# Supplementary: Towards Device Efficient Conditional Image Generation

Nisarg A. Shah
snisarg812@gmail.com
Gaurav Bharaj
first.last@gmail.com

AI Foundation
California, USA

## 1 Detailed Setup and Implementation details

We evaluate and ablated our method on CelebA-HQ dataset [5] that contains 30,000 high-quality face images (resized to $256 \times 256$) and corresponding pixel-level segmentation mask annotations. We evaluate images generated by our optimized $G^*$ using perceptual quality score – FID [7], memory consumption, run-times, and quantitative comparisons. We conduct experiments to translate semantic segmentation mask to face images using pix2pix [3] to compare different methods. The paired dataset is divided into about 23,500 training images, 4,000 validation images, and about 2,500 test images. To verify the efficacy of our algorithm across different autoencoders, we follow the settings in pix2pix and use U-net [6] and ResNet as generators. Like [1, 3], we use PatchGANs, that uses $70 \times 70$ image patches instead of whole images. During optimization of the networks, the objective value is divided by two while optimizing the discriminator. The networks are trained for 200 epochs using Adam [4], and learning rate of $1e^{-4}$.

We use U-net [3] termed as Unet-64, where number of channels is 64 and that gets doubled after every strided convolution with an upper limit of 512. We also evaluate our approach on an *overly-parametrized* Unet-192 to observe its advantages to reduce over-fitting. We also trained Unet-32 and Unet-16 to compare the pruned variants of Unet-64 in an equi-parametric setting. Since the discriminator does not affect inference time, the student and teacher discriminator structure was kept the same. We analyze the performance of different autoencoders – Unet and ResNet, and compare their respective vanilla versions and optimized models using our proposed method. We further show application of $\mathsf{ALAP} - \mathsf{AE}$ on CycleGAN for horse-to-zebra dataset, and on pix2pix to cartoonize the faces to verify the generalizability of our algorithm across different tasks and comparison with state-of-art methods available.

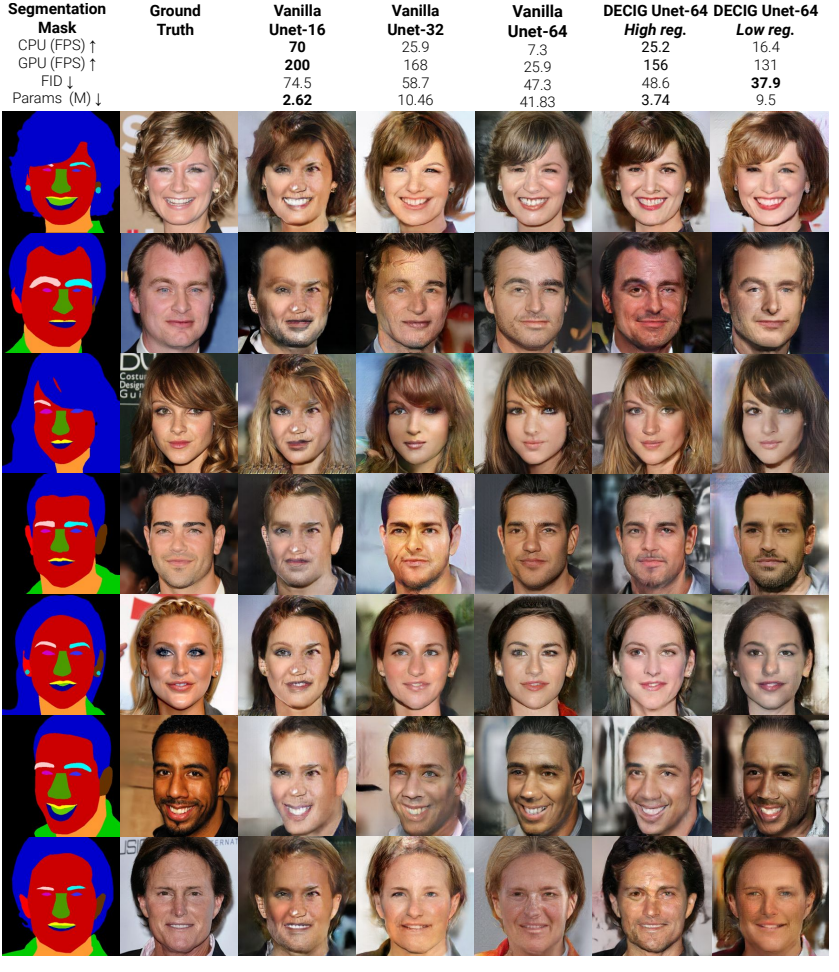


Figure 1: (Left to Right) We take several Vanilla Unet variants as baseline for conditional-GAN based Image generation (Unet-64, Unet-32, Unet-16), and create DECIGversions of the Unet-64 network for high and low regularization setting. Note: While better image generation methods exist, our emphasis is to maintain image quality vs. the baselines autoencoders.

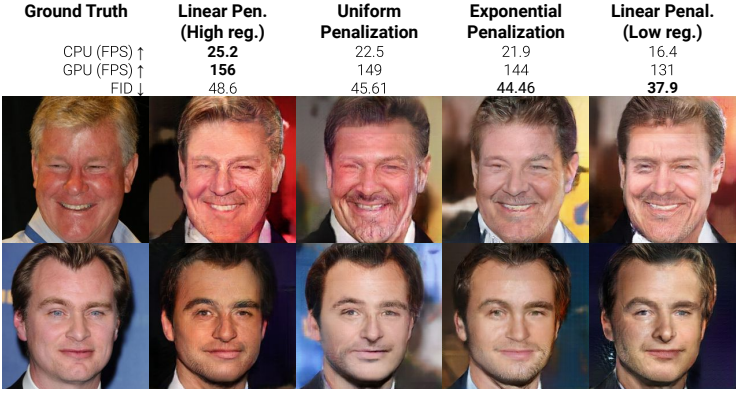


Figure 2: (Left to Right) We create the versions of Unet-64 variant with different Channel weight regularization. ( *Linear*(high-reg.), *Uniform*, *Exponential*, *Linear*(low-reg.) feature channel regularization)

## 2 Detailed Qualitative and Quantitative results of DECIG-UNet over CelebA-HQ dataset

Fig. 1 shows additional results of several variants on U-net[6] architectures for conditional image generation. While satisfactory results are achieved for vanilla generator (Unet-64), it requires significant parameters as well as compute resources 1. Although, *miniature* Unet variants (Unet-16 and Unet-32) have fewer MACs (FLOPs), memory consumption, and parameters, their generated images look austere and blurry with repeated patches; thus making them look fake. While images generated by our proposed condensed generators look sharper and more realistic, at a low inference times. Here, it is important to note, that primary objective of *high-reg* version of DECIG is to develop more compressed model with equivalent perceptual scores, compared to it's vanilla variant, whereas, in *low-reg* versions, higher perceptual quality is preferred over compression metrics. *High* and *low* indicates the amount of penalization in the overall loss function.

## 3 Channel Weight and Layer Device Regularization

Our channel weight regularization method supports several multiplicative functions like Uniform, Linear, and Exponential. Uniform and Exponential factors are used in low reg form. The qualitative and quantitative predictions for different channel weight multiplicative factors discussed in Fig. 2.

We also tested the results for device agnostic layer level regularization discussed in Sec. 3.2 in Fig. 3. Here, specific improvements over inference time of model could be observed for the model optimised for that corresponding device i.e. CPU and GPU.

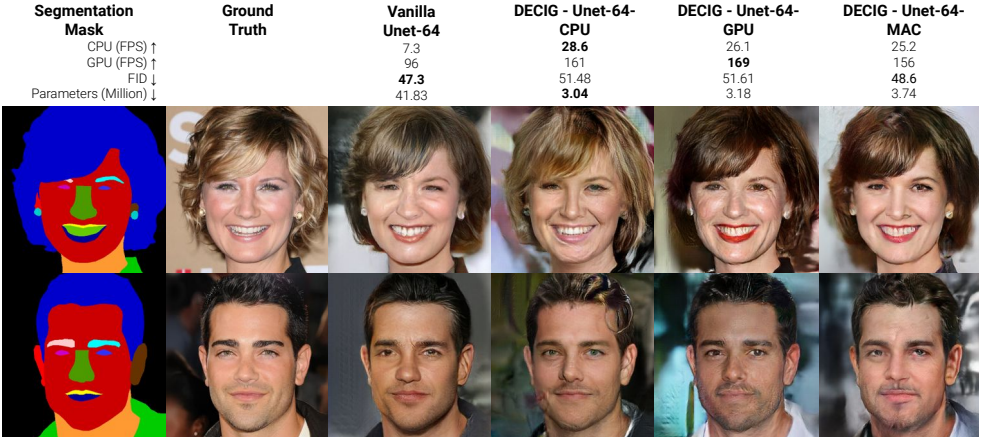


Figure 3: (Left to Right) Comparison of (high-reg) Unet-64 variants condensed for particular type of Device e.g. CPU, GPU and general (MAC) (Segmentation Map, Ground truth, Unet-64, DECIG-Unet-64-[CPU, GPU, MAC]

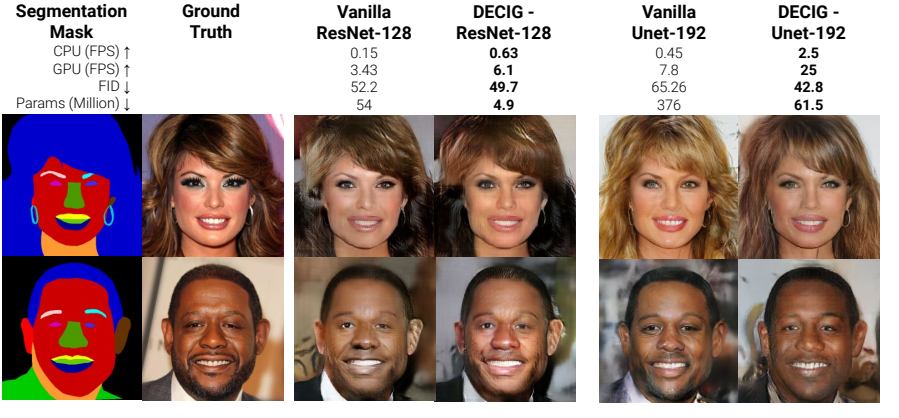


Figure 4: (Left to Right) We take baseline conditional-GAN based AE's (Unet-192, ResNet), and create DECIG versions of these AEs.

# 4 Overparametrization and Regularization effect

On quantitative part of Fig.4, we can observe that, with Unet-192 and DECIG-Unet-192, typically there's a 6× reduction in the number of parameters, and the weight-induced pruned network achieves a lower FID score compared to the original model. Unet-192 has an FID score of 65.3, which is 30% poorer compared to its smaller variant Unet-64's FID of 47.3. These results are produced due to over-fitting of the Unet-192 model on the training dataset. Interestingly, our penalization algorithm solves the over-fitting problem to an extent by achieving FID improvements of 30% with 5× and 3.2× improvements on run-time over CPU and GPU, respectively. We hypothesize this is due to the regularization effect of the penalization algorithm on channels that condenses the features in each layer, and make redundant channel weights zero.

# References

[1] Hanting Chen, Yunhe Wang, Han Shu, Changyuan Wen, Chunjing Xu, Boxin Shi, Chao Xu, and Chang Xu. Distilling portable generative adversarial networks for image translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[5] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.