Self-Supervised Effective Resolution Estimation with Adversarial Augmentations

Manuel Kansy[1,2] *, Julian Balletshofer[2], Jacek Naruniec[2], Christopher Schroers[2], Graziana Mignone[2], Markus Gross[1,2], and Romann M. Weber[2]

[1]Department of Computer Science, ETH Zurich, [2]DisneyResearch|Studios, Zurich

{manuel.kansy, grossm}@inf.ethz.ch, {<first name>.<last name>}@disneyresearch.com

Abstract

High-resolution, high-quality images of human faces are desired as training data and output for many modern applications, such as avatar generation, face super-resolution, and face swapping. The terms high-resolution and high-quality are often used interchangeably; however, the two concepts are not equivalent, and high-resolution does not always imply high-quality. To address this, we motivate and precisely define the concept of effective resolution *in this paper. We thereby draw connections to signal and information theory and show why baselines based on frequency analysis or compression fail. Instead, we propose a novel self-supervised learning scheme to train a neural network for effective resolution estimation without human-labeled data. It leverages adversarial augmentations to bridge the domain gap between synthetic and real, authentic degradations – thus allowing us to train on domains, such as human faces, for which no or only few human labels exist. Finally, we demonstrate that our method outperforms state-of-the-art image quality assessment methods in estimating the sharpness of real and generated human faces, despite using only unlabeled data during training.*

1. Introduction

Generative models that produce high-quality face images, such as human avatar generation, face super-resolution, and face swapping, require high-resolution, high-quality training data sets. Low-resolution images negatively affect the resulting models' output. It is therefore crucial to identify and remove these images prior to training. This seemingly simple task is challenging, however, since high-resolution does not always imply high-quality, so users must often resort to visual inspection.

Diagnosing the quality of a model's output can also be a challenging and subjective task. Consider, for example, a generative model in the form of a neural network that has been progressively trained (*e.g.* [15, 30]). It has been observed that such models can devote more effort to capturing details at lower resolutions, for example resulting in a 1024×1024 image with only minor detail improvements over a 512×512 image [17]. This scenario motivates our definition of *effective resolution*, given more precisely in Section 3, as the lowest-resolution image that is informationally equivalent to a given test image. Figure 1 visualizes different effective resolutions of a face image[1] and shows the scores assigned by our proposed model.

While effective resolution is an aspect of image quality, it refers to something more specific than what is typically covered by general image quality assessment methods [27, 28, 47, 36, 43, 48, 6, 18, 35], which predict scores corresponding to images' perceptual quality. Effective resolution can be affected by the process of capturing data (*e.g.* camera out-of-focus or downscaling) or by the subject of an image itself (*e.g.* content with low detail, such as a smooth, uniformly colored surface). We argue that estimating effective resolution is best handled by domain-aware methods, one of which we describe in this work.

We propose a self-supervised training scheme based on downscaling and upscaling images using random interpolation methods and having a network predict the (inverse of the) downscaling factor. We further propose using adversarial augmentations during training to help bridge the domain gap between training with synthetically degraded images and evaluating on real, authentically degraded images and to generalize to unseen faces. In contrast to most image quality assessment methods, our method does not require any human quality labels during training. To our knowledge, we are the first to propose adversarial augmentations in an image quality assessment context, and we show their usefulness in significantly boosting the performance as well as producing a more stable and meaningful gradient.

While we focus on images of human faces, the method

*Corresponding author.

[1]**Note:** All images in this paper are best viewed digitally and uncompressed as their effective resolutions are affected negatively otherwise.

r	512×512	512×512	512×512	512×512	512×512
r_{eff}	512×512	256×256	128×128	64×64	32×32
$\hat{r_{\text{eff}}}$	512×512	201×201	103×103	46×46	37×37

Figure 1: Visualization of different effective resolutions. We downscale an input image of absolute resolution $r = 512 \times 512$ to different resolutions r_{down} and upscale it back to r using bilinear interpolation. For each image, the absolute resolution $r = 512 \times 512$, the effective resolution $r_{\text{eff}} = r_{\text{down}}$, and $\hat{r_{\text{eff}}}$ denotes our model's predicted effective resolution. **Note:** We first downscaled a high-quality image to resolution 512×512, so that we can assume that $r = r_{\text{eff}} = 512$.

we describe is general and can easily be applied to other image domains. Among the applications of our approach is the automatic selection of high-quality training data; quality assessment of model output; a source of features for downstream tasks; an informative training metric, both as a progress monitor and as a differentiable loss; and even advice regarding architectural choices in cases in which an image distribution is of low average effective resolution relative to the initial modeling assumptions.

Our main contributions are:

1. Proposing the term *effective resolution* as a meaningful measure of image quality and defining it mathematically.
2. Developing a novel self-supervised training scheme using adversarial augmentations to train an effective resolution estimation network.
3. Demonstrating state-of-the-art performance in assessing the quality of human faces despite having no human labels during training.

2. Related work

The term *effective resolution* was mentioned in StyleGAN2 [17] in an intuitive way, but we are the first to define this term precisely in the context of computer vision to the best of our knowledge. Effective resolution is also used in the contexts of spatial resolution in X-ray microtomography [29], super-resolution microscopy [5], and digital photography (to define the resolution required to print an image on a specific physical paper size).

2.1. General image quality assessment

Effective resolution is related to the field of no-reference image quality assessment (IQA), which aims to predict the general quality of an image. Traditional IQA methods, like BRISQUE [27], NIQE [28], or IL-NIQE [47] using natural scene statistics or other hand-crafted features have been superseded by deep learning methods with the seminal work by Kang *et al*. [14] that uses a convolutional neural network to extract features. Bosse *et al*. [1] further showed that deep semantic features can boost the performance. To help train deeper networks and improve the performance on real degradations, large-scale labeled data sets such as CID2013 [38], LIVE Wild [7], PaQ-2-PiQ [43], KonIQ-10k [12], SPAQ [6], and PIPAL [13] were released. This led to the adoption of more sophisticated architectures such as the transformer [37] as seen in [44, 8, 18, 42, 41].

2.2. Blur-specific image quality assessment

Because our goal is to retain images with certain distortions, like difficult lighting or noise, general IQA methods are not suitable in this application. Classical blur detection methods often use spatial features such as edges [25, 31, 19], transform-based features [40, 10] or a combination thereof [39, 22, 21]. These methods often fail in complex, real world cases. Therefore, new algorithms are based on deep learning. Li *et al*. [20] discuss the dependence of sharpness on the semantic context and propose using deep semantic features. Recently, two large-scale data sets, *i.e*. SPAQ [6] and KonIQ++ [35], that include a label for specific distortions such as sharpness/blurriness along with the overall quality were introduced. The authors further propose methods that include this information during training to jointly predict specific distortions and the overall quality.

Some methods such as CONTRIQUE [23] and DB-CNN [48] incorporate synthetic data in their training to improve the performance. However, they still require labeled real images in some stage during training. Since labeled data sets are expensive to obtain and do not exist for certain domains, *e.g*. face images, we propose a

fully self-supervised method. Rather than relying on labels, we leverage synthetic distortions combined with adversarial noise, which is mostly used in the context of adversarial attacks [24], to achieve state-of-the-art results.

3. Effective resolution

Since the absolute resolution of an image, *i.e*. the number of pixels, does not always correlate with its perceived sharpness, we propose the term *effective resolution*.

Definition. *The effective resolution r_{eff} of an image x is*

$$r_{eff} \coloneqq \min_{u \in U, d \in D} \{\, r_{down} \mid u_{r_{down} \to r}(d_{r \to r_{down}}(x)) = x \,\} \quad (1)$$

where D is a class of nonparametric downscaling methods and U is a class of nonparametric upscaling methods.

In simple words, the effective resolution is the minimum resolution to which an image can be downscaled, such that upon upscaling it back to the original (absolute) resolution, the upscaled image is identical to the original image. Intuitively, at this minimum size, *i.e*. at an image's effective resolution, there exist at least some pixels that are used effectively to convey information, and the image appears sharp. In practice, the definition of effective resolution can be loosened to allow small, perceptually negligible differences between the original and down-upscaled image. Unless otherwise specified, we consider square images in order to simplify the notation of effective resolution into a single number, *i.e*. the side length.

Figure 2 visualizes our definition for an 8×8 image a of a 2×2 chessboard pattern. Since a 2×2 chessboard pattern can be fully specified with a 2×2 image, its effective resolution is 2.

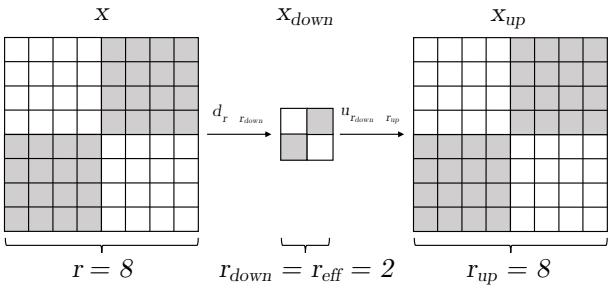


Figure 2: Visualization of the definition of effective resolution. An 8×8 image x of a 2×2 chessboard pattern has an effective resolution $r_{\text{eff}} = 2$ since it can be downscaled to a 2×2 image x_{down} and successively upscaled back to the same 8×8 resolution using nearest neighbor interpolation such that $x_{\text{up}} = x$. The upscaling operation only adds redundant information.

For images with the same content and absolute resolution, a lower effective resolution implies a lower image sharpness or, inversely, a higher blurriness. Note that the terms *sharpness* and *blurriness* are not exact inverses. For example, an in-focus image of a smooth, uniformly colored surface is neither sharp (because there are no details) nor blurry (because the camera is not out-of-focus, and there is no relative motion between camera and object). Nevertheless, in this paper, we assume that sharpness and blurriness are opposites for the domain of face images for the sake of simplicity.

Effective resolution is closely related to and can be investigated using ideas from signal and information theory as described in Section 5. To this end, we construct two baselines: one that considers the maximum frequency of an image and another one centered on the idea that a blurry image can be compressed more. We not only express theoretical shortcomings of the considered baselines but also demonstrate their poor performance compared to our proposed method.

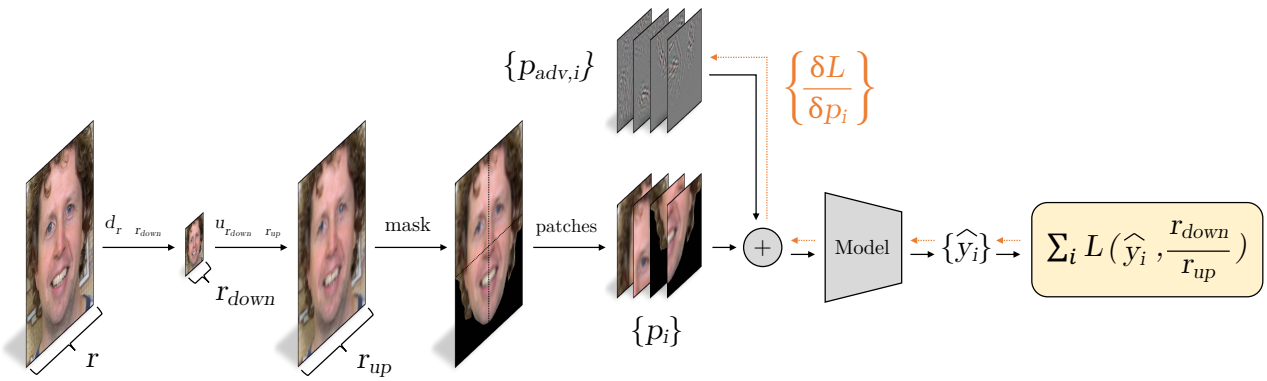
4. Method

We propose to use a neural network to estimate the effective resolution of an image since neural networks are content-aware and learn all relevant thresholds automatically rather than having to set them manually for each image domain. Additionally, due to our training scheme with adversarial augmentations, our model is robust to noise and thus generalizes to unseen degradations.

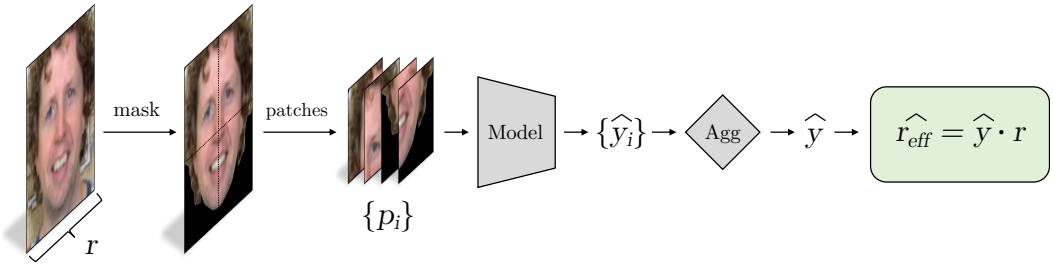
Our method is based on two key concepts: (1) a self-supervised training scheme that randomly down- and upscales images to obtain training samples and (2) a strategy of using adversarial augmentations during training to significantly boost the performance. Figure 3 shows an overview of our method architecture.

4.1. Self-Supervised training scheme

Most image quality assessment methods, even those obtaining image quality representations in a self-supervised manner [23, 48], rely on human labels as ground truth regression targets at some point during training. In contrast, our method works without any human labels by translating our definition of effective resolution from Section 3 into a training target. Specifically, we can generate a training sample by first downscaling an image with absolute resolution r using a random downscaling factor and interpolation method to an absolute resolution r_{down} and then upscaling the image back to the original absolute resolution $r_{\text{up}} = r$ using another random interpolation method. If the downscaling factor is sufficiently large, we can assume that $r_{\text{down}} = r_{\text{eff}}$. As described in Section 3, upscaling an image with a nonparametric upscaling method only adds redundancy, so the effective resolution remains constant when upscaling. The target effective resolution ratio y of the upscaled image is thus the inverse of the downscaling factor,



(a) Training: First, an original image x with a resolution of r is downscaled to a random resolution of r_{down} before being upscaled to a resolution of r_{up} where $r_{\text{up}} = r$. Then, the background is masked out, and patches are extracted. These patches are passed through the model to produce the predicted effective resolution ratio $\hat{y}_i$ for each patch p_i. The loss between the predicted patch-wise scores $\hat{y}_i$ and the target effective resolution ratio $\frac{r_{\text{down}}}{r_{\text{up}}}$ acts as training signal. Furthermore, in the training steps with adversarial augmentations, the gradient of the loss L with respect to the input patches p_i (visualized in orange) is used to produce patch-wise adversarial noise $p_{adv,i}$ which is then added to the input patches p_i before being input into the model.



(b) Inference: First, the background of an input image with a resolution r is masked out. Afterwards, the image is split into patches p_i, and each patch is passed through the model to obtain the patch-wise scores $\hat{y}_i$. These scores are then aggregated to the final score $\hat{y}$. Finally, the aggregated predicted effective resolution ratio $\hat{y}$ is multiplied with the input resolution r to obtain the predicted effective resolution $\hat{r}_{\text{eff}}$.

Figure 3: Method architecture.

$\frac{r_{\text{down}}}{r_{\text{up}}}$, and is regressed by a neural network.

Resizing an image changes the pixel information of an image and might thus destroy crucial quality information. Therefore, we propose to use patches similar to [20, 18] to handle images of varying and large absolute resolutions effectively. We assume that the downscaled training images are uniformly sharp spatially, so each patch inherits the same regression target, similar to [46]. During inference, the patch-wise scores $\hat{y}_i$ of an image are aggregated into one score $\hat{y}$ which is multiplied by the input resolution r to obtain the final predicted effective resolution $\hat{r}_{\text{eff}}$.

The background is masked out with an off-the-shelf method [45, 49] for two reasons. First, the background of face portraits is often blurry. This violates the assumption that downscaled training images are perfectly and uniformly sharp and would lead to faulty training targets for background patches for small downscaling factors. Second, we are only interested in the quality of the face, so the network should ignore the background during inference (which contains complex degradations for generated images).

4.2. Adversarial augmentations

During training, we sample the downscaling factor as well as the interpolation methods randomly for each training sample, so that the model sees a wide range of degradations. To further improve the robustness of the model to small pixel changes as well as to bridge the domain gap between training on artificial degradations and testing on degradations encountered in real or generated images, we propose to use adversarial augmentations during training. Our intuition is that the adversarial augmentations inter- and extrapolate between the different interpolation methods and thus ensure that the model does not overfit to specific interpolation methods. By perturbing the input image values slightly during training, the model sees many images that are all mapped to the same score and thus focuses less on tiny pixel value differences but rather on the overall quality. As the adversarial noise can actually influence the perceived sharpness of an image (see supplementary material), we constrain the strength of the adversarial noise during

training to ensure that it is mostly imperceptible to the human eye. This allows us to assume that a sample's target effective resolution ratio y remains constant.

Note that, while there are some similarities, our training scheme using adversarial augmentations is different from the adversarial training known from generative adversarial networks (GANs) [9]. Specifically, we only have a single neural network and use its own gradient to produce adversarial noise.

5. Experiments and results

As labeling the effective resolution is very difficult and no such labeled data set exists to our knowledge, we evaluate our method on the task of blurriness/sharpness estimation of human face images. We thereby compare to general and blur-specific image quality assessment methods and demonstrate state-of-the-art performance despite having no human labels during training.

5.1. Experimental details

5.1.1 Implementation details

Our network consists of a ResNet50 [11] as a backbone, followed by an average pooling and a fully connected layer. It is trained using mean absolute percentage error (MAPE) as loss function. To generate adversarial augmentations, we use projected gradient descent (PGD) attacks [24] with 10 PGD steps of size 30 in the L_2 norm. During inference, we aggregate the patch scores using the median.

Refer to the supplementary material for a complete explanation of the implementation details.

5.1.2 Training data

For training, we use a proprietary data set of 10777 high-quality human faces from 17 subjects. The images are cropped to the face area, leading to various resolutions (all over 1000×1000). Samples of the training data can be found in the supplementary material.

5.1.3 Evaluation data

For evaluation, we use real and generated images of four subjects that are not seen during training. The real images are similar to those used during training. They cover various types of settings (inside vs. outside, static vs. handheld) and degradations (out-of-focus blur, motion blur). The generated images were extracted at different times during the training of the face swapping method from [30] (trained with images from the complete data set for the same subjects) and all have a resolution of 512×512. We selected two folders with 10 images each per subject and category, resulting in a total number of 160 images. Samples of the evaluation data can be found in the supplementary material.

To obtain ground truth rankings according to human perception, we conducted an experiment with 15 test subjects. Since the quality of an image is difficult to label directly, we asked the subjects to perform pairwise comparisons, similar to [33, 32, 13]. For this, we asked the subjects to choose the sharper image while ignoring the background and degradations such as over-exposure and color shifts. For each batch of 10 images of the same person, we created 45 pairwise comparisons (which equals the number of comparisons that would be required to do one full pairwise comparison). We use the ASAP method [26] rather than doing the full pairwise comparisons because it provides more accurate ranking scores with the same number of comparisons. In total, test subjects performed 10800 pairwise comparisons. Afterwards, we compared the ratings of each subject to the average of the ratings from all of the other subjects to estimate the human performance. The bottom of Table 1 lists the mean and standard deviation of the subjects' ratings.

5.2. Baselines

5.2.1 Frequency baseline

As the resolution of an image determines its maximum representable frequencies, we can construct a "frequency baseline" which predicts the effective resolution based on signal theory using the Fourier transform [2]. According to the Nyquist-Shannon sampling theorem [34], downscaling an image with an ideal sinc filter removes frequencies below $f_{\text{max}} = \frac{r}{2}$ where r is the spatial resolution of the image. Upscaling an image with an ideal sinc filter does not add higher frequencies. Therefore, if we assume that a given blurry image x of resolution r can be obtained by upscaling an image of a smaller resolution, we can calculate the effective resolution as follows: $r_{\text{eff}} = 2 \cdot \hat{f_{\text{max}}}$ where $\hat{f_{\text{max}}}$ is the maximum frequency found in image x. We therefore extract the frequency at which the cumulative energy (*i.e.* squared sum) of all lower frequencies sums up to $(100 - \epsilon_c)\%$ of all the energy of all frequencies below the Nyquist frequency where ϵ_c is a hand-picked threshold and $\epsilon_c \gtrapprox 0$. Unless otherwise stated, we use $\epsilon_c = 0.005$ for the experiments as it results in the best validation performance.

5.2.2 Compression baseline

From an information-theoretic perspective, one can interpret the effective resolution in terms of the compressibility of an image. Blurrier images can be compressed more than sharp images, so effective resolution correlates positively with the compressed file size. We can thus construct a "compression baseline" using JPEG compression, where the score is the compressed file size of an image using OpenCV's [3] default JPEG quality setting of 95.

Method Type	Method	Generated		Real		All	
		SRCC↑	PRA↑	SRCC↑	PRA↑	SRCC↑	PRA↑
Baselines	Frequency baseline	0.3283	0.6132	0.6726	0.7450	0.5005	0.6791
	Compression baseline	0.6864	0.7614	0.7333	0.7876	0.7098	0.7745
Classic general	BRISQUE [27]	0.8424	0.8466	0.5985	0.7232	0.7205	0.7849
	NIQE [28]	0.8773	0.8751	0.3348	0.6283	0.6061	0.7517
	IL-NIQE [47]	0.7061	0.7755	0.3409	0.6212	0.5235	0.6983
Deep learning general	NIMA [36]	0.7636	0.8097	0.5000	0.6811	0.6318	0.7454
	PaQ-2-PiQ [43]	0.6606	0.7535	0.6848	0.7642	0.6727	0.7588
	DB-CNN [48]	0.6939	0.7810	0.6500	0.7465	0.6720	0.7637
	MUSIQ (PaQ-2-PiQ) [18, 43]	0.7758	0.8155	<u>0.9227</u>	<u>0.9124</u>	0.8492	0.8640
	MUSIQ (SPAQ) [18, 6]	0.7061	0.7845	0.6333	0.7464	0.6697	0.7654
	MUSIQ (KonIQ-10k) [18, 12]	0.8545	0.8527	0.8788	0.8825	0.8667	0.8676
Classic blur-specific	CPBD [31]	0.6909	0.7780	0.5424	0.7064	0.6167	0.7422
	Δ DOM [19]	**0.9500**	**0.9321**	0.8288	0.8422	<u>0.8894</u>	<u>0.8871</u>
Deep learning blur-specific	SFA [20]	0.5076	0.6921	0.8121	0.8364	0.6598	0.7643
	MT-A [6]	0.5470	0.6986	0.8106	0.8276	0.6788	0.7631
	KonIQ++ [35]	0.7742	0.8131	**0.9364**	**0.9303**	0.8553	0.8717
Ours	Ours (patch size 256)	<u>0.9258</u>	<u>0.9120</u>	0.9045	0.8847	<u>0.9152</u>	<u>0.8984</u>
	Ours (patch size 128)	<u>0.9318</u>	<u>0.9263</u>	<u>0.9076</u>	<u>0.8952</u>	**0.9197**	**0.9107**
Ground truth	Mean	0.9407	0.9250	0.9466	0.9317	0.9436	0.9284
	Standard deviation	0.0256	0.0213	0.0186	0.0178	0.0172	0.0151

Table 1: Evaluation results (masked). SRCC denotes the Spearman rank correlation coefficient, and PRA denotes the pairwise ranking accuracy. The best score per column is marked in bold, the second- and third-best are underlined.

5.3. Comparison with state-of-the-art methods

We compare our proposed method to general as well as blur-specific state-of-the-art image quality assessment methods. For BRISQUE [27], NIQE [28], IL-NIQE [47], NIMA [36], PaQ-2-PiQ [43], DB-CNN [48], and MUSIQ [18], we use the implementations provided in [4]. For all other methods, *i.e.* for MT-A [6], KonIQ++ [35], CPBD [31], Δ DOM [19], and SFA [20], we use the official implementations and parameters. To ensure the fairness of the evaluation, the background is masked out for all methods since it generally improves the performance. Refer to the supplementary material for the unmasked results.

Table 1 lists the results of all of the methods on our evaluation data for generated images, real images, and overall. We evaluate the methods on the Spearman rank correlation coefficient (SRCC) and the pairwise ranking accuracy (PRA). The SRCC measures the strength of the monotonic association between the method's scores and the ground truth and is in the range $[-1, 1]$ where 1 is a perfect positive correlation. The PRA is the ratio of pairwise comparisons that the method ranks the same as the test subjects. The score range is $[0, 1]$, where a PRA of 1 is the best, and a PRA close to 0.5 indicates that a method is nearly random. Both metrics are independent of the score range and linearity, so no score modification or retraining is necessary. When calculating the above metrics for methods that output the blurriness, we negate the method's output, so that a higher score always refers to better quality.

The frequency baseline does not work well in practice because common interpolation filters approximate the ideal sinc filter very poorly. As an extreme example, when upscaling using nearest neighbor interpolation, pixels are simply repeated, leading to very high frequencies at the original pixel boundaries. Figure 4 shows that the predicted effective resolution of the frequency baseline (using $\epsilon_c = 0.001$) is too high and does not decrease monotonically with larger downscaling factors.

The compression baseline performs better than the frequency baseline but has the major shortcoming that the compressibility is heavily influenced by the content. Therefore, an image of a man with a beard will likely have a larger file size than an image of a shaved man of similar quality.

Our proposed network outperforms all methods overall by more than 2% and lags less than two standard deviations behind the estimated human performance. It performs especially well on generated images, likely because the ad-

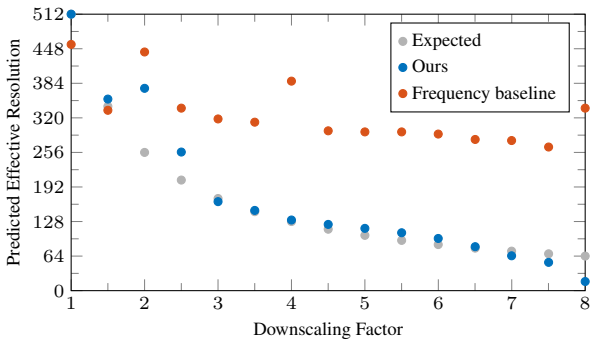


Figure 4: Monotonicity plot for the frequency aseline. The plot shows the predicted effective resolution when downscaling an image using area interpolation and upscaling using nearest neighbor interpolation for different downscaling factors. The frequency baseline (orange, using $\epsilon_c = 0.001$) results in values that are too high and not monotonous, especially as the downscaling factor increases, since nearest neighbor upscaling produces many high frequencies at the original pixel boundaries. In comparison, our proposed network (blue) follows the expected values (gray) much more closely. **Note:** We first downscaled a high-quality image to resolution 512×512 (masked image from Figure 3), so that we can assume that $r = r_{\text{eff}} = 512$.

versarial augmentations lead to a model that is more robust to unseen degradations. It is noteworthy that Δ DOM [19] only outperforms our method when we mask out the background, and its performance degrades tremendously otherwise whereas our method's performance only degrades slightly. Furthermore, the deep learning methods that outperform our method on real images all require labeled data while our method is trained in a completely self-supervised manner and uses a smaller backbone.

5.4. Ablation study

Table 2 shows an ablation of the most important training parameters. Refer to the supplementary material for the complete table. Note that all ablations are performed with a patch size of 256.

5.4.1 Training data set

As our proposed method does not require any labels, we are flexible in choosing the training data set. It appears beneficial to train on a data set that contains many subjects as the model trained on a small subset of the data that contains all subjects performs better than the one trained with only one person but the same number of training images. Our proposed method is fairly robust to low-quality images in the training data set as seen in the relatively small degradation

Category	Setting	All	
		SRCC↑	PRA↑
Training data set	One person	0.8280	0.8446
	Small	0.8803	0.8783
	Adding blurry images	0.9023	0.8896
	Subset of FFHQ [16]	<u>0.9053</u>	0.8924
Adversarial method	None	0.2083	0.5765
	Step size $L_2 = 3$ (*/10)	0.8879	0.8828
	Step size $L_2 = 300$ (* · 10)	0.8098	0.8309
Unmasked	Unmasked training	0.8636	0.8740
	Unmasked inference	<u>0.9129</u>	0.8954
	Unmasked training + inf.	0.8977	<u>0.8968</u>
Pre-processing	Only bicubic interp.	<u>0.9053</u>	<u>0.8956</u>
	Only nearest neighb. interp.	0.7461	0.4512
Ours	Ours (patch size 256)	**0.9152**	**0.8984**

Table 2: Ablation study results (reduced). SRCC denotes the Spearman rank correlation coefficient, and PRA denotes the pairwise ranking accuracy. The best score per column is marked in bold, the second- and third-best are underlined. * indicates the value of the parameter in "Ours (patch size 256)".

in performance when adding around 10% blurry images to the data set and when training on a similarly-sized subset of FFHQ [16] (first 10000 images) that contains many degraded images.

5.4.2 Adversarial method

Without adversarial augmentations, the performance becomes nearly random as the model can overfit to the specific interpolation methods encountered in training. This effect is shown in Figure 5, where even just rounding the pixel values (in the range $[0, 255]$) to the nearest integer leads the model trained without adversarial augmentations to perform very unexpectedly and actually predict higher scores as an image is downscaled more. The fact that the unrounded version looks reasonable also demonstrates that a monotonicity experiment, where a single image is degraded with different strengths, can only really indicate that a method will perform poorly on real images but not that it will perform well. Simply changing the training procedure to also round the pixel values leads to better results for this experiment but does not improve the performance on real images significantly. We hypothesize that adversarial augmentations have such a big impact in our training scheme since the network is tasked to look at the small pixel differences that affect the sharpness of an image rather than the overall content of the image, so tiny changes in the pixel values influence the

score significantly.

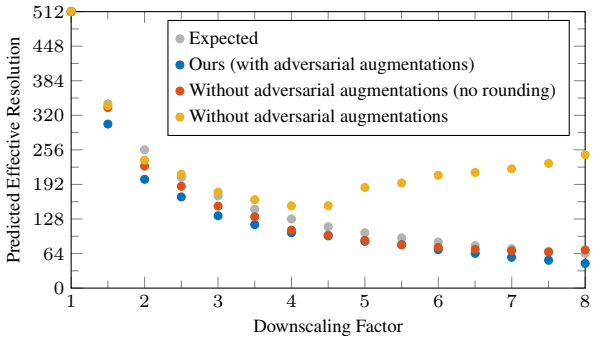


Figure 5: Monotonicity plot for the model without adversarial augmentations. The plot shows the predicted effective resolution when down- and upscaling an image using bilinear interpolation for different downscaling factors. The model trained without adversarial augmentations follows the expected values (gray) closely only if the image values in the range $[0, 255]$ are not rounded to the nearest integer (orange). Otherwise, the predicted effective resolution actually increases beyond a certain downscaling factor. In comparison, our proposed network (blue) performs well with the rounding since it is much more robust to small pixel value differences. **Note:** We first downscaled a high-quality image to resolution 512×512 (masked image from Figure 3), so that we can assume that $r = r_{\text{eff}} = 512$.

Whereas the adversarial noise generated for a non-adversarially trained model appears arbitrary and relatively uniformly distributed, the adversarial noise generated for an adversarially trained model is very meaningful and shows the face structures that influence the sharpness the most, as seen in Figure 6 and the supplementary material. There is a trade-off for the strength of the adversarial augmentations. If the adversarial augmentations are too weak, the model is not robust enough. On the other hand, if the adversarial augmentations are too strong and actually change the perceived sharpness of an image, the network does not receive a clear, consistent training signal since the training assumes the label of an image does not change due to the adversarial noise. This can even lead to training collapse.

5.4.3 Unmasked

Not masking out the background during training degrades the performance more than not masking it out during inference. Interestingly, if the background is not masked out during inference, the model is better if it was trained with masks rather than without despite having a small domain gap between training and evaluation images. We hypothesize that masking during training helps since it ensures

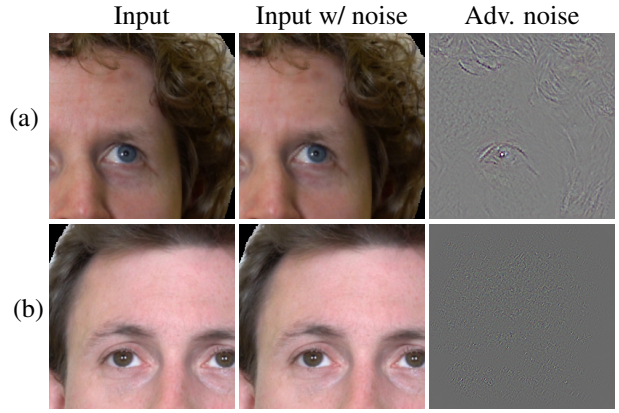


Figure 6: Visualization of the adversarial noise during training. The adversarial noise of our proposed method (a) is meaningful and captures image structures that heavily influence the sharpness whereas the adversarial noise of the model trained without adversarial augmentations (b) appears random.

that the assumption of having sharp training images after the downscaling operation is met.

5.4.4 Preprocessing

When using only bicubic interpolation during training, the performance only degrades slightly, showcasing the usefulness of the adversarial noise in simulating real degradations. However, when using only nearest neighbor interpolation, the performance degrades significantly. We hypothesize that the domain gap between the block artifacts from nearest neighbor upscaling and real degradations is too large to be covered using adversarial augmentations alone.

6. Conclusion

In this paper, we motivate and define the term *effective resolution* that correlates with image sharpness and intuitively defines the minimal resolution to which we can downscale an image without loss of information. We propose a self-supervised training scheme to predict the effective resolution of an image based on downscaling and upscaling images. Thereby, we leverage adversarial augmentations to boost the performance tremendously. This motivates further research in using adversarial noise in the field of image quality assessment to improve generalization as well as in using the resulting meaningful gradient for downstream tasks. Our method achieves state-of-the-art performance on a data set with realistic and generated face images despite only using unlabeled data during training, demonstrating the effectiveness of the proposed approach. While we focus on face images, our method is generic in nature and transfers to other domains.

References

[1] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219, 2017.

[2] Ronald Newbold Bracewell and Ronald N Bracewell. *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York, 1986.

[3] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[4] Chaofeng Chen. Pytorch toolbox for image quality assessment, 6 2022.

[5] A Descloux, Kristin Stefanie Grußmayer, and Aleksandra Radenovic. Parameter-free image resolution estimation based on decorrelation analysis. *Nature methods*, 16(9):918–924, 2019.

[6] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3677–3686, 2020.

[7] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015.

[8] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1220–1230, 2022.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[10] Rania Hassen, Zhou Wang, and Magdy MA Salama. Image sharpness assessment based on local phase coherence. *IEEE Transactions on Image Processing*, 22(7):2798–2810, 2013.

[11] 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] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020.

[13] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *European Conference on Computer Vision*, pages 633–651. Springer, 2020.

[14] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1733–1740, 2014.

[15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[18] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021.

[19] Jayant Kumar, Francine Chen, and David Doermann. Sharpness estimation for document and scene images. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 3292–3295. IEEE, 2012.

[20] Dingquan Li, Tingting Jiang, Weisi Lin, and Ming Jiang. Which has better visual quality: The clear blue sky or a blurry animal? *IEEE Transactions on Multimedia*, 21(5):1221–1234, 2018.

[21] Leida Li, Weisi Lin, Xuesong Wang, Gaobo Yang, Khosro Bahrami, and Alex C Kot. No-reference image blur assessment based on discrete orthogonal moments. *IEEE transactions on cybernetics*, 46(1):39–50, 2015.

[22] Leida Li, Wenhan Xia, Weisi Lin, Yuming Fang, and Shiqi Wang. No-reference and robust image sharpness evaluation based on multiscale spatial and spectral features. *IEEE Transactions on Multimedia*, 19(5):1030–1040, 2016.

[23] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31:4149–4161, 2022.

[24] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[25] Pina Marziliano, Frederic Dufaux, Stefan Winkler, and Touradj Ebrahimi. A no-reference perceptual blur metric. In *Proceedings. International conference on image processing*, volume 3, pages III–III. IEEE, 2002.

[26] Aliaksei Mikhailiuk, Clifford Wilmot, Maria Perez-Ortiz, Dingcheng Yue, and Rafał K Mantiuk. Active sampling for pairwise comparisons via approximate message passing and information gain maximization. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2559–2566. IEEE, 2021.

[27] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

[28] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[29] R Mizutani, R Saiga, S Takekoshi, C Inomoto, N Nakamura, M Arai, K Oshima, M Itokawa, A Takeuchi, K Uesugi, Y Terada, and Y Suzuki. Estimating the resolution of real images. *Journal of Physics: Conference Series*, 849:012042, jun 2017.

[30] Jacek Naruniec, Leonhard Helminger, Christopher Schroers, and Romann M Weber. High-resolution neural face swapping for visual effects. In *Computer Graphics Forum*, volume 39, pages 173–184. Wiley Online Library, 2020.

[31] Niranjan D Narvekar and Lina J Karam. A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). *IEEE Transactions on Image Processing*, 20(9):2678–2683, 2011.

[32] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30:57–77, 2015.

[33] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, Marco Carli, and Federica Battisti. Tid2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10(4):30–45, 2009.

[34] Claude E Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

[35] Shaolin Su, Vlad Hosu, Hanhe Lin, Yanning Zhang, and Dietmar Saupe. Koniq++: Boosting no-reference image quality assessment in the wild by jointly predicting image quality and defects. In *The 32nd British Machine Vision Conference*, volume 2, 2021.

[36] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[38] Toni Virtanen, Mikko Nuutinen, Mikko Vaahteranoksa, Pirkko Oittinen, and Jukka Häkkinen. Cid2013: A database for evaluating no-reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 24(1):390–402, 2014.

[39] Cuong T Vu, Thien D Phan, and Damon M Chandler. $\mathbf{s}_3$: a spectral and spatial measure of local perceived sharpness in natural images. *IEEE transactions on image processing*, 21(3):934–945, 2011.

[40] Phong V Vu and Damon M Chandler. A fast wavelet-based algorithm for global and local image sharpness estimation. *IEEE Signal Processing Letters*, 19(7):423–426, 2012.

[41] Jing Wang, Haotian Fan, Xiaoxia Hou, Yitian Xu, Tao Li, Xuechao Lu, and Lean Fu. Mstriq: No reference image quality assessment based on swin transformer with multi-stage fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1269–1278, 2022.

[42] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022.

[43] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3575–3585, 2020.

[44] Junyong You and Jari Korhonen. Transformer for image quality assessment. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1389–1393. IEEE, 2021.

[45] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[46] Ning Yu, Xiaohui Shen, Zhe Lin, Radomir Mech, and Connelly Barnes. Learning to detect multiple photographic defects. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1387–1396. IEEE, 2018.

[47] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015.

[48] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2018.

[49] zllrunning. face-parsing.pytorch, 10 2019.