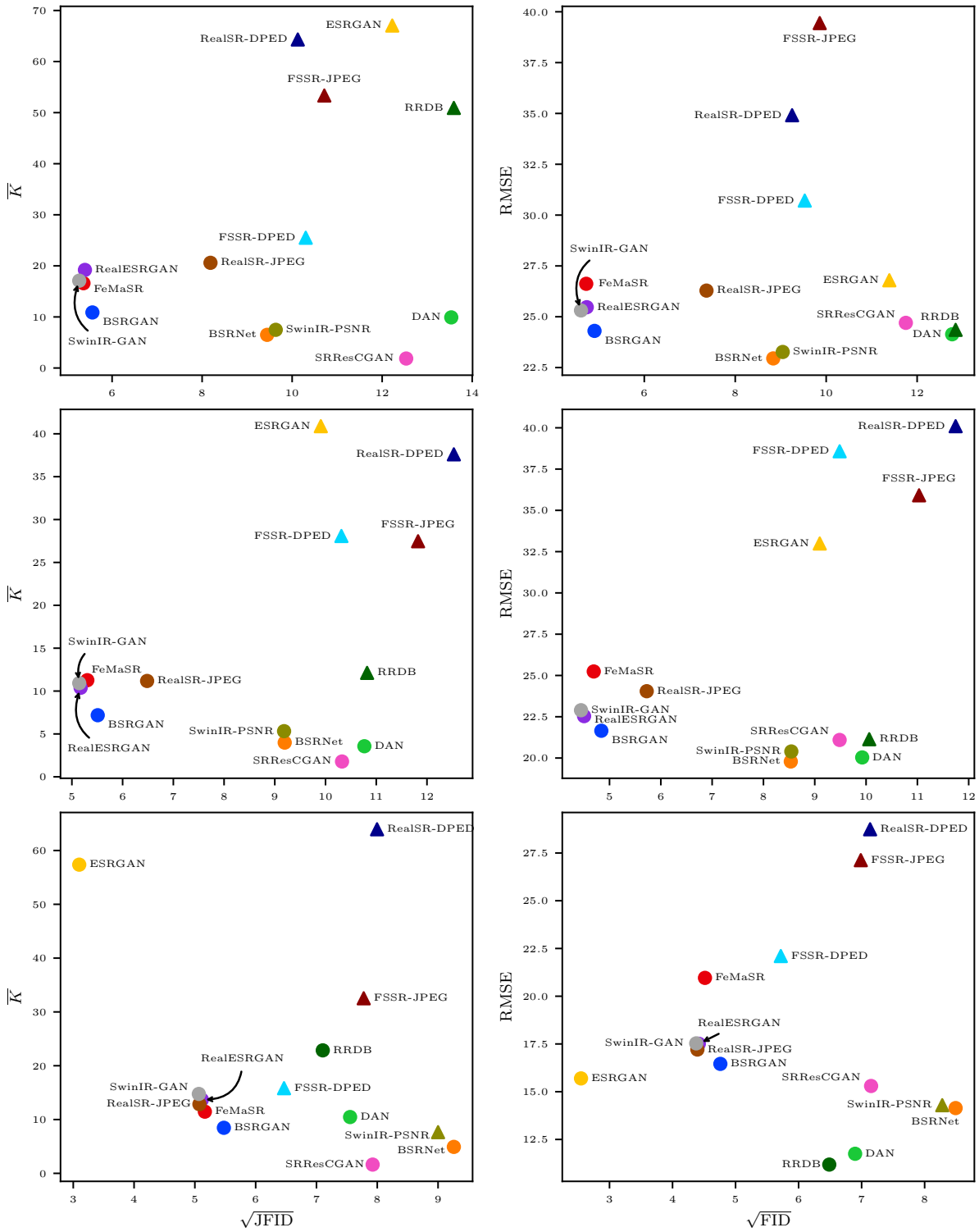

Figure 9: Plots of $\overline{K}$ versus $\sqrt{\text{JFID}}$ and RMSE versus $\sqrt{\text{FID}}$ (distortion-perception plane) of the SISR algorithms evaluated in Section 4.2. These plots are complementary and contain the algorithms which are omitted from Figure 3, which are marked here as triangles. The top figures correspond to the degradation from the Track2 challenge in (Lugmayr et al., 2019), the center figures correspond to the degradation from the Track1 challenge in (Lugmayr et al., 2020), and the bottom figures correspond to the common bicubic degradation.

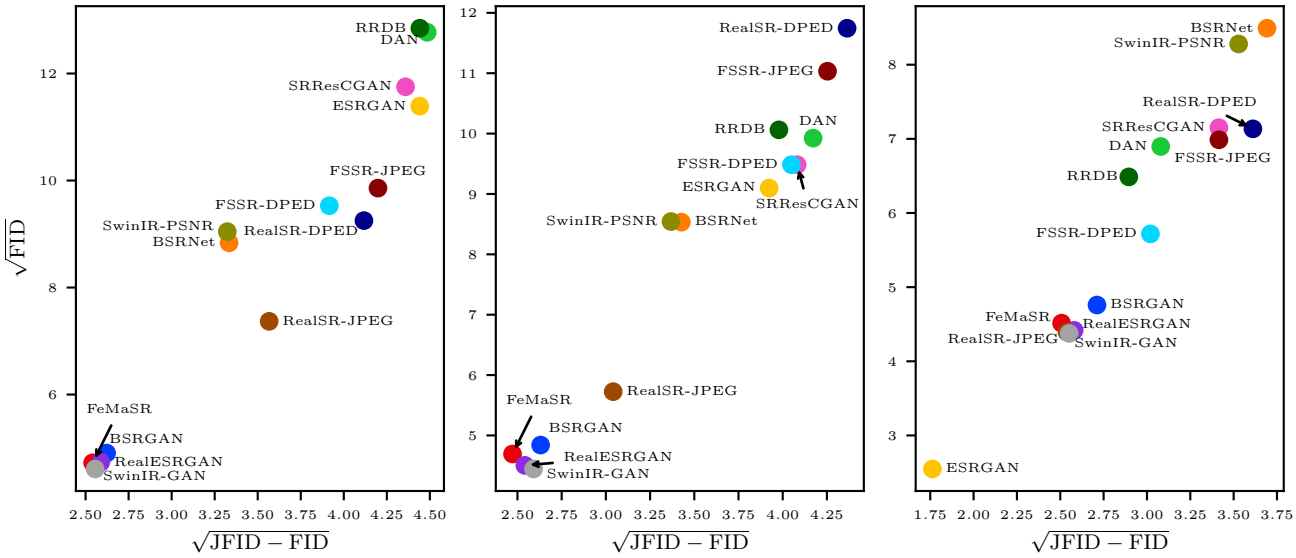

Figure 10: Plots of $\sqrt{\text{FID}}$ versus $\sqrt{\text{JFID-FID}}$. The left plot corresponds to the degradation from the Track2 challenge in (Lugmayr et al., 2019), the center plot corresponds to the degradation from the Track1 challenge in (Lugmayr et al., 2020), and the right plot corresponds to the common bicubic degradation. We see that two algorithms with a similar FID value (similar perceptual quality) may attain different JFID values (different joint perceptual quality). This shows that JFID is not only influenced by perceptual quality, but also by consistency.

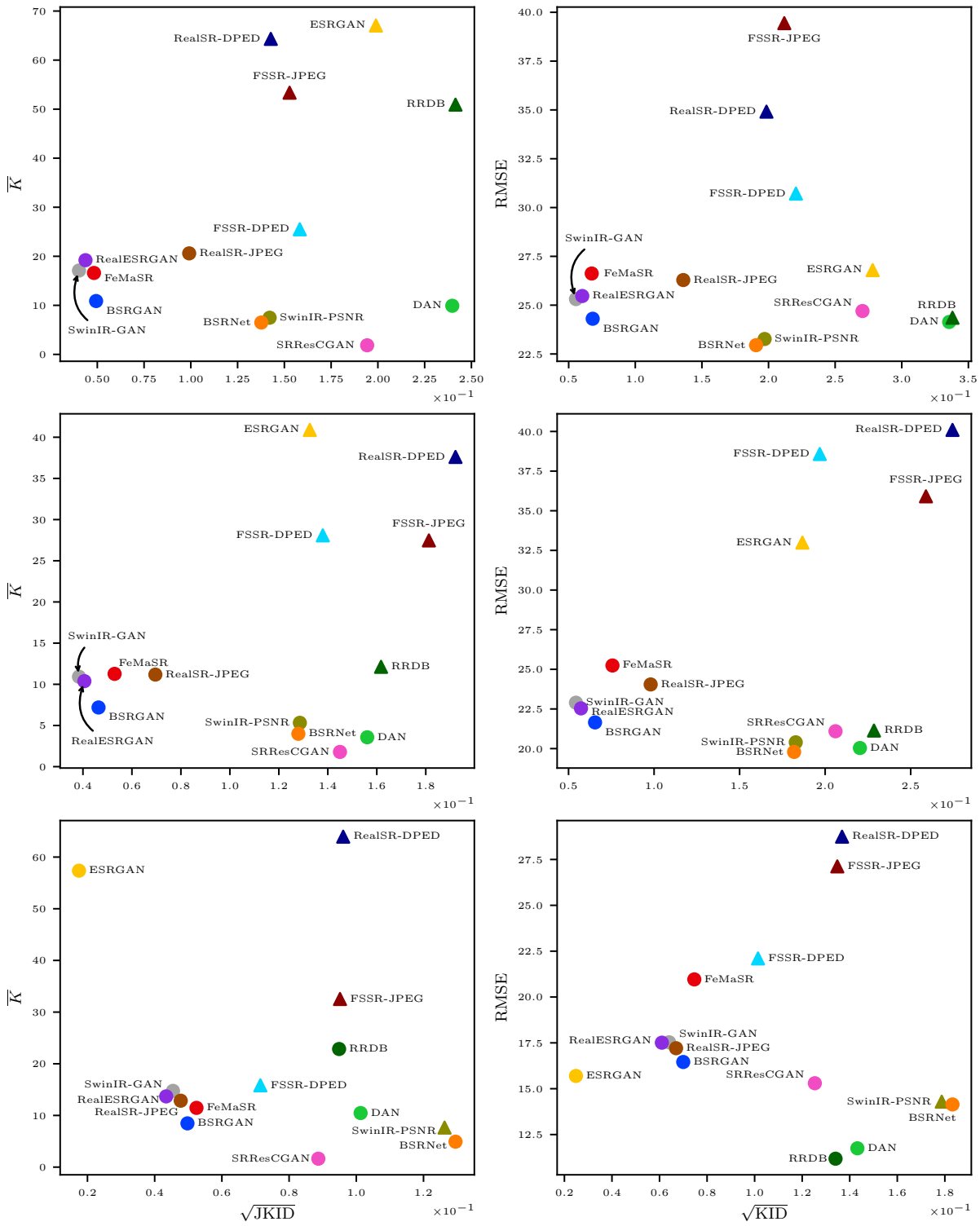

Figure 11: Plots of $\overline{K}$ versus $\sqrt{\text{JKID}}$ and RMSE versus $\sqrt{\text{KID}}$ (distortion-perception plane) of the SISR algorithms evaluated in Section 4.2. The top figures correspond to the degradation from the Track2 challenge in (Lugmayr et al., 2019), the center figures correspond to the degradation from the Track1 challenge in (Lugmayr et al., 2020), and the bottom figures correspond to the common bicubic degradation.

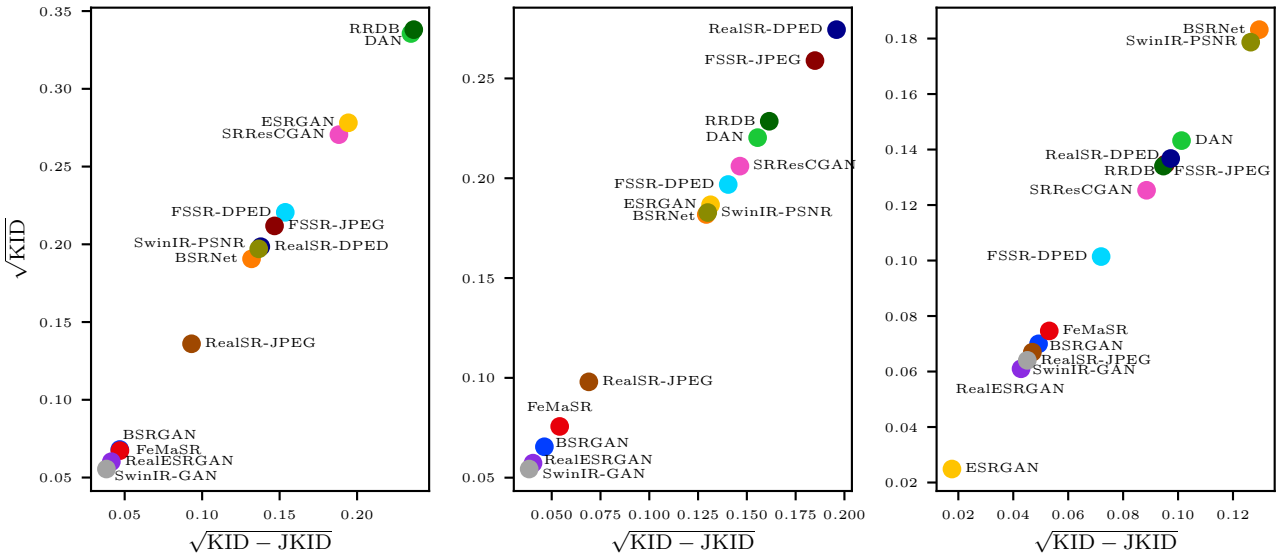

Figure 12: Plots of $\sqrt{\text{KID}}$ versus $\sqrt{\text{KID-JKID}}$. The left plot corresponds to the degradation from the Track2 challenge in (Lugmayr et al., 2019), the center plot corresponds to the degradation from the Track1 challenge in (Lugmayr et al., 2020), and the right plot corresponds to the common bicubic degradation. We observe a linear relationship between $\sqrt{\text{KID}}$ and $\sqrt{\text{KID-JKID}}$, so JKID is influenced solely by perceptual quality and completely ignores consistency. Thus, JKID is an ineffective measure of the joint perceptual quality.

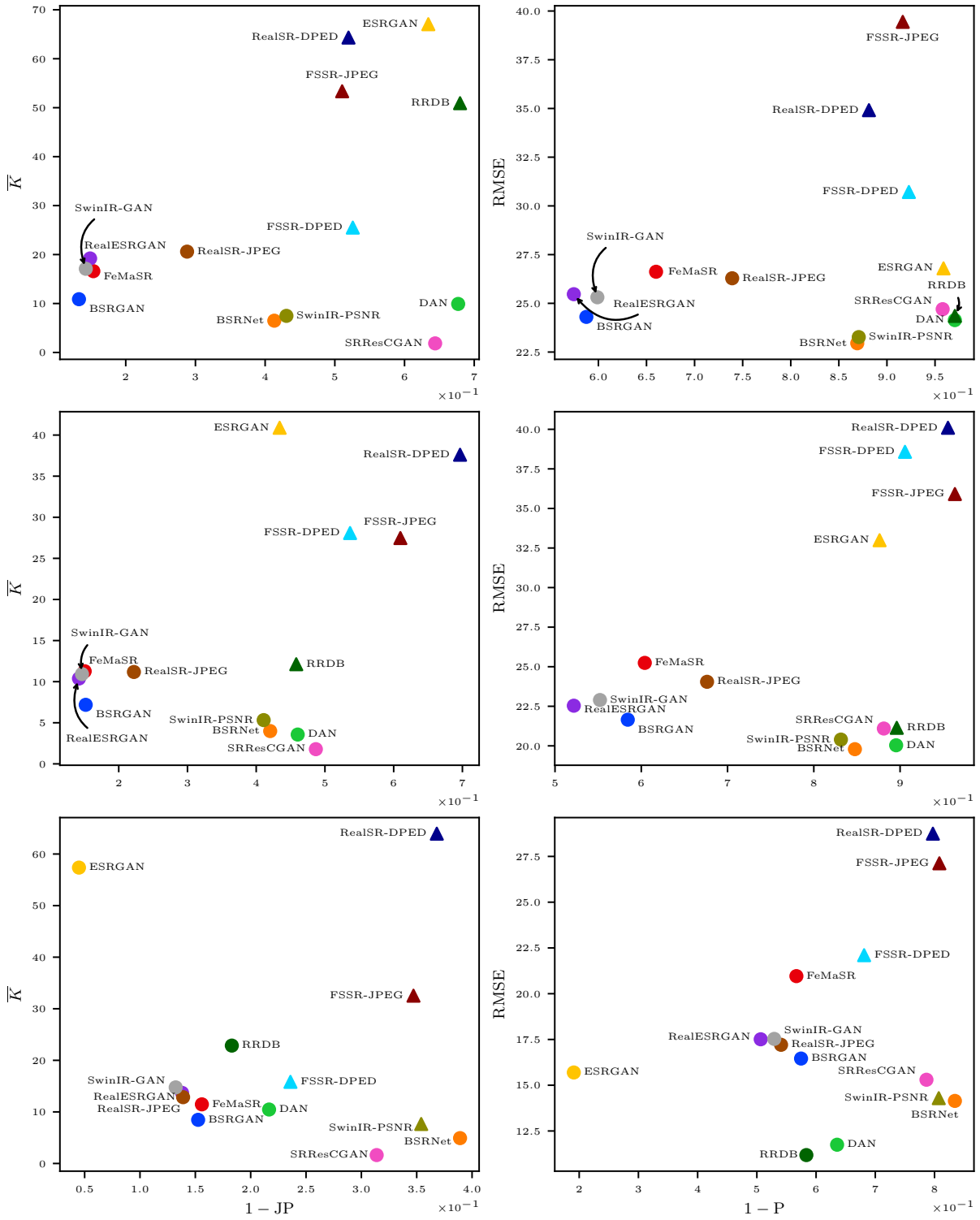

Figure 13: Plots of $\overline{K}$ versus $1 - \text{JP}$ and RMSE versus $1 - \text{P}$ (distortion-perception plane) of the SISR algorithms evaluated in Section 4.2. The top figures correspond to the degradation from the Track2 challenge in (Lugmayr et al., 2019), the center figures correspond to the degradation from the Track1 challenge in (Lugmayr et al., 2020), and the bottom figures correspond to the common bicubic degradation.

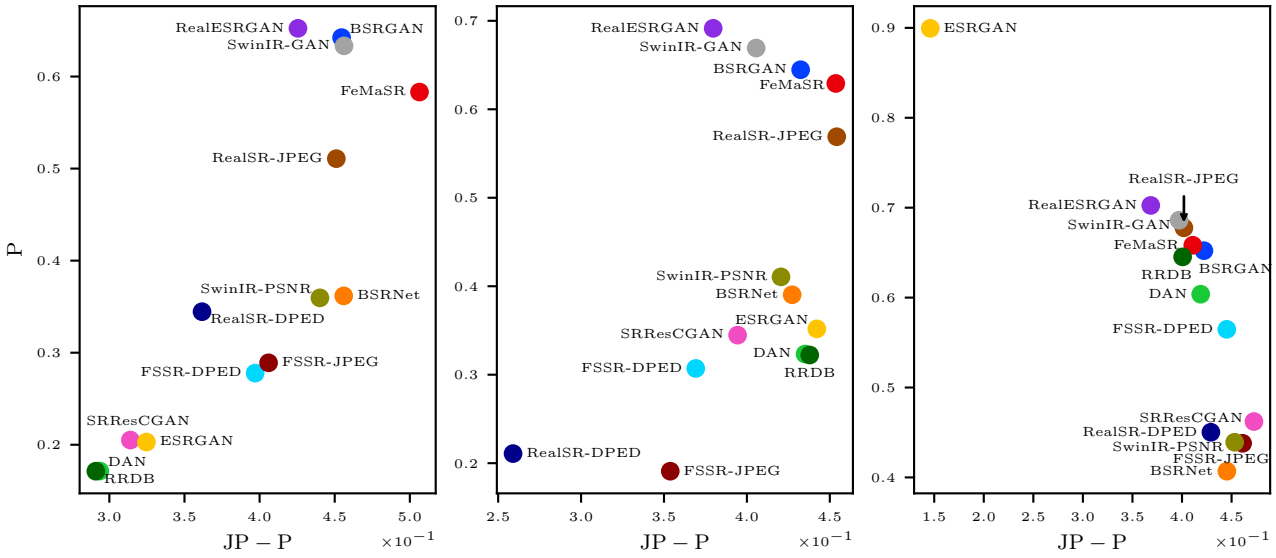

Figure 14: Plots of P versus JP − P. The left plot corresponds to the degradation from the Track2 challenge in (Lugmayr et al., 2019), the center plot corresponds to the degradation from the Track1 challenge in (Lugmayr et al., 2020), and the right plot corresponds to the common bicubic degradation. We see that two algorithms with a similar P value may attain different JP values. This shows that JP is not only influenced by perceptual quality, but also by consistency.

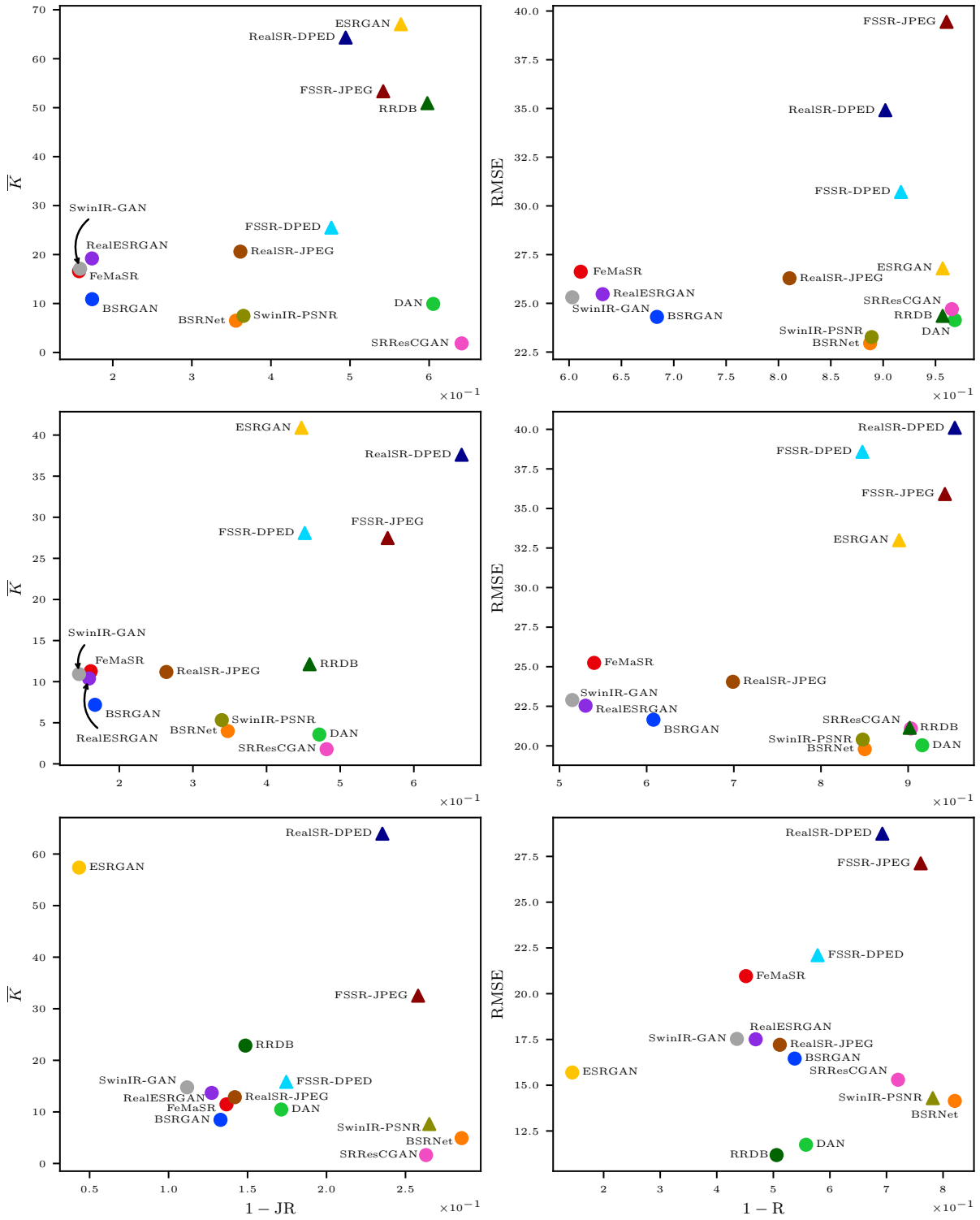

Figure 15: Plots of $\overline{K}$ versus $1 - \text{JR}$ and RMSE versus $1 - \text{R}$ (distortion-perception plane) of the SISR algorithms evaluated in Section 4.2. The top figures correspond to the degradation from the Track2 challenge in (Lugmayr et al., 2019), the center figures correspond to the degradation from the Track1 challenge in (Lugmayr et al., 2020), and the bottom figures correspond to the common bicubic degradation.

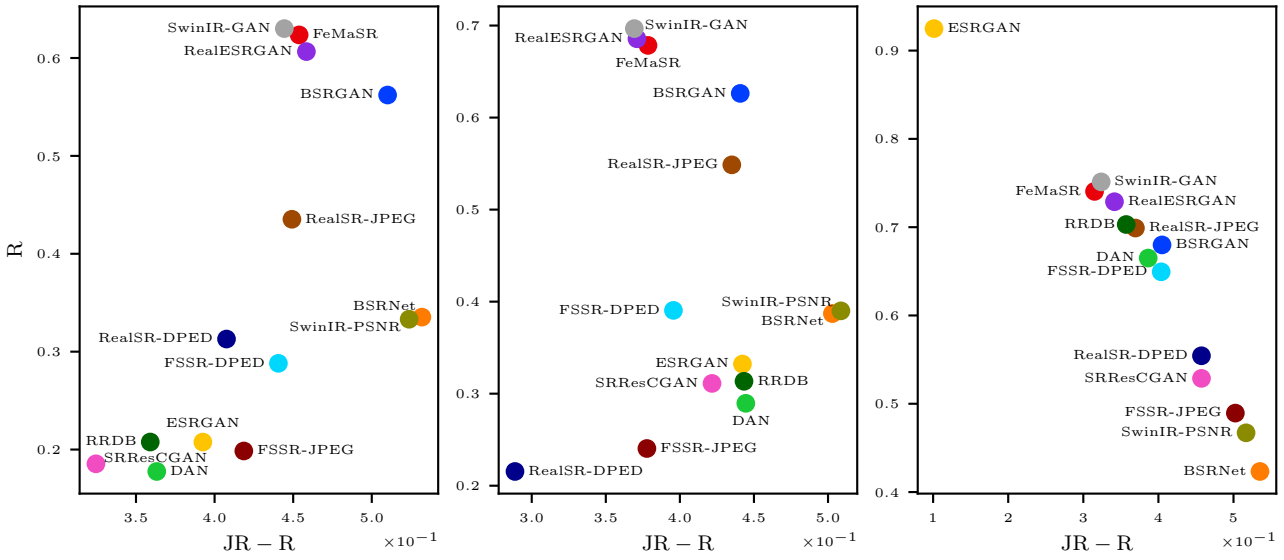

Figure 16: Plots of R versus JR − R. The left plot corresponds to the degradation from the Track2 challenge in (Lugmayr et al., 2019), the center plot corresponds to the degradation from the Track1 challenge in (Lugmayr et al., 2020), and the right plot corresponds to the common bicubic degradation. We see that two algorithms with a similar R value may attain different JR values. This shows that JR is not only influenced by perceptual quality, but also by consistency.

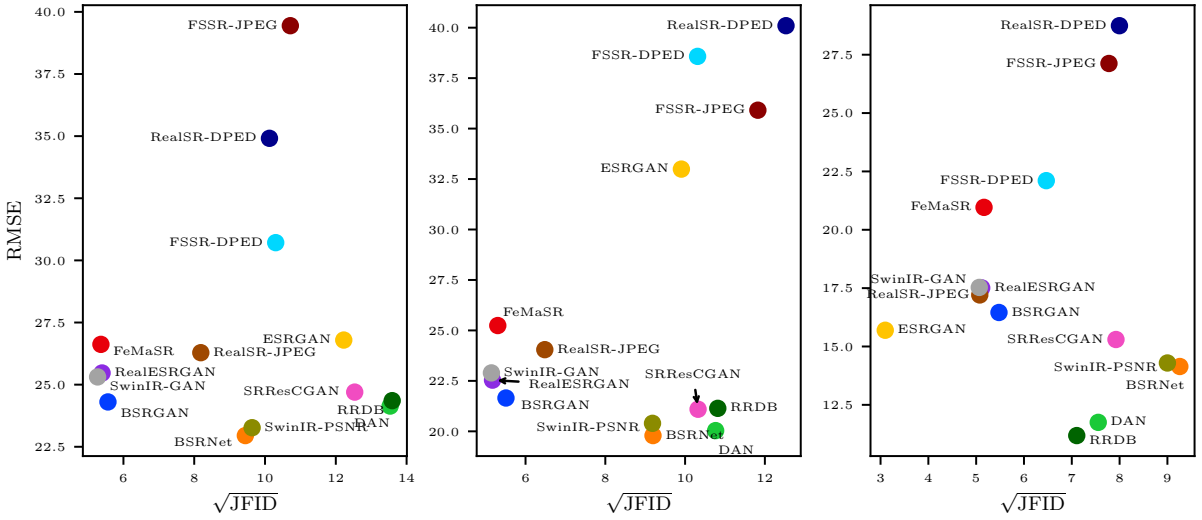

Figure 17: Plots of RMSE versus $\sqrt{\text{JFID}}$ of the SISR algorithms evaluated in Section 4.2. The left plot corresponds to the degradation from the Track2 challenge in (Lugmayr et al., 2019), the center plot corresponds to the degradation from the Track1 challenge in (Lugmayr et al., 2020), and the right plot corresponds to the common bicubic degradation.

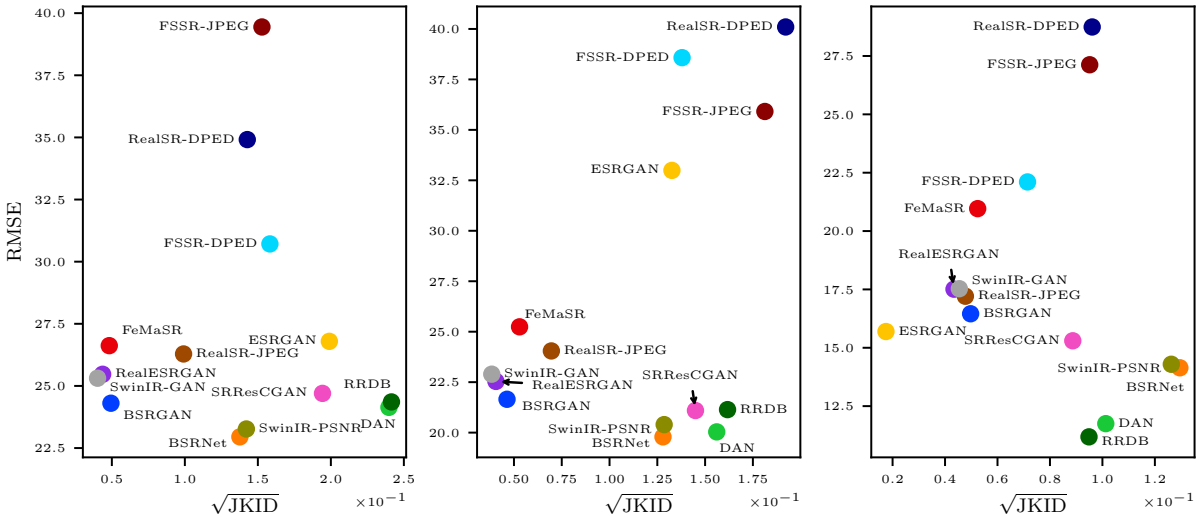

Figure 18: Plots of RMSE versus $\sqrt{\text{JKID}}$ of the SISR algorithms evaluated in Section 4.2. The left plot corresponds to the degradation from the Track2 challenge in (Lugmayr et al., 2019), the center plot corresponds to the degradation from the Track1 challenge in (Lugmayr et al., 2020), and the right plot corresponds to the common bicubic degradation.

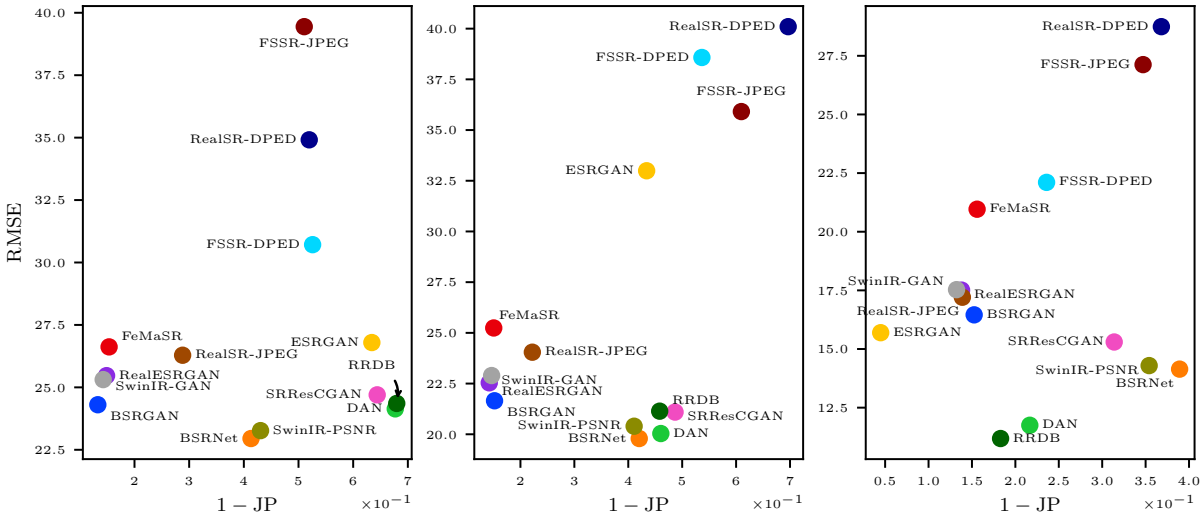

Figure 19: Plots of RMSE versus $1 - \text{JP}$ of the SISR algorithms evaluated in Section 4.2. The left plot corresponds to the degradation from the Track2 challenge in (Lugmayr et al., 2019), the center plot corresponds to the degradation from the Track1 challenge in (Lugmayr et al., 2020), and the right plot corresponds to the common bicubic degradation.

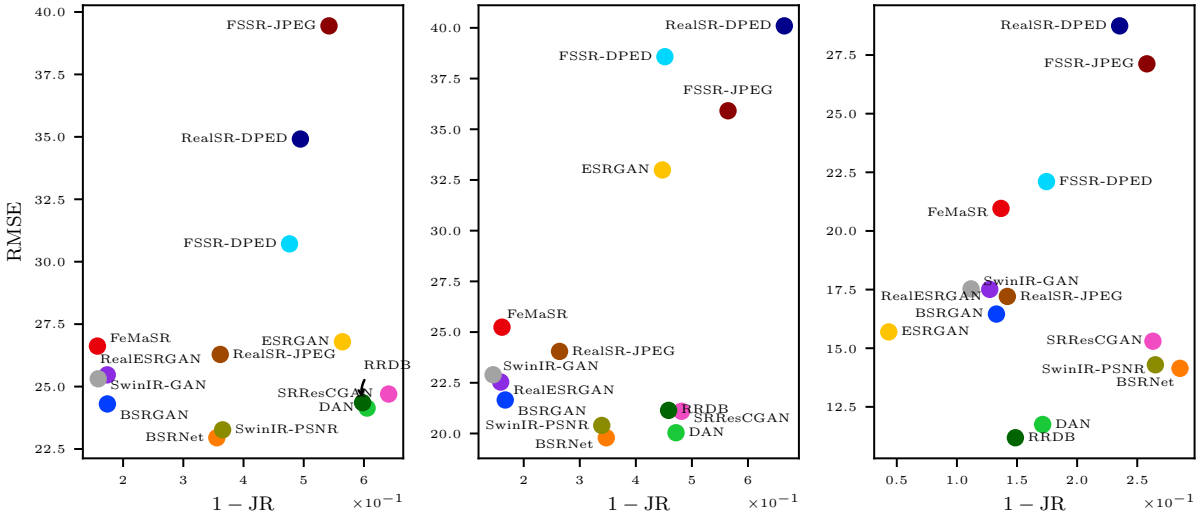

Figure 20: Plots of RMSE versus $1 - \text{JR}$ of the SISR algorithms evaluated in Section 4.2. The left plot corresponds to the degradation from the Track2 challenge in (Lugmayr et al., 2019), the center plot corresponds to the degradation from the Track1 challenge in (Lugmayr et al., 2020), and the right plot corresponds to the common bicubic degradation.

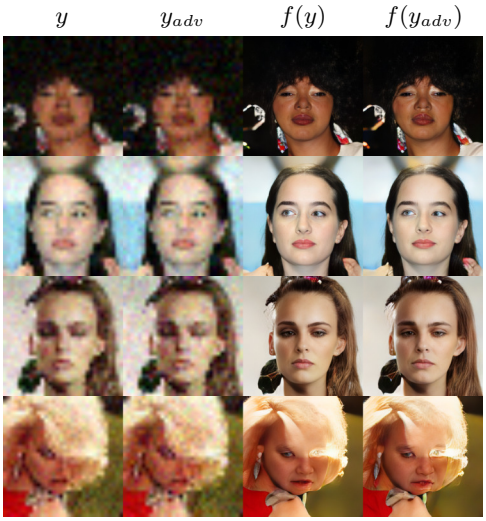

Figure 21: Adversarial attacks on low-resolution face images intended to alter the output of GFPGAN (Wang et al., 2021b) to produce a female face. **Left**: Successful results, where $f(y)$ is classified as "male" and $f(y_{adv})$ is classified as "female" by the gender classification model we use for evaluation. The attacks successfully change the classified gender of 9.3% of the images from "male" to "female". **Right**: Unsuccessful results, where $f(y)$ is classified as "female" and $f(y_{adv})$ is classified as "male". For GFPGAN, there are only 4 such unsuccessful examples from the 1000 images we use for evaluation (0.4% of the test data set).

Figure 22: Adversarial attacks on low-resolution face images intended to alter the output of the RRDB (Wang et al., 2018c) model to produce a female face. **Left**: Successful results, where $f(y)$ is classified as "male" and $f(y_{adv})$ is classified as "female" by the gender classification model we use for evaluation. The attacks successfully change the classified gender of 5.3% of the images from "male" to "female". **Right**: Unsuccessful results, where $f(y)$ is classified as "female" and $f(y_{adv})$ is classified as "male". 7.5% of the images that are originally classified as "female" are being classified as "male" after the adversarial attack.

Figure 23: Adversarial attacks on low-resolution face images intended to alter the output of the GFPGAN and RRDB models to produce a face of an older person. **Left**: Results using the GFPGAN model. **Right**: Results using the RRDB model.