EGIC: Enhanced Low-Bit-Rate Generative Image Compression Guided by Semantic Segmentation

Nikolai Körber[1,2], Eduard Kromer[2], Andreas Siebert[2], Sascha Hauke[2], Daniel Mueller-Gritschneder[3], and Björn Schuller[1]

[1] Technical University of Munich, Munich, Germany
{nikolai.koerber, schuller}@tum.de
[2] University of Applied Sciences Landshut, Landshut, Germany
{eduard.kromer, andreas.siebert, sascha.hauke}@haw-landshut.de
[3] TU Wien, Vienna, Austria
daniel.mueller-gritschneder@tuwien.ac.at

Abstract. We introduce EGIC, an enhanced generative image compression method that allows traversing the distortion-perception curve efficiently from a single model. EGIC is based on two novel building blocks: i) OASIS-C, a conditional pre-trained semantic segmentation-guided discriminator, which provides both spatially and semantically-aware gradient feedback to the generator, conditioned on the latent image distribution, and ii) Output Residual Prediction (ORP), a retrofit solution for multi-realism image compression that allows control over the synthesis process by adjusting the impact of the residual between an MSE-optimized and GAN-optimized decoder output on the GAN-based reconstruction. Together, EGIC forms a powerful codec, outperforming state-of-the-art diffusion and GAN-based methods (*e.g.*, HiFiC, MS-ILLM, and DIRAC-100), while performing almost on par with VTM-20.0 on the distortion end. EGIC is simple to implement, very lightweight, and provides excellent interpolation characteristics, which makes it a promising candidate for practical applications targeting the low bit range.

Keywords: Generative Image Compression · Transformer · GANs

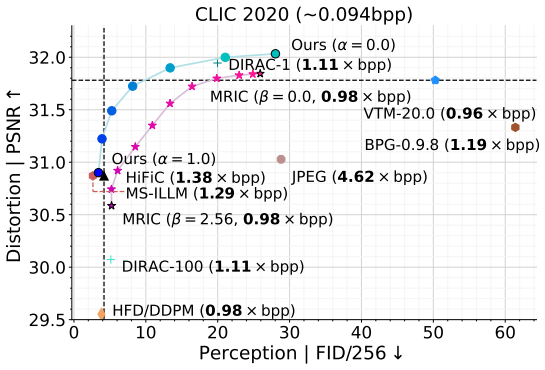

Fig. 1: Distortion-perception comparison (top left is best)

1 Introduction

Neural image compression methods incorporating generative models (*e.g.*, Generative Adversarial Networks, short GANs [18]) have been able to achieve comparable perceptual quality at considerably lower bit-rates [3, 37], hence being a promising direction for storage-efficient and bandwidth-constrained applications. Their underlying principle is that missing information can be realistically synthesized (*e.g.*, textures), therefore allowing more control over highly-sensitive information. Formally, these methods fall into the category of lossy compression with high perception [7, 65, 69], *i.e.*, we are interested in the lowest possible distortion for a given bit-rate with the constraint that the reconstructions follow the underlying data distribution. Note that this definition implies that low distortion alone does not per se yield good perceptual quality. In fact, it has been shown that perception and distortion are at odds with each other [7].

With the rise of diffusion models [14], there has been increasing efforts to rival GANs in the context of generative image compression [17, 24, 57, 66]. While these models have garnered attention for their better training dynamics and for their competitive or even higher image sample quality, they often come at considerably increased computational cost and inference latency. For example, the denoising network DIRAC-n [17] requires additional 108.4M parameters and up to $n = 100$ sampling steps compared to the base codec. In HFD/DDPM [24], Hoogeboom *et al.* even considered a larger setup, using more than 1B additional model parameters and up to 250 sampling steps. In this work, we challenge the prevailing belief in the superiority of diffusion models over GANs for generative image compression.

We propose EGIC, an **E**nhanced **G**enerative **I**mage **C**ompression method. EGIC is based on a novel conditional pre-trained semantic segmentation-guided discriminator (OASIS-C), which provides both spatially and semantically-aware gradient feedback to the generator, conditioned on the latent image distribution. Semantic segmentation-guided discriminators have been originally proposed for semi-supervised semantic segmentation [56] and later been adopted for the task of semantic image synthesis [54]. Both fields have played an important role for generative image compression early on. For example, the extreme learned image compression method GC (D^+) [3] largely builds upon the success of pix2pixHD [25, 60], a powerful image-to-image translation method. In [3], (D^+) further refers to a specific configuration, in which the discriminator D is presented with semantic labels as additional side information. Back then, however, only restrictive image domains were considered (the Cityscapes dataset [12]). In this work, we go one step further and demonstrate that semantic segmentation-guided discriminators, through careful design, can also considerably boost the generative compression performance across the image domain.

A common criticism of generative image compression methods is the lack of transparency in the underlying generation process. As pointed out by Agustsson *et al.* [1], users might worry that the reconstructions deviate too much from the original image. Multi-realism image compression algorithms try to address this by providing the user with a choice: from a single compressed repre-

sentation, we can either obtain a reconstruction that resembles more the traditional compression setting (low distortion), a visually appealing reconstruction (high perception), or anything in between. Unfortunately, existing solutions [1, 15, 17, 26, 27, 61, 69] typically come at the expense of considerably increased model size, decoding latency and/ or reduced overall performance.

We propose Output Residual Prediction (ORP), an efficient retrofit solution for multi-realism image compression. ORP is inspired by recent theoretical findings that simple image interpolation between an MSE-optimized decoder and a perfect perceptual decoder is sufficient to achieve any point on the distortion-perception (D-P) curve [69]. The main idea is to (implicitly) predict the residual R between an MSE-optimized and GAN-optimized decoder output, which allows control over the synthesis process by adjusting the impact of the residual ($\alpha \in [0, 1]$) on the GAN-based reconstruction, see Fig. 1. Compared to existing solutions, ORP only requires a fraction of additional model parameters (*e.g.*, $0.15\times$ compared to MRIC [1]) and only requires a single inference-cycle (as opposed to DIRAC-n [17]), revealing that more sophisticated methods may in fact be unnecessary. In summary our contributions are:

1. We introduce EGIC, a novel generative image compression method that allows traversing the D-P curve efficiently from a single model. EGIC is based on two core building blocks (Sec. 4): i) OASIS-C, a conditional pre-trained semantic segmentation-guided discriminator, and ii) ORP, a lightweight retrofit solution for multi-realism image compression.
2. We conduct a thorough study to identify suitable discriminator architectures/ GAN formulations for the task of generative compression (Sec. 5).
3. We empirically evaluate the effectiveness of our method on both convolutional (HiFiC [37]) and transformer-based (SwinT-ChARM [72]) backbones, on three challenging benchmark datasets (Sec. 6). We find that EGIC is particularly well-suited for the low bit range. On the perception end, EGIC outperforms a wide-variety of diffusion and GAN-based methods (*e.g.*, HiFiC, MS-ILLM, DIRAC-100), while being considerably more storage-efficient (*e.g.*, $0.03\times$ model parameters compared to HFD/DDPM). On the distortion end, EGIC almost matches VTM-20.0, the state-of-the-art for non-learned image codecs, while providing excellent interpolation characteristics for all other operating modes in between.

2 Related Work

Generative image compression. Agustsson *et al.* [3] demonstrated that an extreme learned image compression method combined with a multi-scale PatchGAN discriminator [25] can achieve compression rates far beyond the prior state-of-the-art while maintaining similar perceptual quality. Their work was later refined and extended by a hyper-prior [6], formally known as HiFiC [37]. In MRIC [1], the authors further pushed the rate-distortion-perception frontier by incorporating more powerful building blocks into their system [20, 38].

Yan *et al*. [65] proposed an allegedly optimal training framework that achieves the lowest possible distortion under the perfect perception constraint for a given bit-rate. Essentially, the authors state that a perceptual decoder can be trained using solely a GAN conditioned on an encoder optimized under the traditional rate-distortion objective. In [65], WGAN-GP [5,19] is employed using the vanilla concatenation-based conditioning scheme presented in [39]. While theoretically appealing, the authors have only been able to demonstrate superior performance on the MNIST dataset. Their ideas were later refined in [69], but still did not reach the performance of HiFiC. From both works, it appears that their success is highly dependent on the underlying conditional GAN framework; it is interesting to note that most current works [1, 21, 37, 41, 65, 69] use a concatenation-based conditioning scheme, which is known to be inferior to projection [40].

Although there are numerous other works [7,21,26,48,51,59,65,69], we argue that the fundamental GAN principles have barely changed[4]. For example, in PO-ELIC, a recent work, He *et al*. [21] use the same PatchGAN discriminator architecture as in HiFiC and MRIC, but with hinge-loss. Recent advances in the field of generative image compression can therefore mainly be attributed to improved building blocks, but not to more powerful generative models.

An exception to this line of work are diffusion-based methods [17, 24, 57, 66]. Diffusion models have recently rivaled GANs [14], often achieving competitive or higher image sample quality. While this direction is promising, their practical use is currently hindered by the high computational cost.

Finally, a common criticism of generative image compression methods is the lack of transparency in the underlying generation process. As pointed out by Agustsson *et al*. [1], users might worry that the reconstructions deviate too much from the original image. However, this concern is not a limitation in general and can be addressed via universal rate-distortion-perception representations [70]. Existing solutions have considered image/ weight interpolation [26, 61, 64, 69], denoising diffusion probabilistic models for residual prediction [17] and loss-conditional training [1, 15, 27].

Semantic image synthesis/ generative models. Semantic image synthesis has played an important role in generative image compression from the very beginning [3, 60]. While its use has been primarily advertised for constrained application domains with semantic label maps available (*e.g*., the Cityscapes dataset [12]), recent work [4, 45, 54] as well as better semantic segmentation models [10, 11, 63], show its generation ability across the image domain. Of particular importance to us are semantic segmentation-guided discriminators, which have been originally proposed for semi-supervised semantic segmentation [56] and later been adopted for the task of semantic image synthesis [54]. The general idea is to convert the discriminator to a multi-class classifier, where the additional classes correspond to regular semantic labels.

Other promising candidates we consider in this work are the SESAME [42], the U-Net [53], and the projected discriminators [52]. The SESAME discrimina-

[4] An exception is MS-ILLM [41], a concurrent work which we became aware of only during the completion of this work. We provide a short comparison in Sec. 4.

tor has been introduced as a multi-scale and improved variant of the PatchGAN discriminator, while the use of U-Net and projected discriminators have led to significant advances over BigGAN [8] and StyleGAN [29], respectively, arguably the two most popular GAN families.

Another interesting line of work are frequency-aware GANs [16,28,55]. These methods are based on the observation that the statistics of GAN-generated images often differ considerably from real images in the frequency domain. In this work, we use the Focal Frequency Loss (FFL) [28] as a tool to quantify the frequency awareness of each method.

3 Background

Traditional rate-distortion trade-off. We follow the same notation as in previous works [37]: a neural image compression method consists of three components, an encoder E, a decoder G (hereafter referred to as generator) and an entropy model P. Specifically, E encodes x to a quantized latent representation $y = E(x)$, while G creates a reconstruction of the original image $x' = G(y)$. The learning objective is to minimize the rate-distortion trade-off [13], with $\lambda > 0$:

$$\mathcal{L}_{RD} = \mathbb{E}_{x\sim p_X}[\lambda r(y) + d(x, x')]. \tag{1}$$

In Eq. (1), the bit-rate is estimated using the cross entropy $r(y) = -\log P(y)$, where P represents a probability model of y and $d(x, x')$ is a full-reference metric. In practice, an entropy coding method based on P is used to obtain the final bit representation, *e.g.*, using adaptive arithmetic coding. For a more general overview of neural compression, we refer the interested reader to [68].

Rate-distortion-perception trade-off. In Mentzer *et al.* [37], a discriminator D is further added to navigate the triple trade-off [7] using the non-saturating loss [18]:

$$\mathcal{L}_{RDP} = \mathbb{E}_{x\sim p_X}[\lambda r(y) + d(x, x') - \beta \log(D(x', y))], \tag{2}$$

$$\mathcal{L}_{disc} = \mathbb{E}_{x\sim p_X}[-\log(1 - D(x', y))] + \mathbb{E}_{x\sim p_X}[-\log(D(x, y))]. \tag{3}$$

In Eq. (2), $d(x, x')$ is decomposed into $d = k_M\text{MSE} + k_P\text{LPIPS}$ [71], where k_M and k_P are hyper-parameters. We keep this formulation to make use of the same hyper-parameters as in HiFiC. It is worth noting that the discriminator D is conditioned on y, identical to the formulation under the optimal training framework [65]. In both lines of work, a concatenation-based conditioning scheme [39] is chosen to model $P_{X|Y}$. We will study this design decision later on.

4 Our Approach

Learning objective. Inspired by recent advances in the field of semi-supervised semantic segmentation/ semantic image synthesis [54, 56], we redesign the discriminator to a $(N+1)$-class semantic segmentation task:

$$\mathcal{L}_{ours} = \mathbb{E}_{x\sim p_X}[\lambda r(y) + d(x, x') + \beta \mathcal{L}_{wce}(x', y)], \tag{4}$$

$$\mathcal{L}_{seg} = \mathbb{E}_{x\sim p_X}[\mathcal{L}_{wce}(x, y)] + \mathbb{E}_{x\sim p_X}[- \sum_{i,j}^{H\times W} \log D(x', y)_{i,j,c_{ij}=N+1}]. \tag{5}$$

In our formulation, $D \in \mathbb{R}^{H\times W\times N+1}$ represents a probability distribution over all semantic classes $\{1, .., N+1\}$, with $N+1$ being the fake label. Note that D is conditioned on y following theoretical and empirical results in [37, 65]. $\mathcal{L}_{wce}(x, y) = -\sum_{i,j}^{H\times W} w_{ij} \log D(x, y)_{i,j,c_{ij}}$ denotes the weighted $(N+1)$-class cross entropy loss over all pixel locations $(i, j) \in H \times W$, where c_{ij} is the index of the prediction for the correct semantic class and w_{ij} is a pixel weighting scheme. Different from the approach in [54], however, we employ the more commonly used pixel loss weighting scheme presented in [67], which puts more emphasis on small instances ($w_{ij} = 3$ for area size smaller than 64×64 px, $w_{ij} = 1$ everywhere else). This change is primarily due to practical considerations which will be motivated later on.

Similar to the non-saturating loss, G tries to fool D by generating realistic and semantically correct reconstructions, whereas D tries to differentiate between x and x'. This is essentially achieved by assigning the fake label as correct semantic class ($c_{ij} = N+1$).

We additionally regularize the discriminator in Eq. (5) with the LabelMix (LM) consistency loss [54], adapted to the compression setting:

$$\mathcal{L}_{cons} = \|D_{\text{logits}}(\text{LM}(x, x', M), y) - \text{LM}(D_{\text{logits}}(x, y), D_{\text{logits}}(x', y), M)\|_2^2, \tag{6}$$

with $\text{LM}(x, x', M) = M \odot x + (1 - M) \odot x'$. In Eq. (6), M is a randomly generated binary mask that respects the underlying semantic boundaries of x and $\text{LM}(x, x', M)$ corresponds to the resulting mixed real-fake image. The discriminator predictions are constrained to be equivariant under the LM operation, *i.e.*, the discriminator prediction of the mixed image ($D_{\text{logits}}(\text{LM}(x, x', M), y)$) should be identical to the mixed discriminator predictions of the real and fake images, respectively ($\text{LM}(D_{\text{logits}}(x, y), D_{\text{logits}}(x', y), M)$), thus forcing the discriminator to focus more on content and structure. This is essentially achieved by applying the L2 norm on the unnormalized discriminator predictions D_{logits}. We use a fixed LM coefficient of 10 for all experiments.

Our approach shares some similarities with the discriminator presented in MS-ILLM [41], with the main difference being the labels. We directly employ human-annotated semantic labels, whereas Muckley *et al.* propose to replace these with codebook indices from a pre-trained VQ-VAE [44,47] model. Our work is motivated by recent findings that pixel-level supervision of the discriminator is crucial in obtaining artifact-free images [55]. Codebook entries of VQ-VAEs on the other hand refer to patch-based supervision (32×32); the codebook size also typically exceeds the number of semantic labels by at least one order of magnitude (*e.g.*, 134 vs 1024), which can be more challenging to train. Both

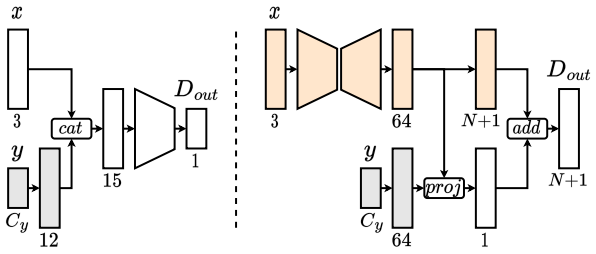

Fig. 2: Schematic comparison between PatchGAN (l.) and OASIS-C (r.)

approaches do not require labels during inference and can thus be considered as an enhanced version of the GC (D^+)-variant introduced in [3].

Training strategies. We employ a two-stage training strategy. In the first stage, we use Eq. (4) without adversarial supervision (*i.e.*, $\mathcal{L}_{RD} = \mathbb{E}_{x\sim p_X}[\lambda r(y) + d(x, x')]$), using the same configuration as in the original work [37]. For the second stage, we consider two variants: strategy-I denotes the full learning objective as described in Eq. (4) and Eq. (5), whereas strategy-II is based on pure adversarial supervision inspired by [65]. In both cases, we only fine-tune G and fix the pre-trained E, P from stage one. The latter is motivated by the observation that an encoder trained under the traditional rate-distortion optimization is also well-suited for the perceptual compression task [65]. The outputs of our training procedure are shared weights for E, P, G_1, and G_2, from stages one and two.

OASIS-C. In this section, we introduce OASIS-C, our novel conditional pre-trained semantic segmentation-guided discriminator. The name OASIS-C encapsulates both its origin (we base our work on the U-Net architecture [49] presented in OASIS [54]) and its target (**C**ompression). We provide a visual comparison to the widely adopted PatchGAN discriminator (HiFiC [37]) in Fig. 2.

In HiFiC y is pre-processed by a 3×3 Convolution-SpectralNorm-LeakyReLU layer with 12 filters and stride 1, upsampled and concatenated with x and subsequently fed into a PatchGAN discriminator [25]. In OASIS-C, i) we use the same latent pre-processing blocks (highlighted gray) but instead employ a pixel-wise projection-based conditioning scheme. ii) We replace spectral norm with weight norm, which increases the overall model capacity. iii) We pre-train the (unconditional) discriminator using DeepLab2 [62] to accelerate training (highlighted orange); the projection layers are initialized randomly. Note that this is motivated by the fact that we also start from a pre-trained E, G_1, P state (training stage two). Our formulation shares some characteristics with projected GANs [52], which projects samples to a pre-trained feature space, prior to classification. While in [52] the pre-trained feature space is fixed, we fine-tune the whole model to learn the conditional distribution $P_{X|Y}$. We justify all our design decisions in more detail in Sec. 5.

Output residual prediction (ORP). We start by discussing existing solutions for multi-realism image compression, see Fig. 3 for an overview.

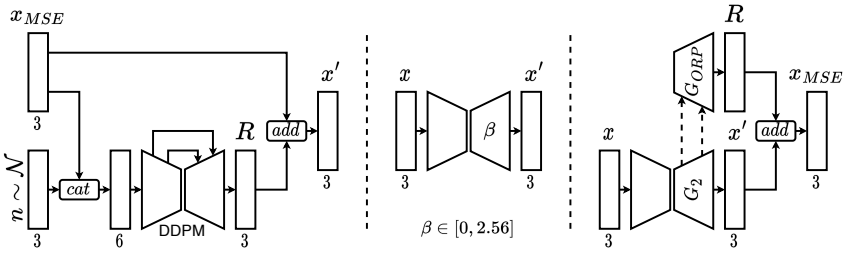

Fig. 3: DIRAC-n [17] vs Beta Conditioning (MRIC) [1] vs ORP

Dirac-n [17] uses a conditional denoising probabilistic model [23] (DDPM) for residual prediction R, conditioned on an initial reconstruction x_{MSE} (*e.g.*, JPEG or learned codec). By adjusting the number of sampling steps, $n \in [1, 100]$, DIRAC-n smoothly interpolates the D-P curve. Despite good performance, Dirac-n has two main drawbacks: parameter overhead and speed (up to 100 sampling steps), caused by the additional denoising network.

MRIC [1] proposes a variant of loss-conditional training [15] (in their paper referred to as Beta conditioning) to target different operating modes. Intuitively, the idea is to train on a distribution of losses, rather than on some specific configuration. At test time, the trained model can then be conditioned to generate outputs based on user-preferences (β). In order to work well, loss-conditional training requires a sophisticated conditioning mechanism. In practice, it also may be difficult to cover the whole spectrum of learning objectives.

ORP on the other hand is a lightweight retrofit solution that is inspired by recent theoretical findings, that simple image interpolation between an MSE-optimized decoder and a perfect perceptual decoder is sufficient to achieve any point on the D-P curve [69]. The main idea is to (implicitly) predict the residual R between a MSE-optimized and GAN-optimized decoder output:

$$x' = G_2(x) + (1 - \alpha)R, \tag{7}$$

$$\mathcal{L}_{ORP} = \mathbb{E}_{x \sim p_X}[MSE(x, x')] \tag{8}$$

with $R = G_{ORP}(F)$, where F are feature maps extracted from $G_2(x)$. During training $\alpha = 0$, which constraints x' to the traditional MSE-optimized decoder output. During inference, $\alpha \in [0, 1]$ allows traversing any point on the D-P curve using an implicitly encoded variant of image interpolation [26, 69]. Compared to existing solutions, ORP has several key advantages: i) The ORP is model agnostic and can be added to any pre-trained generative image compression method. ii) For $\alpha = 1$, we always get the regular GAN-based output (R is canceled out), which in practice is more difficult to obtain. iii) We only need to finetune G_{ORP} (we keep E, P, G_2 frozen), which considerably speeds up training.

5 Exploring GANs for Compression

In this section, we study and compare suitable discriminator architectures/ GAN formulations for the task of generative image compression. We consider PatchGAN [25], SESAME [42], U-Net [53], projected [52], and OASIS [54] discriminators. We employ training strategy-II for most parts, motivated by the observation that a perceptual decoder can be trained using solely a GAN conditioned on an encoder optimized under the traditional rate-distortion objective [65]. Intuitively, a good candidate should be able to generate high-fidelity reconstructions based on pure adversarial supervision.

Setup. For a fair comparison, we base our experiments on the official code base of HiFiC [37] and DeepLab2 [62], a TensorFlow library for deep labeling[5]. We use the same encoder, decoder, and entropy architecture as in [37], and only change the discriminator/ GAN learning objective. The latter is motivated by recent theoretical arguments that the critic is decisive in matching the distribution of the training data [52]. All our experiments start from stage two, using the same pre-trained E, G_1, and P. As a baseline, we use the official training configuration from [37], *i.e.*, we use adversarial supervision+distortion terms, however with fixed E and P to enable identical bit-rates across all experiments.

Datasets. We use the Coco2017 panoptic dataset [35] with 118,287 training images and 133 semantic classes for stage one and our main experiments. To evaluate the generalization ability across the image domain, we use the following benchmarks: DIV2K [2], CLIC 2020 [58], and Kodak [31]. DIV2K and CLIC 2020 are both high-resolution image datasets, which contain 100 and 428 images, respectively (see [37, A.9]); Kodak contains 24 images and is widely used as an image compression benchmark. For our preliminary study, we additionally consider a down-sized (factor 2) version of Cityscapes [12], which contains 19 semantic classes, 2975 training, and 500 validation images, respectively.

Training and evaluation. Again, for a fair comparison, we use the same hyper-parameters as in HiFiC, except the learning rate, which we fix to 1e$-$5 for stage two. For strategy-II, we set $\beta = 1$ to rebalance G and D. We further reduce the number of optimization steps on Cityscapes from 1M to 150k, considering the reduced size and complexity. We use PSNR and the FID-score [22] as a measure for distortion and perception, following recent work [1, 37]. For our preliminary experiments, we compute the patched FID-score (FID/256 [37, A.7]) based on the clean-FID implementation [46]; for our main experiments, we have switched to torch-fidelity [43] to ease comparison to recent methods, where a recalculation based on clean-FID is prevented due to restrictive data access. As common in the literature, we pad all images and crop the resulting reconstructions.

5.1 Comparing GAN Approaches

Fairly comparing discriminator architectures/ GAN formulations for generative image compression is difficult due to the large variety in model design, condi-

[5] We partially translate DeepLab2 to TensorFlow 1.15 to directly integrate the code base into HiFiC.

Table 1: Comparing GAN approaches on Cityscapes (0.092bpp)

Method	Disc. (D)	Cond.	GAN objective	Distortion		Perception	
				PSNR ↑	rel-PSNR	FID ↓	rel-FID
baseline	PatchGAN [25]	cat	non-saturating	**32.71**	-	**10.62**	-
conf-a	PatchGAN	cat	non-saturating	24.32	-25.6%	112.29	+957.3%
conf-b	SESAME [42]	cat	hinge	29.43	-10.0%	75.65	+612.3%
conf-c	U-Net [53]	cat	non-saturating	29.46	-9.9%	87.02	+719.4%
conf-d	Projected [52]	cat	hinge	29.64	-9.5%	50.66	+377.0%
conf-e	OASIS [54]	cat	CE (N+1)	30.03	-8.2%	16.50	+55.4%
conf-f	DeepLabV3+ [11]	cat	CE (N+1)	21.53	-34.2%	20.61	+94.1%
conf-c	U-Net	proj.	non-saturating	29.38	-10.2%	30.80	+190.0%
conf-e	OASIS	proj.	CE (N+1)	29.79	-8.9%	13.54	+27.5%
conf-e	OASIS	impl.	CE (N+1)	29.90	-8.6%	15.30	+44.1%

tioning scheme and regularization terms. It is also important to note that these methods were primarily co-designed with their respective generator structures, while we consider them in isolation. We make no claim to the superiority of one method over another, but rather are interested in their general suitability for generative image compression.

We consider the low bit range (HiFiC-lo), where the influence of generative models is arguably greatest. For each method, we use the best configuration following the original work, including regularization terms (see supplementary material). We start our experiments by applying the same HiFiC-based conditioning scheme in all model variants, if possible. We make the following observations: unsurprisingly, no pure GAN-based configuration exceeds the performance of the baseline method (32.71dB PSNR, 10.62 FID-score). The performance gaps range from -34.2% to -8.2% and $+957.3\%$ to $+94.1\%$ decrease in PSNR and FID-score, respectively. The PatchGAN discriminator performs worst, which is to be expected since it was designed primarily to penalize high-frequency structure in addition to the commonly used L1/L2 loss functions [25]. Indeed, when paired with an additional distortion loss, this formulation works remarkably well as demonstrated by the baseline configuration as well as by previous work [1, 3, 21, 26, 37].

We find that both the SESAME, U-Net, and projected discriminators produce similar strong PSNR values ($29.43 - 29.64$dB), with varying degrees of artifacts (see Fig. 4). The SESAME discriminator improves upon the PatchGAN variant due to its inherent multi-scale nature as well as access to additional semantic side information. For projected GANs, we find that the perceptual quality largely depends on the image resolution of the efficientnet-lite feature extractors. We suppose that the (well hidden) gridding artifacts are due to a resolution mis-

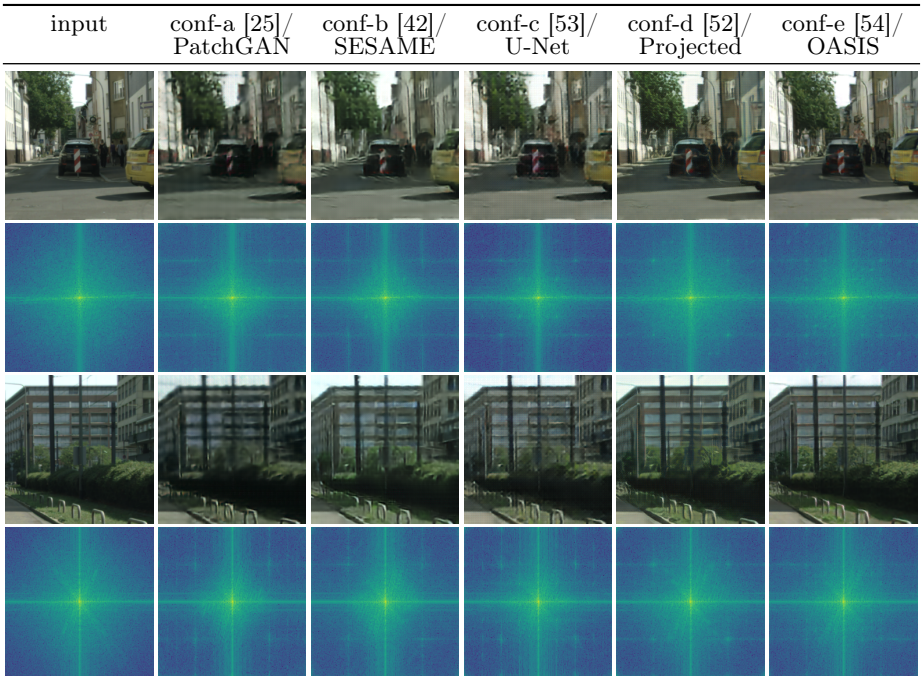

Fig. 4: Comparing various purely adversarially optimized generative image compression methods at low bit-rate (HiFiC-lo). Rows one and three show examples of reconstructed cropped images (256×256), while rows two and four show the corresponding spectra of the images. **Best viewed electronically.**

match; *i.e.*, the efficientnet-lite variants are based on low resolutions images, *e.g.*, 224×224px for efficientnet-lite0, whereas HiFiC mostly targets high-resolution images up to 2000×2000px.

The best purely adversarial method is achieved by OASIS, which considerably exceeds all its competitors in terms of perception (FID-score of 16.50). We attribute its better performance to the spatially and semantically-aware pixel-level supervision, which implicitly provides a strong conditioning mechanism (see Tab. 1, conf-e/ impl. conditioning). Note that a stronger semantic segmentation model (*e.g.*, DeeplabV3+ [11]) does not per se lead to better performance (conf-f). This finding is consistent with the observation that a stronger feature extractor does not necessarily lead to lower FID scores [52].

5.2 Investigating the Conditioning Scheme

We use the same setup as before, but now replace the vanilla concatenation-based conditioning scheme with projection [40] for some selected methods. For OASIS and U-Net, we employ a pixel-wise projection-based conditioning scheme; for U-Net we use an additional projection for the global output.

Table 2: Improving OASIS [54] step-by-step

Method	PSNR ↑	FID ↓
OASIS [54]	29.90	15.30
+ pre-trained	30.01	9.75
+ weight norm [50]	30.20	7.96
+ projection [40] (ours w/o d)	29.97	7.74
ours w/ d	**32.24**	**6.36**

It can be observed that all projection-based configurations improve their base configurations while largely reducing image artifacts (see supplementary material). For optimization strategies with distortion, these considerations probably play a minor role, since d already provides for a strong implicit conditioning mechanism. However, for approaches based on pure adversarial optimization, such as in the case of the optimal training framework [65, 69], our findings shed some light on the lack of generalizability beyond the MNIST dataset.

5.3 Improving OASIS

A major concern we had prior to the adoption was the tremendous amount of hardware resources required for training OASIS. Specifically, Schönfeld *et al.* trained their model on COCO-stuff for 4 weeks, in a multi-GPU environment (4× Tesla V100 GPUs). Instead, we target a single-GPU setup.

We attribute the slow training process to the spectral norm in its default configuration, which we have found to severely hinder learning progress. We note that similar observations have been reported recently, *e.g.*, Lee *et al.* [32] proposed to multiply the normalized weight matrix with the spectral norm at initialization, which increases the often untuned Lipschitz constant and hence the overall model capacity. For our specific use case, we have found that weight normalization [50] combined with a pre-trained discriminator produces a good training speed/ stability/ compression performance trade-off (Tab. 2). In the supplementary material we further show that simply replacing the PatchGAN discrimiantor with OASIS alone is not sufficient to improve over the state-of-the-art. This is especially true for highly complex learning tasks (Coco2017), where OASIS exhibits sever training instabilities.

6 Comparison to the State-of-the-Art

For our main experiments, we use SwinT-ChARM [72] as backbone architecture with an additional prediction head (we clone the last Swin-transformer block [36] of the pre-trained G_2, which corresponds to G_{ORP} in Sec. 4). Exact model configurations as well as extended experiments on HiFiC and additional datasets (Kodak and DIV2K) are summarized in the supplementary material.

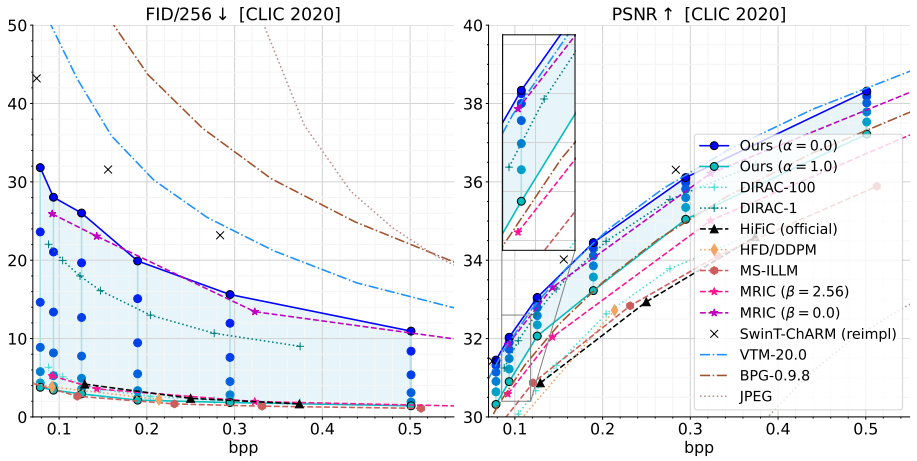

Fig. 5: Comparison to the state-of-the-art on CLIC 2020

Baselines. We consider both diffusion (SwinT-ChARM/ DIRAC-100, short DIRAC-100 [17], HFD/DDPM [24]) and GAN-based (HiFiC [37], MS-ILLM [41], MRIC ($\beta = 2.56$)) methods for perception, *i.e.*, Ours ($\alpha = 1.0$). Similarly, we consider VTM-20.0, BPG-0.9.8, JPEG, SwinT-ChARM/ DIRAC-1, short DIRAC-1 [17], and MRIC ($\beta = 0.0$) [1] for distortion, *i.e.*, Ours ($\alpha = 0.0$). We further add SwinT-ChARM (reimpl), our TensorFlow-reimplementation of SwinT-ChARM, which can be considered an upper bound for distortion.

We provide objective and subjective comparisons on CLIC 2020 in Figs. 1, 5 and 6. We observe that Ours ($\alpha = 1.0$) is most effective in the low to medium bit range, being competitive or outperforming strong baselines (HiFiC, MS-ILLM, MRIC ($\beta = 2.56$), DIRAC-100, HFD/DDPM) in terms of FID, while having considerably better PSNR-scores in all cases. This becomes even more pronounced on DIV2K, where Ours ($\alpha = 1.0$) dominates MS-ILLM in terms of perception. Noteworthy, EGIC outperforms HiFiC-lo, the long standing previous state-of-the-art, even when using 30% fewer bits. For Ours ($\alpha = 0.0$), we find that ORP is surprisingly effective. In terms of distortion, our method outperforms MRIC and DIRAC, while (almost) matching VTM-20.0, the state-of-the-art for non-learned codecs. This is despite using only a fraction of additional model parameters (*e.g.*, 0.15× compared to Beta conditioning in MRIC [1]) and only requiring a single inference-cycle (as opposed to DIRAC-n [17]), revealing that more sophisticated methods may in fact be unnecessary. In the supplementary material we further discuss other formulations for ORP and provide a detailed comparison to more involved image and weight interpolation techniques.

Finally, it is worth mentioning that EGIC is considerably more storage-efficient compared to all other methods. EGIC only requires a fraction of the number of model parameters compared to HFD/DDPM (0.03×), HiFiC/ MS-ILLM (0.18×), DIRAC (0.24×), and MRIC (0.55×); in contrast to DIRAC and HFD/DDPM, EGIC also only requires a single inference cycle.

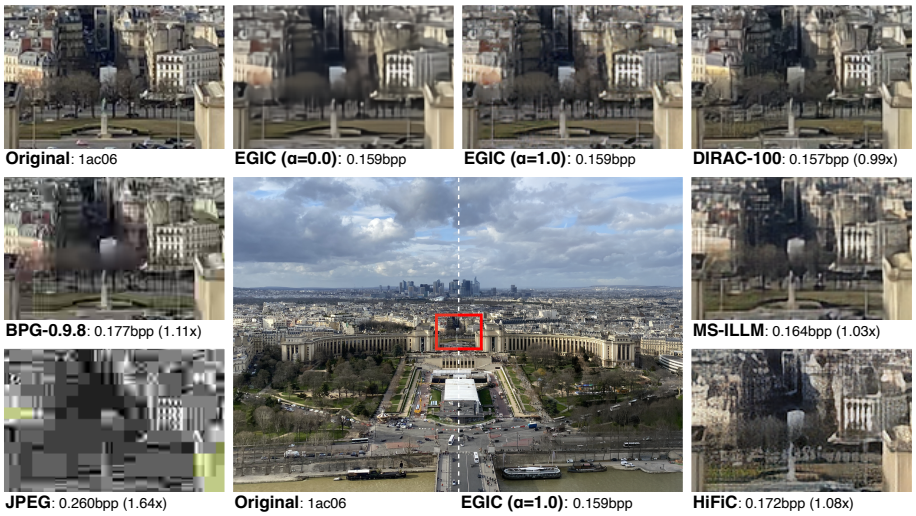

Fig. 6: Visual comparison of EGIC ($\alpha \in \{0.0, 1.0\}$) with state-of-the-art distortion-oriented (l.) and perception-oriented (r.) codecs. Please visit the supplementary material for more impressions. **Best viewed electronically.**

7 Conclusion

We have developed EGIC, a novel generative image compression method that allows traversing the D-P curve efficiently from a single model on the receiver side. We find that EGIC is highly competitive, outperforming a wide-variety of diffusion and GAN-based methods (*e.g.*, HiFiC, MS-ILLM, DIRAC-100), while operating almost on par with VTM-20.0 on the distortion-oriented end. EGIC enjoys a simple and lightweight design with excellent interpolation characteristics, which makes it a promising candidate for practical applications targeting the low bit range. Our code will be made publicly available[6] upon publication to facilitate further research.

Limitations. Gradient feedback from OASIS-C is currently applied to the whole image, which in some cases might lead to sub-optimal preservation of small faces and text. This can be addressed via content-weighted learning mechanisms [33, 34]. Furthermore, EGIC at this stage requires large labeled training data, which can be alleviated by switching to semi-supervised approaches or by incorporating a powerful prior for semantic segmentation tasks, *e.g.*, SAM [9,30]. Finally, despite promising quantitative results, we find that HFD/DDPM and DIRAC-100 produce in some cases more pleasing (not necessarily more accurate) reconstructions. Whether this can be attributed to the superiority of diffusion models in general, or simply to the substantial difference in model sizes (1B in the case of HFD/DDPM), remains an open subject of debate.

[6] `https://github.com/Nikolai10/EGIC`

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research under the funding program Forschung an Fachhochschulen - FKZ 13FH019KI2. The authors would further like to thank Auke Wiggers, Matthew Muckley, Eirikur Agustsson and Lucas Theis for providing evaluation data.

References

1. Agustsson, E., Minnen, D., Toderici, G., Mentzer, F.: Multi-realism image compression with a conditional generator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
2. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2017)
3. Agustsson, E., Tschannen, M., Mentzer, F., Timofte, R., Gool, L.V.: Generative adversarial networks for extreme learned image compression. In: Proceedings of the IEEE International Conference on Computer Vision (2019)
4. Arad Hudson, D., Zitnick, L.: Compositional transformers for scene generation. In: Advances in Neural Information Processing Systems. vol. 34 (2021)
5. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70 (2017)
6. Ballé, J., Minnen, D., Singh, S., Hwang, S.J., Johnston, N.: Variational image compression with a scale hyperprior. In: International Conference on Learning Representations (2018)
7. Blau, Y., Michaeli, T.: Rethinking Lossy Compression: The Rate-Distortion-Perception Tradeoff. In: Proceedings of the 36th International Conference on Machine Learning (2019)
8. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: International Conference on Learning Representations (2019)
9. Chen, J., Yang, Z., Zhang, L.: Semantic segment anything. `https://github.com/fudan-zvg/Semantic-Segment-Anything` (2023)
10. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
11. Cheng, B., Collins, M.D., Zhu, Y., Liu, T., Huang, T.S., Adam, H., Chen, L.C.: Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In: CVPR (2020)
12. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
13. Cover, T.M., Thomas, J.A.: Elements of information theory. John Wiley & Sons (2012)
14. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: Advances in Neural Information Processing Systems. vol. 34 (2021)
15. Dosovitskiy, A., Djolonga, J.: You only train once: Loss-conditional training of deep networks. In: International Conference on Learning Representations (2020)

16. Gal, R., Hochberg, D.C., Bermano, A., Cohen-Or, D.: Swagan: A style-based wavelet-driven generative model. ACM Trans. Graph. **40**(4) (2021)
17. Ghouse, N.F., Petersen, J., Wiggers, A., Xu, T., Sautière, G.: A Residual Diffusion Model for High Perceptual Quality Codec Augmentation. arXiv: 2301.05489 (2023)
18. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. vol. 27 (2014)
19. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
20. He, D., Yang, Z., Peng, W., Ma, R., Qin, H., Wang, Y.: Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
21. He, D., Yang, Z., Yu, H., Xu, T., Luo, J., Chen, Y., Gao, C., Shi, X., Qin, H., Wang, Y.: Po-elic: Perception-oriented efficient learned image coding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2022)
22. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems (2017)
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33 (2020)
24. Hoogeboom, E., Agustsson, E., Mentzer, F., Versari, L., Toderici, G., Theis, L.: High-Fidelity Image Compression with Score-based Generative Models. arXiv: 2305.18231 (2023)
25. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. CVPR (2017)
26. Iwai, S., Miyazaki, T., Sugaya, Y., Omachi, S.: Fidelity-controllable extreme image compression with generative adversarial networks. In: 2020 25th International Conference on Pattern Recognition (ICPR). Los Alamitos, CA, USA (2021)
27. Iwai, S., Miyazaki, T., Omachi, S.: Controlling rate, distortion, and realism: Towards a single comprehensive neural image compression model. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2900–2909 (January 2024)
28. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Focal frequency loss for image reconstruction and synthesis. In: ICCV (2021)
29. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: Advances in Neural Information Processing Systems. vol. 33 (2020)
30. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
31. Kodak, E.: Kodak lossless true color image suite (PhotoCD PCD0992)
32. Lee, K., Chang, H., Jiang, L., Zhang, H., Tu, Z., Liu, C.: ViTGAN: Training GANs with vision transformers. In: International Conference on Learning Representations (2022)
33. Li, M., Gao, S., Feng, Y., Shi, Y., Wang, J.: Content-oriented learned image compression. In: Computer Vision – ECCV 2022 (2022)

34. Li, M., Zuo, W., Gu, S., Zhao, D., Zhang, D.: Learning convolutional networks for content-weighted image compression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
35. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
36. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
37. Mentzer, F., Toderici, G.D., Tschannen, M., Agustsson, E.: High-fidelity generative image compression. Advances in Neural Information Processing Systems (2020)
38. Minnen, D., Singh, S.: Channel-wise autoregressive entropy models for learned image compression. In: 2020 IEEE International Conference on Image Processing (ICIP) (2020)
39. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv: 1411.1784 (2014)
40. Miyato, T., Koyama, M.: cGANs with projection discriminator. In: International Conference on Learning Representations (2018)
41. Muckley, M.J., El-Nouby, A., Ullrich, K., Jegou, H., Verbeek, J.: Improving statistical fidelity for neural image compression with implicit local likelihood models. In: Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202 (2023)
42. Ntavelis, E., Romero, A., Kastanis, I., Van Gool, L., Timofte, R.: SESAME: Semantic Editing of Scenes by Adding, Manipulating or Erasing Objects. In: Computer Vision – ECCV 2020. Cham (2020)
43. Obukhov, A., Seitzer, M., Wu, P.W., Zhydenko, S., Kyl, J., Lin, E.Y.J.: High-fidelity performance metrics for generative models in pytorch (2020)
44. van den Oord, A., Vinyals, O., kavukcuoglu, k.: Neural discrete representation learning. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
45. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
46. Parmar, G., Zhang, R., Zhu, J.Y.: On aliased resizing and surprising subtleties in gan evaluation. In: CVPR (2022)
47. Razavi, A., van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. In: Advances in Neural Information Processing Systems. vol. 32 (2019)
48. Rippel, O., Bourdev, L.: Real-time adaptive image compression. In: Proceedings of the 34th International Conference on Machine Learning (2017)
49. Ronneberger, O., P.Fischer, Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). LNCS, vol. 9351 (2015)
50. Salimans, T., Kingma, D.P.: Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In: Advances in Neural Information Processing Systems. vol. 29 (2016)
51. Santurkar, S., Budden, D., Shavit, N.: Generative compression. In: 2018 Picture Coding Symposium (PCS) (2018)
52. Sauer, A., Chitta, K., Müller, J., Geiger, A.: Projected gans converge faster. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)
53. Schonfeld, E., Schiele, B., Khoreva, A.: A u-net based discriminator for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)

54. Schönfeld, E., Sushko, V., Zhang, D., Gall, J., Schiele, B., Khoreva, A.: You only need adversarial supervision for semantic image synthesis. In: International Conference on Learning Representations (2021)
55. Schwarz, K., Liao, Y., Geiger, A.: On the frequency bias of generative models. In: Advances in Neural Information Processing Systems. vol. 34 (2021)
56. Souly, N., Spampinato, C., Shah, M.: Semi supervised semantic segmentation using generative adversarial network. In: 2017 IEEE International Conference on Computer Vision (ICCV) (2017)
57. Theis, L., Salimans, T., Hoffman, M.D., Mentzer, F.: Lossy compression with gaussian diffusion (2023)
58. Toderici, G., Theis, L., Johnston, N., Agustsson, E., Mentzer, F., Ballé, J., Shi, W., Timofte, R.: Clic 2020: Challenge on learned image compression (2020)
59. Tschannen, M., Agustsson, E., Lucic, M.: Deep generative models for distribution-preserving lossy compression. In: Advances in Neural Information Processing Systems. vol. 31 (2018)
60. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
61. Wang, X., Yu, K., Dong, C., Tang, X., Loy, C.C.: Deep network interpolation for continuous imagery effect transition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
62. Weber, M., Wang, H., Qiao, S., Xie, J., Collins, M.D., Zhu, Y., Yuan, L., Kim, D., Yu, Q., Cremers, D., Leal-Taixe, L., Yuille, A.L., Schroff, F., Adam, H., Chen, L.C.: DeepLab2: A TensorFlow Library for Deep Labeling. arXiv: 2106.09748 (2021)
63. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems (2021)
64. Xu, T., Zhang, Q., Li, Y., He, D., Wang, Z., Wang, Y., Qin, H., Wang, Y., Liu, J., Zhang, Y.: Conditional perceptual quality preserving image compression. arXiv: 2308.08154 (2023)
65. Yan, Z., Wen, F., Ying, R., Ma, C., Liu, P.: On perceptual lossy compression: The cost of perceptual reconstruction and an optimal training framework. In: Proceedings of the 38th International Conference on Machine Learning (2021)
66. Yang, R., Mandt, S.: Lossy image compression with conditional diffusion models. arXiv: 2209.06950 (2023)
67. Yang, T.J., Collins, M.D., Zhu, Y., Hwang, J.J., Liu, T., Zhang, X., Sze, V., Papandreou, G., Chen, L.C.: Deeperlab: Single-shot image parser. arXiv: 1902.05093 (2019)
68. Yang, Y., Mandt, S., Theis, L.: An introduction to neural data compression. arXiv: 2202.06533 (2022)
69. Zeyu Yan, Fei Wen, P.L.: Optimally controllable perceptual lossy compression. In: Proceedings of the International Conference on Machine Learning (ICML) (2022)
70. Zhang, G., Qian, J., Chen, J., Khisti, A.J.: Universal rate-distortion-perception representations for lossy compression. In: Advances in Neural Information Processing Systems (2021)
71. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
72. Zhu, Y., Yang, Y., Cohen, T.: Transformer-based transform coding. In: International Conference on Learning Representations (2022)