GDP: Stabilized Neural Network Pruning via Gates with Differentiable Polarization — Supplementary Materials

Yi Guo[1], Huan Yuan,[1], Jianchao Tan[1], Zhangyang Wang[2], Sen Yang[1], Ji Liu[1]
[1]Kuaishou Technology, [2]University of Texas at Austin
{guoyi03,yuanhuan,jianchaotan,senyang,jiliu}@kuaishou.com, {atlaswang}@utexas.edu

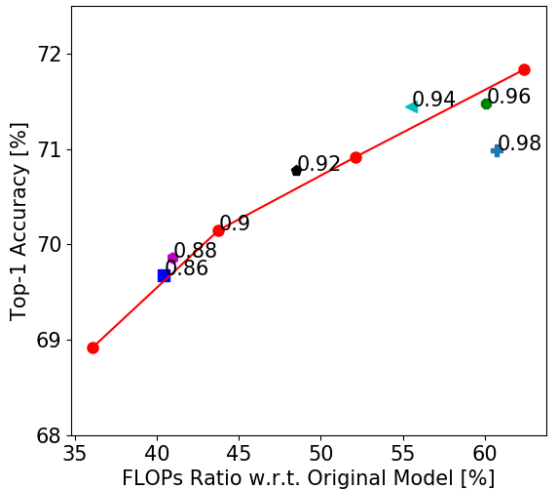

Figure 1. This figure shows different decaying rates of ϵ and their corresponding compression ratios and Top-1 accuracy. The experiment is conducted on ImageNet with MobileNet-V2. All the discrete dots have all the same hyper-parameters except the decaying rate, and the attached numerical values represent the decaying rate for each epoch. The red dashed polyline is from the main body of the paper only for reference. All the initial values of ϵ is 0.1. We can clearly see that GDP is somewhat robust to the decaying rate of ϵ ranging from 0.86 to 0.96, which results in the final ϵ ranging from $0.1 * 0.86^{140} \approx 6e^{-11}$ to $0.1 * 0.96^{140} \approx 3e^{-4}$.

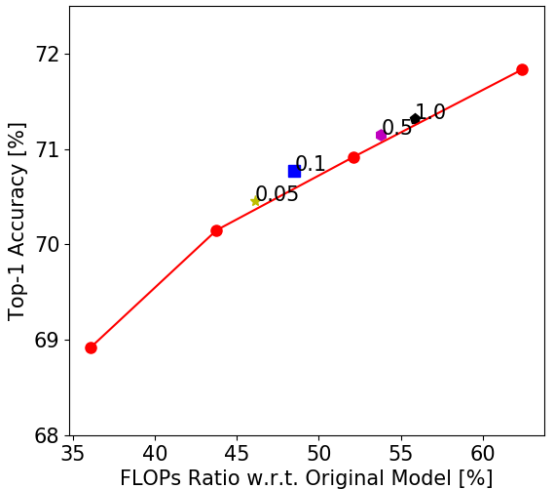

Figure 2. This figure shows different initial values of ϵ and their corresponding compression ratios and Top-1 accuracy. The experiment is conducted on ImageNet with MobileNet-V2. All the discrete dots have all the same hyper-parameters except the initial value of ϵ, and the attached numerical values represent the initial value. The red dashed polyline is from the main body of the paper only for reference. All the decaying rate of ϵ is 0.92 for each epoch. We can clearly see that GDP is somewhat robust to the initial value of ϵ.

1. Ablation study for ϵ

The smooth L0 function is defined as:

$$g_\epsilon(x) = \frac{x^2}{x^2 + \epsilon} \tag{1}$$

The polarization depends on the value of ϵ. A larger ϵ results in a smoother gate with less polarization; a smaller ϵ makes gate function closer to L_0 norm but with less numerical stability. Fig. 1 shows different decaying rates of ϵ with their corresponding compression ratios and Top-1 accuracy. We can see that GDP is robust to the final value of ϵ ranging from $0.1 * 0.86^{140} \approx 6e^{-11}$ to $0.1 * 0.96^{140} \approx 3e^{-4}$, which is a really big range.

Besides the decaying rate, we also study the effect of initial value of ϵ, as shown in Fig. 2. We can also see the robustness of GDP with respect to different initial values.

2. Visual results for style transfer

It can be seen that our pruned model can achieve similar results without hurting perceptual quality, as shown in Fig. 3 and Fig. 4

3. Visual results for semantic segmentation

We compare baseline and ground-truth with our GDP pruned models. We can find our pruned model can even achieve better boundary quality compared with baseline, as shown in Fig. 5

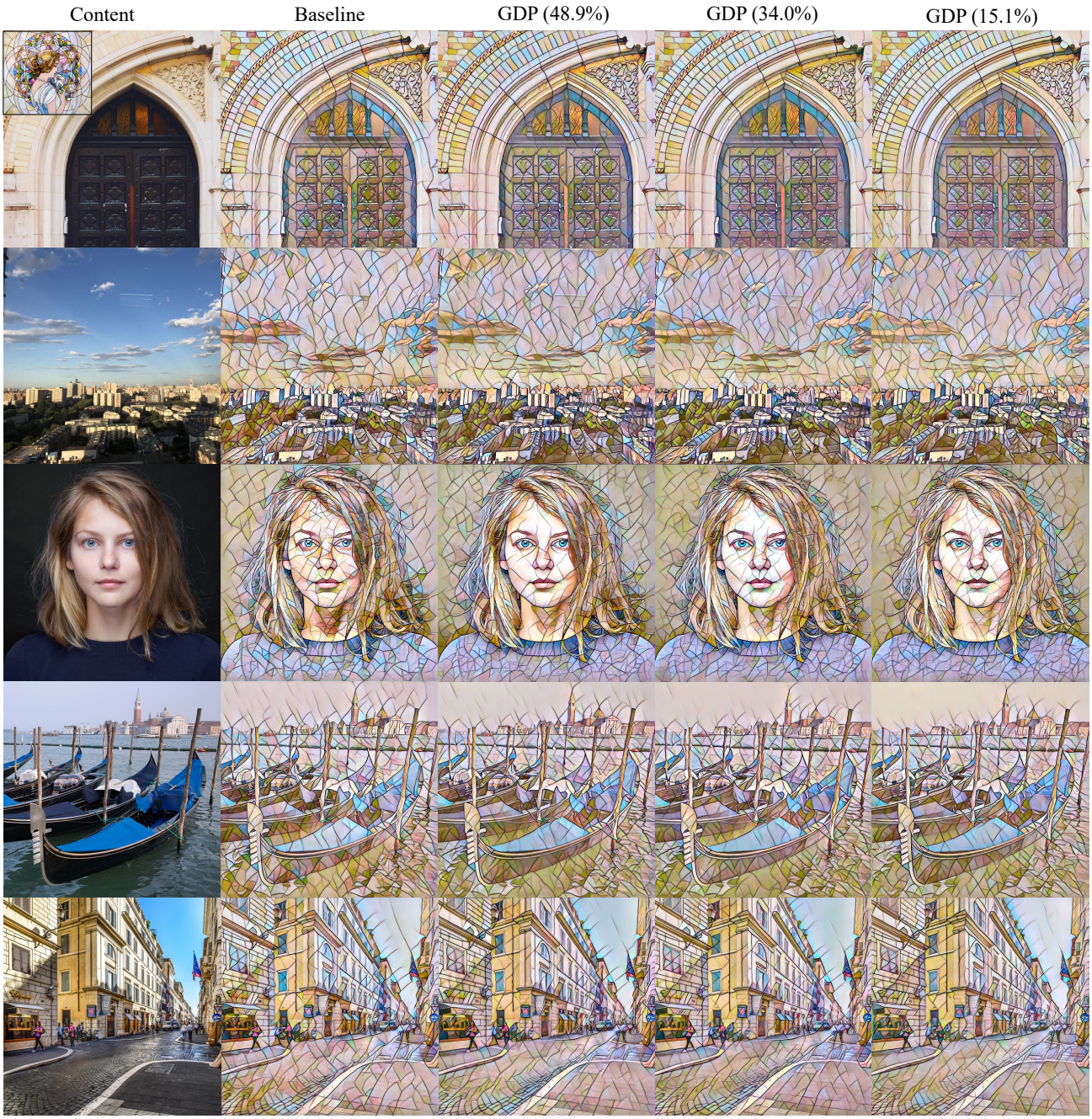

Figure 3. Visual comparisons between baseline and our GDP pruned models. GDP (48.9%) means that the remaining flops ratio of the sub-net pruned by GDP is 48.9%

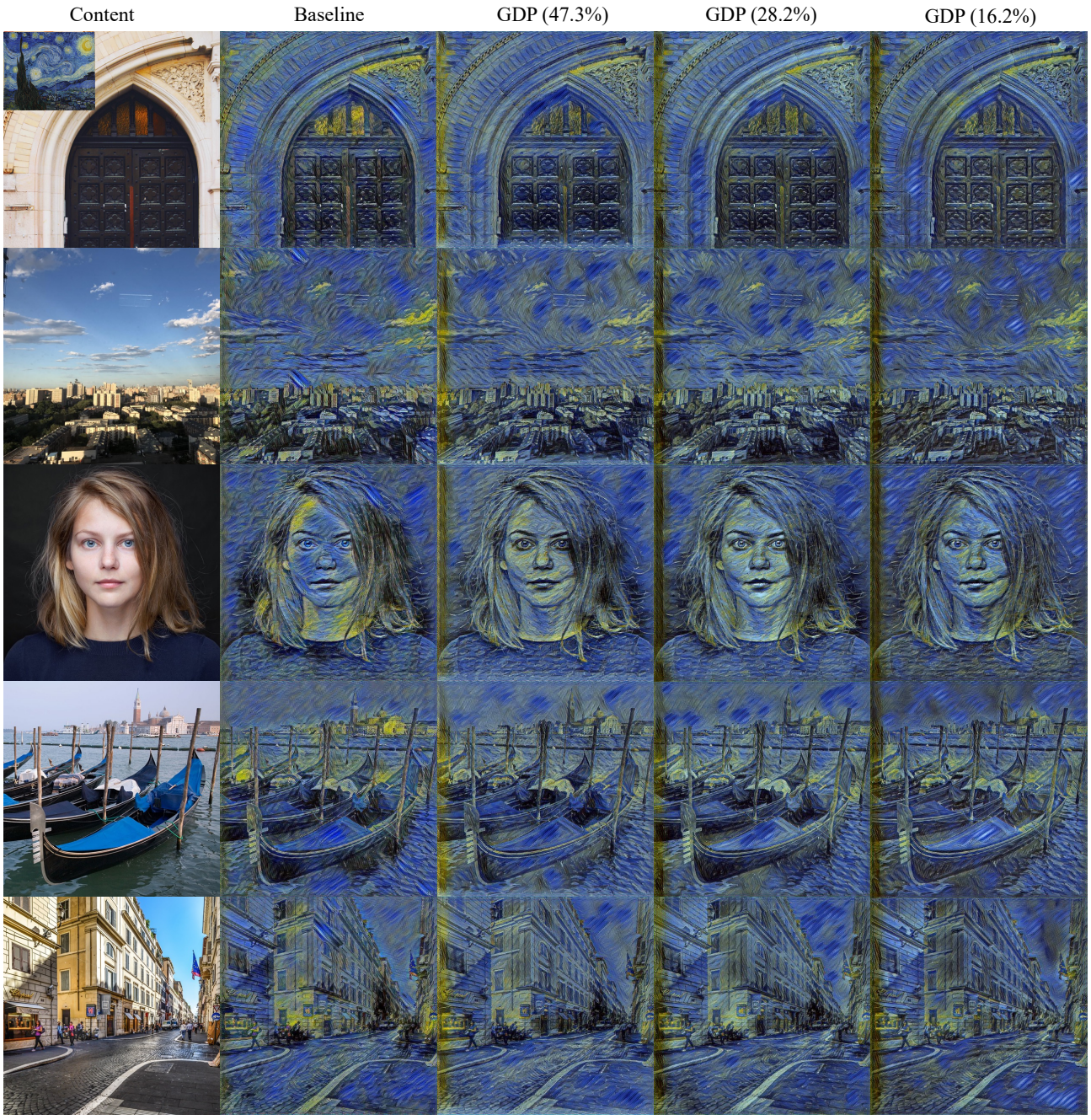

Figure 4. Visual comparisons between baseline and our GDP pruned models on style transfer. GDP (47.3%) means that the remaining flops ratio of the sub-net pruned by GDP is 47.3%

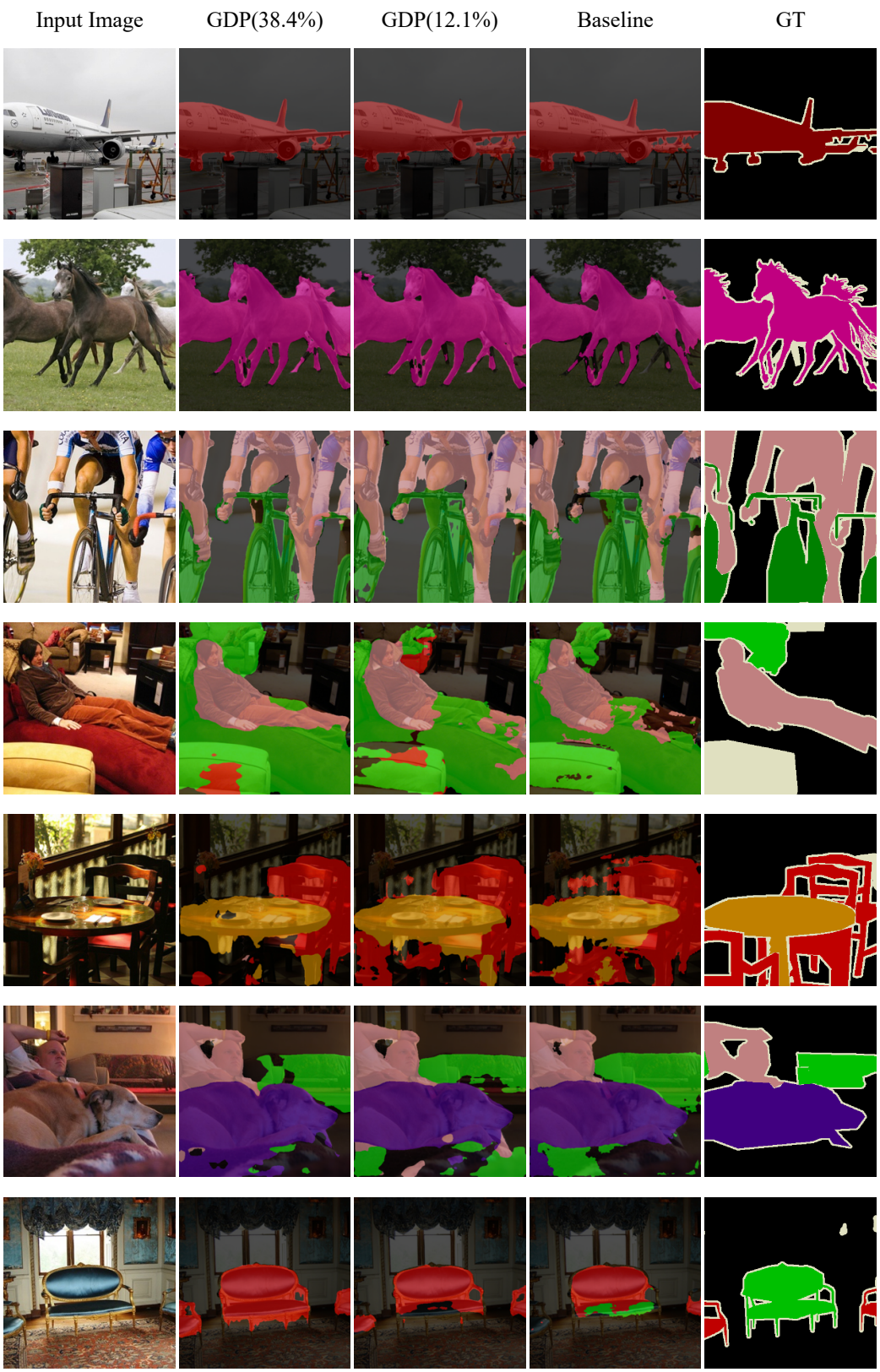

Figure 5. Visual comparisons between baseline and our GDP pruned models for Pascal VOC with DeepLabv3+. We can find our pruned model can even achieve better boundary quality.