High-Perceptual Quality JPEG Decoding via Posterior Sampling

Sean Man Guy Ohayon Theo Adrai Michael Elad
Technion - Israel Institute of Technology, Haifa, Israel
{sean.man,ohayonguy,theoad}@campus.technion.ac.il, elad@cs.technion.ac.il

Abstract

JPEG is arguably the most popular image coding format, achieving high compression ratios via lossy quantization that may create visual artifacts degradation. Numerous attempts to remove these artifacts were conceived over the years, and common to most of these is the use of deterministic post-processing algorithms that optimize some distortion measure (e.g., PSNR, SSIM). In this paper we propose a different paradigm for JPEG artifact correction: Our method is stochastic, and the objective we target is high perceptual quality – striving to obtain sharp, detailed and visually pleasing reconstructed images, while being consistent with the compressed input. These goals are achieved by training a stochastic conditional generator (conditioned on the compressed input), accompanied by a theoretically well-founded loss term, resulting in a sampler from the posterior distribution. Our solution offers a diverse set of plausible and fast reconstructions for a given input with perfect consistency. We demonstrate our scheme's unique properties and its superiority to a variety of alternative methods on the FFHQ and ImageNet datasets.

1. Introduction

JPEG (Joint Photographic Experts Group) [53] is one of the most popular *lossy* image compression techniques, extensively used in digital cameras, internet communications and more. JPEG reduces image file-size by discarding information that is supposed to be less valuable to a human observer. To achieve high compression ratios, JPEG often discards noticeable visual details, which may lead to strong artifacts in the decompressed image, such as blockiness.

Since its conception in the late 80's, numerous post-processing algorithms were proposed for removing JPEG artifacts.[1] Such methods start with the given compressed image and somehow provide an estimation of the source image. A common estimation approach attempts to minimize the average discrepancy between the source image

[1] While it is outside the scope of this paper to reference this vast literature, we do provide links to leading such techniques in Section 2.

Figure 1. Reconstruction examples of highly compressed JPEG images using [5] and our proposed method. Our method produces stochastic outputs, all of which are perfectly consistent with the compressed image, like [5], but with far better perceptual quality.

and the recovered one, *e.g.*, minimizing the Mean-Squared-Error (MSE). While this strategy may improve the quality of the compressed input and remove the blockiness effect, it still evidently leads to visually unpleasing images that are often accompanied by blurriness. As shown in [6], this is a direct manifestation of the perception-distortion tradeoff, which is apparently noticeable and strict for natural images and common distortion measures – low MSE contradicts high perceptual quality.

Another key aspect in JPEG restoration processes is the consistency of the reconstructed image with the compressed input: The compressed version of the output image should be identical to the compressed input. This is a reasonable requirement as any inconsistent image could never be the true source. Interestingly, while older classical methods tend to be consistent (*e.g.* using iterative projection algorithms [31, 38, 39, 49]), this feature has been overlooked in recent learning-based recovery techniques (such as [8, 14, 15, 18]).

In this work we propose a new JPEG reconstruction algorithm that produces consistent outputs, while also attaining very high perceptual quality. The produced images by our algorithm are free of artifacts such as blockiness, blurriness, etc. However, due to the perception-distortion tradeoff [6], this comes at the unavoidable cost of reduced distortion performance. Example reconstructions from our method are presented in Figure 1. Our approach is based on a Generative Adversarial Network (GAN) [17], conditioned on the compressed image, and having a stochastic synthesis process. As such, the generator is able to produce a variety of plausible and consistent reconstructions for a given compressed input. This conditional GAN is trained with a loss comprised of *(i)* an adversarial term that promotes high perceptual quality, *(ii)* a penalty term adopted from [37] that promotes output variability (per input), and *(iii)* a consistency term that promotes a compliance with the compressed input. Our optimization task theoretically admits the posterior sampler as a unique solution, leading to perfect perceptual quality and perfect consistency with the compressed input.

The contributions of this paper are the following: *(i)* We propose a novel algorithm that produces consistent JPEG reconstructions with high perceptual quality, leading to state-of-the-art (SoTA) results on ImageNet; *(ii)* We identify an empirical perception-consistency tradeoff of stochastic estimators, extending the results presented in [36]. Through a consistency penalty term (instead of enforcing perfect consistency) we are able to traverse this empirical tradeoff; *(iii)* We analyze and compare our method with previous approaches that targetted the same problem, while not being aware of this empirical tradeoff.

2. Related work

A plethora of ideas on JPEG's artifacts removal were published since the late 80's, typically offering a post-processing mechanism within the decoding stage. These include techniques such as simple spatial filters [11, 28, 41], MSE optimization of codebooks [22, 57], MAP optimization using Gibbs priors [29, 39, 49], sparse representations [10, 25] and many other techniques (*e.g.*, see the review by Shen and Kuo [46]).

In recent years, deep-learning-based methods took the lead in post-processing JPEG compressed images. Dong *et al*. [14] were the first to suggest a CNN-based regression model, followed by work such as [8] that improved the architecture and the loss function. Guo *et al*. [18] combined pixel domain and DCT domain sub-networks to take advantage of JPEG's mode of operation, followed by work such as [56, 58] that improved the dual-domain concept and [15, 50] that took a step further by operating solely in the DCT domain. All of these methods adopt a supervised approach, training a network to best fit a given compressed image to its desired ideal output.

As noted in Section 1, a common difficulty of these techniques (classical or deep-learning based) is the overly smoothed reconstructed images that lack fine texture and high frequency details, which occurs due to the inherit tradeoff between perceptual quality and an optimization of a distortion criterion [6]. Indeed, some methods [15, 16, 19] incorporated adversarial training [17] to produce sharp details while not being aware of this tradeoff, hence kept incorporating distortion measures as part of their loss, and thus compromising on the perceptual quality of their outcome as well.

Among the deep-learning-based works, only a few addressed the desire that the reconstructed images should be consistent with the compressed inputs. The work reported in [50] enforced this requirement as a constraint on the resulting DCT coefficients, while [19] encouraged the reconstructions to obey this requirement using a penalty as part of the optimization criterion.

We now turn to discuss two recent and relevant works that inspired this project [5, 37]. PSCGAN [37] is a recently proposed image denoiser that attempts to attain a stochastic estimator that samples from the posterior distribution. While [37] bares some similarities to our work, it differs in the task addressed (denoising vs. JPEG reconstruction), the conditional GAN loss used [17, 34], the fact that it disregards consistency, and the very different neural network architectures deployed. Nevertheless, [37] introduce a first moment loss that encourages acceptable distortion but bypasses the perception-distortion tradeoff, which we leverage in our JPEG recovery algorithm.

The work by Bahat *et al*. [5] deserves a special mention as it is the only prior work, to the best of our knowledge, that addresses both the perception-distortion tradeoff and the consistency with the compressed input. [5] imposes consistency by predicting bounded residuals in the DCT domain and optimizing a GAN loss that bypasses a distortion criteria, hence avoiding the pitfall of the mentioned tradeoff. Nonetheless, our proposed approach achieves much better perceptual quality results compared to [5] while retaining perfect consistency. We give an explanation to this improvement in Section 4 based on the empirical perception-consistency tradeoff, and conduct extensive experiments to compare both methods in Section 5.

3. Method: fundamentals

We assume that a natural image X is a multivariate random variable with probability density function p_X. We denote by Y the JPEG-compressed-decompressed version of X, and from Y we provide an estimate of X, denoted by $\hat{X}$. We do so by attempting to sample from the posterior distribution $p_{X|Y}$. Such an estimator would be highly effective since *(i)* it's outputs are consistent with the measurements[2], i.e. $\mathit{JPEG}(\hat{X}) = Y$; and *(ii)* it attains perfect perceptual quality [6], i.e., $p_{\hat{X}} = p_X$.

3.1. JPEG

The JPEG compression algorithm [53] operates using a pre-defined quantization matrices $Q \in \mathbb{Z}^{8\times 8}$ in the DCT domain on an image X in the YCbCr color-space, which, for simplicity, we assume is of size $8m \times 8n \times 3$. The algorithm starts by dividing the image's channels into 8×8 blocks $\{X_i\}^{m\times n\times 3}$. Each block X_i is transformed by the 2D-DCT transformation $X_i^D = \text{2D-DCT}(X_i)$ and divided element-wise by the corresponding quantization matrix Q to achieve $X_i^Q = X_i^D \oslash Q$. Finally, each entry of the block is rounded to achieve $X_i^R = \lfloor X_i^Q \rceil$. All the blocks $\{X_i^R\}^{m\times n\times 3}$ are stored using a lossless entropy-coding algorithm alongside the matrix Q. Decompression is obtained by multiplying these rounded values by Q and applying an inverse 2D-DCT on the resulting blocks. We denote the above compression-decompression process by $\mathit{JPEG}_Q(\cdot)$.

Note that rounding is a lossy operation, and the matrices Q controls the amount of lost information. These matrices are a function of the quality-factor (QF) chosen by the user – an integer between 0 to 100, where 0 means maximum compression (in this paper we use the baseline matrices defined in [2]).

[2]In our notations throughout the paper, $Y = \mathit{JPEG}(X)$ implies a compression-decompression operation, and thus Y is an image of the same size as X.

3.2. Consistency

To measure the consistency of $\hat{X}$ with Y, we define the consistency index by

$$c(p_{\hat{X}}, p_Y) = \mathbb{E}_Y\left[\mathbb{E}_{\hat{X}|Y}\left[\left\|Y - \mathit{JPEG}_Q(\hat{X})\right\| |Y\right]\right], \quad (1)$$

The estimator $\hat{X}$ is perfectly consistent with the compressed inputs Y if $Y = \mathit{JPEG}_Q(\hat{X})$, or equivalently, iff $c(p_{\hat{X}}, p_Y) = 0$. Interestingly, as we show next, prior JPEG restoration models that attempt to minimize the MSE loss implicitly striven (at least theoretically) to become perfectly consistent, since the MMSE estimator produces consistent restoration:

Theorem 1. *Let $\bar{X}(Y)$ be an MMSE estimator for JPEG artifact removal, i.e $\bar{X}(Y) = \mathbb{E}[X|Y]$. Then $\bar{X}(Y)$ is necessarily perfectly consistent with the compressed input Y.*

See proof in Appendix B .

3.3. Perceptual quality

While there are several definitions for the notion of perceptual quality, we follow the notion developed in [6], which measures for an estimator $\hat{X}$ the index

$$d(p_X, p_{\hat{X}}), \quad (2)$$

where $d(p, q)$ is some divergence between the two distributions, *e.g.*, Kullback-Leibler's (KL), Jensen-Shanon's (JS) divergence, or Wasserstein distance. $\hat{X}$ attains perfect perceptual quality when $d(p_X, p_{\hat{X}}) = 0$, i.e., $p_X = p_{\hat{X}}$.

3.4. Posterior sampling - our goal

The work in [36] ties the perceptual quality and the consistency of estimators via the following theorem:

Theorem 2. *Let $X \sim p_X$ be a random multivariate variable, and let $Y = D(X)$ be a deterministic degradation of X. If $\hat{X}$ is an estimator such that $p_X = p_{\hat{X}}$ and $Y = D(\hat{X})$, then $p_{\hat{X}|Y} = p_{X|Y}$, i.e., $\hat{X}$ is a necessarily posterior sampler.*

Proof. Found in [36]. □

In our case, $Y = D(X) = \mathit{JPEG}_Q(X)$. Recall that X cannot be uniquely recovered from Y since $\mathit{JPEG}_Q(\cdot)$ is a non-invertible degradation, and thus the support of $p_{X|Y}$ is not a singleton – there is a variety of possible sources that correspond to the same compressed image. Since we aim to sample from the posterior, our method must be stochastic, capable of providing many reconstructed samples given the same compressed image it. This is in opposition to most prior work, which adopts a deterministic recovery strategy.

Our solution is comprised of a conditional GAN [35] that takes Y as an input and produces high perceptual quality outputs that are consistent with Y. Posing our task as finding a sampler from $p_{\hat{X}|Y}$, we form a loss function that has two ingredients. Just as with all GAN methods, we use an adversarial loss $\mathcal{L}_{\text{Adv}}$ to minimize the divergence between the real and the generated data, matching p_X and $p_{\hat{X}}$ as best as possible. To promote consistency, we include another objective that penalizes any discrepancy between Y and $JPEG_Q(\hat{X})$, leading to the following final loss:

$$\begin{aligned} &\min_{p_{\hat{X}|Y}} \mathcal{L}_{\text{Adv}}(p_X, p_{\hat{X}}) \\ &+\lambda \mathbb{E}_Y \mathbb{E}_{\hat{X}|Y}\left[\left\|Y - JPEG_Q(\hat{X})\right\|_2^2\right]. \end{aligned} \tag{3}$$

According to Theorem 2, Eq. 3 admits a single optimal solution for any $\lambda > 0$: $p_{X|Y}$. This is true since any optimal solution would attain perfect perceptual quality and produce perfectly consistent restorations. Practically, however, we solve Eq. 3 with parametric neural networks and with a data set of finite size, and thus our solution may not be the true posterior. Moreover, due to the nature of practical optimization, the choice of λ may also affect the obtained solution, as discussed in the following section.

4. Method: practice

4.1. Achieving the posterior

The work in [36] revealed that a posterior sampler is the only consistent restoration algorithm that attains perfect perceptual quality. This leads to a tradeoff between consistency and perceptual quality for deterministic estimators, as any such estimator cannot be a posterior sampler. This theoretical tradeoff **does not** affect our method, as our algorithm is stochastic. In practice, however, it is very likely that a suboptimal optimization procedure, a highly non-convex loss surface, a finite size data set, and a limited capacity architecture, would all make it extremely hard to attain the perfect solution – a sampler from the posterior. Thus, we expect to improve both the consistency index (Eq. 1) and the perceptual index (Eq. 2) up to a certain point, beyond which we shall observe an *empirical* tradeoff, where the improvement of one quality comes at the expense of the other. This empirical tradeoff is not revealed or discussed in [36]. By changing λ in our optimization task (Eq. 3), such a tradeoff can be controlled, i.e., we can decide to attain higher perceptual quality or better consistency.

As such an empirical tradeoff has not been revealed in prior work on stochastic estimators, the balance between perceptual quality and consistency has been implicitly addressed. Bahat *et al.* [5] imposed a consistency requirement as a constraint using their generator architecture and not as a penalty. This can be interpreted as choosing $\lambda \to \infty$, requiring perfect consistency at the cost of lower perceptual quality. On the other hand, prior work that attempted to attain the posterior without paying attention to consistency in any way, such as [37], controlled the tradeoff implicitly by the choice of architecture, loss, and optimizers. In fact, we argue that any attempt to sample from the posterior would most likely compromise on either consistency or perceptual quality, since attaining the true posterior is highly challenging.

4.2. Training method

We denote our estimator by $G_\theta(Z, Y)$, where G_θ is a neural network, Y is the input JPEG-compressed image, and Z is a random seed that enables a diverse set of outputs for any input image Y. Our training procedure consists of a weighted sum of several objectives. First, a non-saturating adversarial loss term [17], accompanied by an R_1 gradient penalty for the critic [34]. We denote these GAN losses by $V(D_\omega, G_\theta)$ and $R_1(D_\omega)$, where D_ω is our critic. Second, we use a consistency penalty term $C(G_\theta)$, as in Eq. 3:

$$C(G_\theta) = \mathbb{E}_{Y,Z}\left[\|Y - JPEG_Q(G_\theta(Z, Y))\|_2^2\right]. \tag{4}$$

Third, we incorporate a first-moment penalty term $FM(G_\theta)$, originally proposed in [37]:

$$FM(G_\theta) = \mathbb{E}_{X,Y}\left[\|X - \mathbb{E}_Z[G_\theta(Z, Y)|Y]\|_2^2\right], \tag{5}$$

which specifies that the average of many outputs $G_\theta(Z, Y)$ that refer to a fixed Y while varying Z should be close to the ideal image X. If indeed this multitude of outputs form a fair sampling from the posterior, this penalty leads exactly to the MMSE estimation – $\mathbb{E}_Z[G_\theta(Z, Y)|Y] = \mathbb{E}_X[X|Y]$. This term is a natural force that replaces the more intuitive supervised distortion penalty $\mathbb{E}_{X,Z}\left[\|X - G_\theta(Z, Y)\|_2^2\right]$. As we have already mentioned, a distortion penalty typically hinders the perceptual quality, while the alternative in Eq. 5 does not, further strengthening the overall optimization. More specifically, without using Eq. 5 we observe mode-collapse during training – as we only have one X per given Y in our dataset, the generator is not incentivized to generate stochastic reconstructions, hence, it almost completely ignores Z (as explained in [37]).

On top of the above, in complex general content datasets such as ImageNet [13] we empirically find that further guidance is required in order to achieve satisfactory results. In these scenarios we include a VGG "perceptual" loss [24, 48], which promotes the generation of fine details in severely compressed images:

$$P(G_\theta) = \mathbb{E}_{X,Y,Z}\left[\|\text{VGG}_{5,4}(X) - \text{VGG}_{5,4}(G_\theta(Z, Y))\|_2^2\right]. \tag{6}$$

Where $\text{VGG}_{5,4}(\cdot)$ are the features of a trained VGG-19 network at the specified convolutional layer. Lastly, in order to increase the variation in the estimator's reconstructions, we introduce a new second-moment penalty:

$$SM(G_\theta) = \mathbb{E}_{X,Y}\left[\left\|\left(X - \bar{X}(Y)\right)^2 - \text{Var}_Z[G_\theta(Z,Y)|Y]\right\|\right]. \tag{7}$$

This term specifies that the variance of the generated images for a given Y should be close to the sample variance using a single **ground-truth** sample and a pre-trained MMSE estimator $\bar{X}(Y)$. In Appendix C we give further rational behind this penalty.

In Subsection 5.3 we present an ablation study to show the importance of the VGG loss and the second-moment penalty. Note that while we do not have a theoretical guaranty that the VGG loss does not introduce a perception-distortion tradeoff, we see in our experiments that both the perceptual quality and the consistency of the generated images improves.

All of the above forces result in the following unified minimax game:

$$\begin{aligned}\min_\theta \max_\omega & V(D_\omega, G_\theta) + \lambda_{R_1} R_1(D_\omega) + \lambda_C C(G_\theta) \\ & + \lambda_{FM} FM(G_\theta) + \lambda_P P(G_\theta) + \lambda_{SM} SM(G_\theta).\end{aligned} \tag{8}$$

We solve this task using a block-coordinate optimization, resulting in alternating optimization tasks:

$$\max_\omega V(D_\omega, G_\theta) + \lambda_{R_1} R_1(D_\omega), \tag{9}$$

$$\begin{aligned}\min_\theta & V(D_\omega, G_\theta) + \lambda_C C(G_\theta) + \lambda_{FM} FM(G_\theta) \\ & + \lambda_P P(G_\theta) + \lambda_{SM} SM(G_\theta).\end{aligned} \tag{10}$$

We should note that the consistency penalty term requires a differentiable implementation of JPEG, a concept introduced in prior work such as [32, 47, 51]. We opt to approximate the backward pass of the rounding operation using $\nabla\lfloor X\rceil = 1$.

4.3. Projection

As we enforce consistency through the training objective and not via the architecture, we can expect the results of our trained models to be inconsistent to some degree. Post training, though, we can still produce perfect consistency by projecting the DCT coefficients of any reconstructed image $\hat{X}$ to the range permitted by the rounding operation that created Y. We denote the projected results by $\tilde{X}$. Per block, and in the DCT domain, the projection operation is defined as

$$\tilde{X}_i^Q = Y_i^Q + \max\left(\min\left(\hat{X}_i^Q - Y_i^Q, 0.5\right), -0.5\right). \tag{11}$$

Note that it is expected that our perceptual quality would degrade as a result of the projection operation, since it is not guaranteed to result in a natural image and also due to empirical perception-consistency tradeoff for stochastic estimators. However, for a model that attains high perceptual quality and satisfying consistency, this degradation should be minor. We demonstrate the projection's effect on our models in Section 5.

5. Experiments

Methods: We compare our method (denoted **Ours**) with several alternative post-processing methods: *(1)* **QGAC** and **QGAC-GAN** [15], a SoTA regression method and a GAN method fine-tuned from it; *(2)* **SwinIR** [30], a SoTA regression methods trained separately for each QF; *(3)* **FBCNN** [23], a SoTA regression method; *(4)* **Bahat *et al.*** [5], as noted in Section 2, the only other work that attempts to attain perfect consistency and perceptual quality; and *(5)* **Ours-MSE**, our very same architecture trained as a regression model using solely an MSE loss as a baseline. Following Theorem 1, this model should produce consistent reconstructions. Moreover, we test the results obtained by our method after projection (as explained in Subsection 4.3), and denote these by **Ours-P**.

Unless mentioned otherwise, when possible we use checkpoints as published by the authors. In other cases that require training, we use commonly available hardware.[3] Please refer to Appendix A for details on training and architectures.

Metrics: To measure consistency, we compute the RMSE between the compressed-decompressed versions of the original and the restored images using the same JPEG settings. Note that the value shown is per-pixel and in units of gray-levels.

To evaluate perceptual quality we compute the FID [21, 43] between the real uncompressed images and the restored ones. For stochastic methods (ours and Bahat *et al.*'s) we present the mean and standard deviation of 64 repeated FID evaluations, where in each we generate one restoration per compressed test image. This makes sure that our model consistently performs with high perceptual quality regardless of the seed. We also present a "ground truth" score, which is the FID between the training and the validation images (in Subsection 5.1) or between the validation and the test images (in Subsection 5.2).

5.1. Results: FFHQ

To showcase the empirical perception-consistency tradeoff we use the FFHQ [26] thumbnails dataset, in which each image is of size 128×128. We compress the images with QF=5 and use the same train-validation-test split as in [37]. We compare our method to Bahat *et al.* (trained using their official implementation and following the training method described in their paper) and the base-

[3] The networks were trained on a single NVIDIA A6000 GPU.

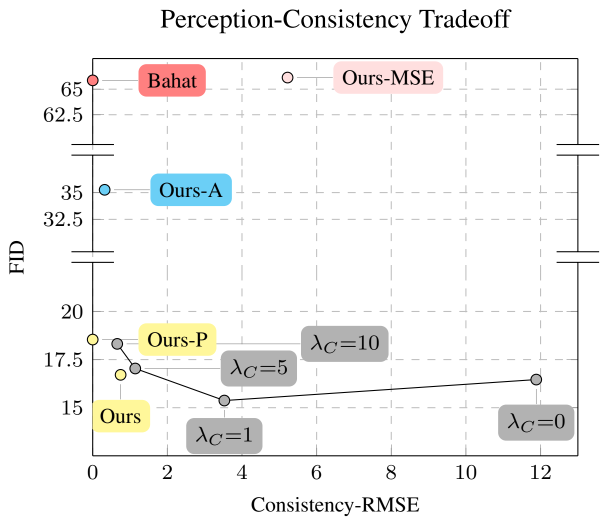

Figure 2. FID versus Consistency of different methods on FFHQ-128 with QF=5. By adjusting λ_C, our method exposes explicit control over the perception-consistency tradeoff of the trained estimator. Observe that when λ_C is increased, the consistency of the estimators improves up to a point where consistency comes at the expense of perceptual quality. By carefully tuning λ_C during training we are able to reach an estimator (Ours) with similar perceptual quality as the quality attained when using λ_C=0, but with a much improved consistency. The consistency of constrained methods (Ours-P, Bahat [5]) are practically zero – see Subsection 5.1 for more details.

line Ours-MSE method. To show the importance of the consistency regularization, we also compare the performance of our method trained by setting $\lambda_C = 0$ in the loss formulation (Eq. 10). As a proxy to an MMSE estimator, we also compare the performance of an average of 64 realizations per input from our model, denoted as **Ours-A**. The empirical perception-consistency tradeoff of the different methods on the FFHQ-128 dataset is visualized in Figure 2. Further visual results and quantitative FID, consistency and PSNR results are summarized in Appendix E (Table 4).

Consistency: As shown in Figure 2, Ours with λ_C=0 and Ours-MSE both produce unsatisfactory consistency levels, which suggests that to attain high consistency, some type of supervision is required. Just by activating our penalty term we are able to improve the consistency-RMSE by more than 11 gray-levels.

Note that both Ours-A and Ours-MSE are approximations of an MMSE estimator, and they differ only in their loss functions. While Ours-MSE attains slightly higher PSNR (see Appendix E), Ours-A attains significantly better consistency. Following Theorem 1 we know that an MMSE estimator should be perfectly consistent, so Ours-A is closer to a true MMSE estimator in that sense, even thought it attains a slightly lower PSNR compared to the regression model.

By projecting the restored images (Ours-P) we improve the consistency significantly, bringing our method to be on-par with Bahat's. While the projection should have resulted in perfect-consistency, we get a slight deviation due to numerical approximations in the JPEG algorithm (in the color-space conversion). This is also apparent in the results of Bahat's method that enforce the consistency as part of the architecture. In Appendix D we further investigate this phenomenon and show that it affects even the standard JPEG implementation *libjpeg* [1].

In Appendix E we present more visual and quantitative results regarding the consistency of the different methods.

Perception-consistency tradeoff: By controlling λ_C in Eq. 10 we can incentivize our generator to produce more consistent reconstructions with the compressed input, as evident in Figure 2. Starting with λ_C=0, we converge to an estimator with good perceptual quality but with lacking consistency. By choosing a small penalty, such as λ_C=1, we are able to converge to an estimator that achieves both better perceptual quality and better consistency. Yet, the consistency of the results are far from being satisfactory. Cranking the penalty coefficient up to λ_C=5 improves the consistency significantly, but the perceptual-quality deteriorate compared to λ_C=0. This trend continues as we further increase λ_C.

By carefully adjusting λ_C during training (please refer to Appendix A.2 for details) we are able to achieve a balanced result – the perceptual quality of λ_C=0 with the improved consistency of a large penalty term. This exploration demonstrates our main point in Subsection 4.1 – the chosen penalty term gives us explicit control over the empirical perception-consistency tradeoff, which lets us converge to a better local-minima, as opposed to the implicit or no control in previous methods. We can expect further improvements with more hyper-parameter tuning.

5.2. Results: general content

To showcase the performance of our method we train our model on 128×128 patches extracted from DIV2k [3] and Flickr2k [52] datasets at multiple QFs in the range $[5, 50]$ and test all the mentioned methods on the ImageNet-ctest10k dataset, as proposed in [42]. QGAC [15] and FBCNN [23] are also trained on DIV2K and Flickr2K, SwinIR is also trained on BSDS500 [4] and WED [33], while Bahat *et al.* [5] is trained on ImageNet [13]. In Figure 3 we present the FID, consistency and PSNR of the different methods across a range of QFs from 5 to 50 on ImageNet-ctest10k, and in Figures 1 and 4 we present reconstruction examples from ImageNet-ctest10k of some of the different methods. Further visual results on ImageNet, LIVE1 [44,45] and BSDS500 are presented in Appendix E.

Perceptual quality & consistency: Following the trends

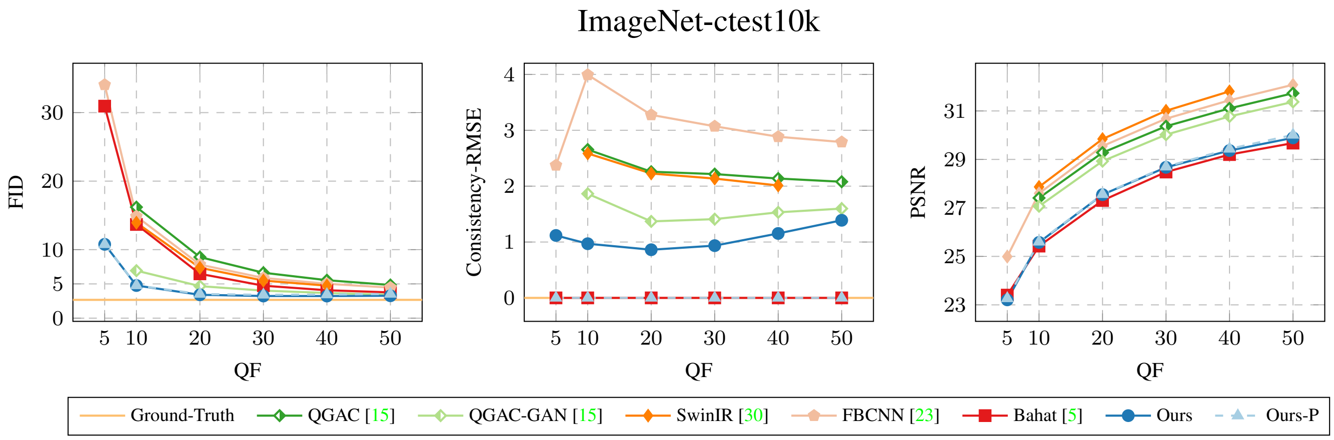

Figure 3. FID, Consistency and PSNR results of different methods on ImageNet-ctest10k across multiple QFs. By using explicit control over the perception-consistency tradeoff and by avoiding the perception-distortion tradeoff, our method provides SoTA FID results across a wide-range of QFs while being consistent with the compressed inputs. Each method is evaluated only on QFs it was trained on. The consistency of Ours-P and Bahat are practically zero – see Subsection 5.1 for more details.

shown on FFHQ-128, Ours-P creates consistent reconstructions across a wide range of QFs similarly to Bahat's, but with much improved perceptual quality.

When comparing to QGAC, SwinIR and FBCNN we see a predictable result – the regression models achieves the worst perceptual quality and QGAC-GAN model shows better results compared to Bahat's model, but they all fail to generate consistent reconstructions. This is a direct result of the perception-distortion tradeoff [6] as also manifested in the PSNR results.

While QGAC, SwinIR and FBCNN are not trained to minimize an MSE loss, they achieve SoTA results in terms of PSNR on JPEG reconstruction, hence, they could be seen as a compelling MMSE estimator candidates. As such, they should have near-perfect consistency following Theorem 1, but clearly this is not the case. Similarly to the experiments shown in Subsection 5.1 on Ours-A and Ours-MSE, it seems that explicit supervision on the consistency of the reconstructed images, or enforcing it via the architecture, is necessary to achieve consistent results.

5.3. Ablation

As explained in Subsection 4.2 our method is trained using several loss terms. For the FFHQ-128 data set we use a GAN loss, a consistency loss, and the first-moment penalty, while for ImageNet we also use a VGG “perceptual” loss and a second-moment penalty. In Table 1 we present an ablation study of training the model with the different terms on the ImageNet data set, and here we provide a detailed explanation for each row in the table: **Baseline:** By using GAN and consistency losses alone (Eq. 3) we attain better perceptual quality (FID) than Bahat [5] but with lower variability (Per-Pixel STD); + **First-Moment:** As noted in [37], the first-moment penalty alleviates the mode-collapse issue of conditional GANs, and indeed, it significantly improves the output variation of our model; + **VGG:** A perceptual loss further improves the perceptual quality at the cost of lower variability; + **Second-Moment:** The new penalty increases output variation without hindering perceptual quality, hence suggesting that the increased output variation is of meaningful details; + **Ours:** With increased second-moment coefficient we further improve the output variability; + **Ours-P:** By enforcing perfect consistency via projection we still attain much better perceptual quality and output variation compared to Bahat [5]; **Perceptual Baseline:** Without explicitly requiring consistency we achieve similar FID to our method but without consistency with the measurements.

6. Summary

In this work we approach the JPEG decompression task from an uncommon direction – generate visually pleasing and consistent reconstructions by leveraging recent advancements in image restoration, such as the perception-distortion and the perception-consistency theoretical tradeoffs. Using these tools we surpass prior work and provide decompressed JPEG images of tunable consistency and high perceptual quality.

The proposed solution is based on a stochastic conditional GAN with carefully tailored loss function that promotes detailed and vivid results, consistency to the measurements, proximity to the training data without scarifying quality, and a spread of the randomized results. Our future work will focus on better diversifying the obtained solutions and on a quest for a tractable computational method for evaluating the proximity of the obtained model to the ideal posterior sampler.

Table 1. Ablation study on the ImageNet-ctest10k dataset at QF=10. The VGG loss is crucial for good-perceptual quality without hurting consistency. The second-moment penalty increases per-pixel standard deviation (a value between 0 and 1) without hurting either perceptual quality nor consistency. The consistency of constrained methods, marked by *, are practically zero – see Subsection 5.1 for more details.

Method	GAN	C	FM	P	SM	FID (↓)	Consistency (↓)	PSNR (↑)	Per-Pixel STD (↑)
Perceptual Baseline	✓		✓	✓		4.54	6.9724	26.0612	0.0111
Baseline	✓	✓				9.12	0.9650	25.5061	0.0021
+ First-Moment	✓	✓	✓			8.83	0.9694	25.5767	0.0064
+ VGG	✓	✓	✓	✓		4.71	0.9411	25.8906	0.0026
+ Second-Moment	✓	✓	✓	✓	✓ (10^1)	4.70	0.9452	25.8539	0.0063
Ours	✓	✓	✓	✓	✓ (10^3)	4.76	0.9696	25.5758	0.0194
Ours-P	✓	✓	✓	✓	✓ (10^3)	4.78	≈ 0*	25.6008	0.0188
Bahat						13.71	≈ 0*	25.4243	0.0040

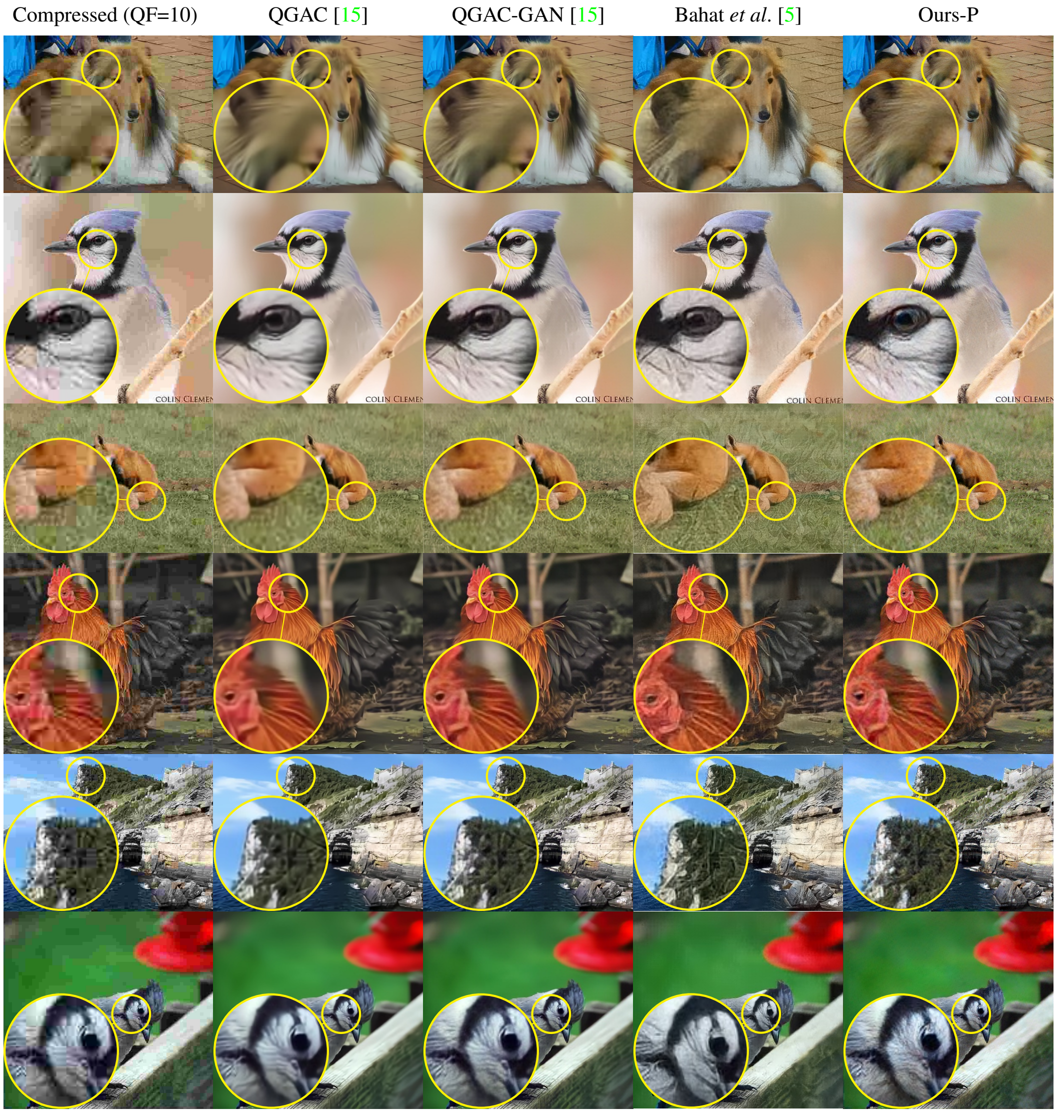

Figure 4. Reconstruction examples on ImageNet-ctest10k at QF=10; Bahat is the only other method that generates consistent reconstructions and aims for high perceptual quality; QGAC-GAN's reconstructions are sharp but not consistent, hence could not have created the input image; Our method generates finer details while still being consistent, as seen in Figure 3; Zoom-in on interesting regions are shown.

References

[1] Independent JPEG Group. http://www.ijg.org/. 6, 13

[2] Recommendation T.81: Information technology - Digital compression and coding of continuous-tone still images - Requirements and guidelines, Sept. 1992. 3

[3] Eirikur Agustsson and Radu Timofte. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–135, 2017. 6

[4] Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour Detection and Hierarchical Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, May 2011. 6, 15

[5] Yuval Bahat and Tomer Michaeli. What's in the Image? Explorable Decoding of Compressed Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2908–2917, 2021. 1, 2, 3, 4, 5, 6, 7, 8, 13, 16

[6] Yochai Blau and Tomer Michaeli. The Perception-Distortion Tradeoff. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, June 2018. 2, 3, 7

[7] G. Bradski. The OpenCV library. *Dr. Dobb's Journal of Software Tools*, 2000. 13

[8] Lukas Cavigelli, Pascal Hager, and Luca Benini. CAS-CNN: A deep convolutional neural network for image compression artifact suppression. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 752–759, May 2017. 2

[9] Lucy Chai, Michaël Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-Resolution Training for High-Resolution Image Synthesis. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, volume 13676, pages 170–188. Springer Nature Switzerland, Cham, 2022. 12

[10] Huibin Chang, Michael K. Ng, and Tieyong Zeng. Reducing Artifacts in JPEG Decompression Via a Learned Dictionary. *IEEE Transactions on Signal Processing*, 62(3):718–728, Feb. 2014. 2

[11] T. Chen, H.R. Wu, and B. Qiu. Adaptive postfiltering of transform coefficients for the reduction of blocking artifacts. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(5):594–602, May 2001. 2

[12] Alex Clark. Pillow (PIL fork), 2015. 13

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. 4, 6

[14] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. Compression Artifacts Reduction by a Deep Convolutional Network. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 576–584, Dec. 2015. 2

[15] Max Ehrlich, Larry Davis, Ser-Nam Lim, and Abhinav Shrivastava. Quantization Guided JPEG Artifact Correction. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 293–309, Cham, 2020. Springer International Publishing. 2, 5, 6, 7, 8, 14

[16] Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, and Alberto Del Bimbo. Deep Universal Generative Adversarial Compression Artifact Removal. *IEEE Transactions on Multimedia*, 21(8):2131–2145, Aug. 2019. 2

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 2, 4

[18] Jun Guo and Hongyang Chao. Building Dual-Domain Representations for Compression Artifacts Reduction. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 628–644, Cham, 2016. Springer International Publishing. 2

[19] Jun Guo and Hongyang Chao. One-To-Many Network for Visually Pleasing Compression Artifacts Reduction. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4867–4876, July 2017. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 12

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 5

[22] Sung-Wai Hong, Yuk-Hee Chan, and Wan-Chi Siu. A practical real-time post-processing technique for block effect elimination. In *Proceedings of 3rd IEEE International Conference on Image Processing*, volume 2, pages 21–24 vol.2, Sept. 1996. 2

[23] Jiaxi Jiang, Kai Zhang, and Radu Timofte. Towards Flexible Blind JPEG Artifacts Removal. Sept. 2021. 5, 6, 7

[24] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 694–711, Cham, 2016. Springer International Publishing. 4

[25] Cheolkon Jung, Licheng Jiao, Hongtao Qi, and Tian Sun. Image deblocking via sparse representation. *Signal Processing: Image Communication*, 27(6):663–677, July 2012. 2

[26] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, June 2019. 5

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 12

[28] A. Kundu. Enhancement of JPEG coded images by adaptive spatial filtering. In *Proceedings., International Conference on Image Processing*, volume 1, pages 187–190 vol.1, Oct. 1995. 2

[29] Jin Li and C.-C.J. Kuo. Coding artifact removal with multi-scale postprocessing. In *Proceedings of International Conference on Image Processing*, volume 1, pages 45–48 vol.1, Oct. 1997. 2

[30] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image Restoration Using Swin Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 5, 7

[31] Jiebo Luo, Chang Wen Chen, K.J. Parker, and T.S. Huang. Artifact reduction in low bit rate DCT-based image compression. *IEEE Transactions on Image Processing*, 5(9):1363–1368, Sept. 1996. 2

[32] Xiyang Luo, Hossein Talebi, Feng Yang, Michael Elad, and Peyman Milanfar. The Rate-Distortion-Accuracy Tradeoff: JPEG Case Study. *arXiv:2008.00605 [cs, eess]*, Aug. 2020. 5

[33] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo Exploration Database: New Challenges for Image Quality Assessment Models. *IEEE Transactions on Image Processing*, 26(2):1004–1016, Feb. 2017. 6

[34] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which Training Methods for GANs do actually Converge? In *Proceedings of the 35th International Conference on Machine Learning*, pages 3481–3490. PMLR, July 2018. 2, 4

[35] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets. *arXiv:1411.1784 [cs, stat]*, Nov. 2014. 3

[36] Guy Ohayon, Theo Adrai, Michael Elad, and Tomer Michaeli. Reasons for the Superiority of Stochastic Estimators over Deterministic Ones: Robustness, Consistency and Perceptual Quality. *arXiv:2211.08944 [cs, eess]*, Nov. 2022. 2, 3, 4

[37] Guy Ohayon, Theo Adrai, Gregory Vaksman, Michael Elad, and Peyman Milanfar. High Perceptual Quality Image Denoising With a Posterior Sampling CGAN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1805–1813, 2021. 2, 4, 5, 7

[38] T.P. O'Rourke and R.L. Stevenson. Improved image decompression for reduced transform coding artifacts. *IEEE Transactions on Circuits and Systems for Video Technology*, 5(6):490–499, Dec. 1995. 2

[39] T. Ozcelik, J.C. Brailean, and A.K. Katsaggelos. Image and video compression algorithms based on recovery techniques using mean field annealing. *Proceedings of the IEEE*, 83(2):304–316, Feb. 1995. 2

[40] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. 12, 13

[41] H. Reeve and Jae Lim. Reduction of blocking effect in image coding. In *ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 8, pages 1212–1215, Apr. 1983. 2

[42] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-Image Diffusion Models. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, pages 1–10, New York, NY, USA, July 2022. Association for Computing Machinery. 6

[43] Maximilian Seitzer. Pytorch-fid: FID score for PyTorch, Aug. 2020. Version 0.2.1. 5

[44] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006. 6, 15

[45] HR Sheikh, Z Wang, L Cormack, and AC Bovik. LIVE image quality assessment database. http://live.ece.utexas.edu/research/quality. 6, 15

[46] Mei-Yin Shen and C. C. Jay Kuo. Review of Postprocessing Techniques for Compression Artifact Removal. *Journal of Visual Communication and Image Representation*, 9(1):2–14, Mar. 1998. 2

[47] Richard Shin and Dawn Song. JPEG-resistant Adversarial Images. In *NIPS 2017 Workshop on Machine Learning and Computer Security*, volume 1, page 8, 2017. 5

[48] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, Apr. 2015. 4

[49] R.L. Stevenson. Reduction of coding artifacts in transform image coding. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 401–404 vol.5, Apr. 1993. 2

[50] Mengdi Sun, Xiaohai He, Shuhua Xiong, Chao Ren, and Xinglong Li. Reduction of JPEG compression artifacts based on DCT coefficients prediction. *Neurocomputing*, 384:335–345, Apr. 2020. 2

[51] Hossein Talebi, Damien Kelly, Xiyang Luo, Ignacio Garcia Dorado, Feng Yang, Peyman Milanfar, and Michael Elad. Better Compression With Deep Pre-Editing. *IEEE Transactions on Image Processing*, 30:6673–6685, 2021. 5

[52] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 114–125, 2017. 6

[53] G.K. Wallace. The JPEG still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1):xviii–xxxiv, Feb. 1992. 1, 3

[54] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training Real-World Blind Super-Resolution With Pure Synthetic Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1905–1914, 2021. 12, 13

[55] Xintao Wang, Liangbin Xie, Ke Yu, Kelvin C.K. Chan, Chen Change Loy, and Chao Dong. BasicSR: Open source image and video restoration toolbox, 2022. 12

[56] Zhangyang Wang, Ding Liu, Shiyu Chang, Qing Ling, Yingzhen Yang, and Thomas S. Huang. D3: Deep Dual-Domain Based Fast Restoration of JPEG-Compressed Images. In *2016 IEEE Conference on Computer Vision and*

Pattern Recognition (CVPR), pages 2764–2772, June 2016. 2

[57] S.-W. Wu and A. Gersho. Improved decoder for transform coding with application to the JPEG baseline system. *IEEE Transactions on Communications*, 40(2):251–254, Feb. 1992. 2

[58] Xiaoshuai Zhang, Wenhan Yang, Yueyu Hu, and Jiaying Liu. Dmcnn: Dual-Domain Multi-Scale Convolutional Neural Network for Compression Artifacts Removal. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 390–394, Oct. 2018. 2