Supplementary materials: Generic Perceptual Loss for Modeling Structured Output Dependencies

Anonymous CVPR 2021 submission

Paper ID 2643

In the supplementary materials, we first give the details of the pilot experiments described in the article. The we show more exploration results about the generic perceptual loss. We first use a toy example to show that random network has the ability to capture the structure information from a global view compared with the pixel-wise L_2 loss. Then we show the impact of the down-sampling operators.

S1. Details of the Pilot Experiment

Most of previous works attribute the success of perceptual loss to the CNN filters pretrained with a large amount of samples. They assume that training for image classification may allow the network to capture high-level features which are coincident to human perception. Although He et al. [9] showed that perceptual loss with the random weight network can work well in solving optimizing problem in the style transfer, it is still interesting to find out if the random weights can work well as a perceptual regularization in training CNNs.

In this pilot experiment, we simply follow the settings in [14], and conduct experiments on a typical image synthesis task, i.e., image super-resolution (SR). We use the popular SRResNet [16] as an SR network. Assume that the input low quality image is I_l, a fully convolutional network can transfer I_l to the estimated high-resolution image $\hat{I}_h$, and we try to minimize the difference between $\hat{I}_h$ and the real image I_h with per-pixel and perceptual losses following Eq. 1.

$$\ell_{feat}^{\phi,j}(\hat{y}, y) = \frac{1}{C_j H_j W_j} \|\phi_j(\hat{y}) - \phi_j(y)\|_2^2, \qquad (1)$$

The per-pixel loss is defined as $\ell_{pixel}(\hat{I}_h, I_h) = \|\hat{I}_h - I_h\|_2^2/C/H/W$. We assign the perceptual loss network with random weights or pretrained weights for comparison. The super-resolution results are shown in Figure 1 of the main article. We can see that adding the perceptual loss with pretrained and randomized networks can both combat the blurry results in utilizing the per-pixel loss alone.

These observations suggest that the network structure, instead of the pretrained weights, contributed to the success of the perceptual loss. Through convolution operations in multiple layers, the CNN itself has captured the hierarchical dependencies of variable statistics. By comparing the difference on the perceptual features, a perceptual loss term can be computed to investigate structured dependencies of the inputs. This motivates us to apply the perceptual loss to more structured output learning tasks, in which applying such randomzied perceptual loss is non-trivial.

S2. More Explorations

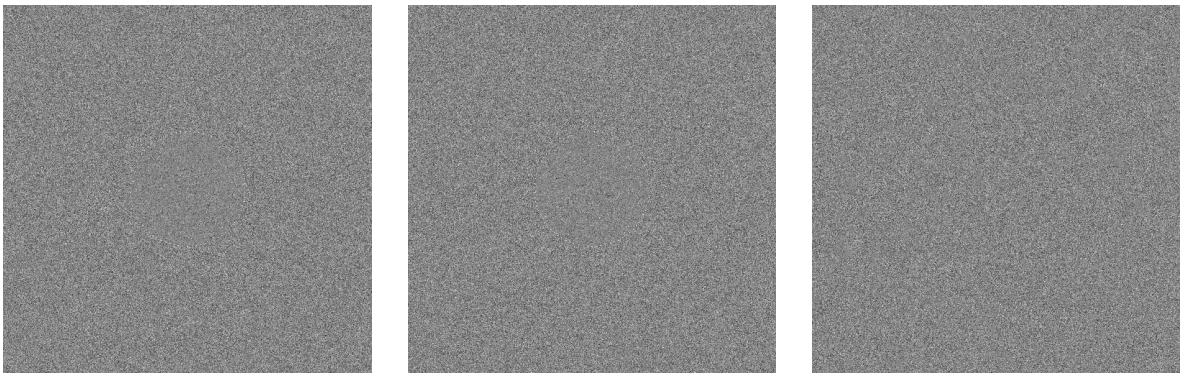
S2.1. Toy Examples

We conduct a toy example to show that the multi layer CNNs have the ability of modeling the structure outputs than a pixel-level L_2 loss. First of all we generate two different backgrounds with the shape of 512×512 using Gaussian noise. The mean is set to 0 and the standard deviation is set to 1. Then, 10000 points with zero values are sampled within the range of a circle with $r = 40$ at the middle of Figure 1a. The exactly same 10000 points are pasted to Figure 1b. Finally, we divided Figure 1a into 16×16 patches, and then random shuffle the patch to generate Figure 1c. Thus, we have a similar pattern of the circle in Figure 1a and Figure 1b.

We evaluate the difference between these images under the pixel-level L_2 loss and the genetic perceptual loss initialized with random weights. We run this toy examples for 50 times. In all cases, under evaluation of the pixel-level L_2 loss, Figure 1a and Figure 1c are closer than Figure 1a and Figure 1b, which is contradicted to the perceptual recognition. However, under the evaluation of the perceptual loss, 37 out of 50 trails successfully recognize the similar pattern in Figure 1a and Figure 1b. This indicates that the multi layer CNNs have a stronger ability of modeling the structure outputs than a pixel-level L_2.

S2.2. Effect of Down-sampling Operators

We also explore the ability of down-sampling operators in each block of the perceptual network. Three different



(a) Low level noise with a Circle **(b)** High level noise with a Circle **(c)** Patch Shuffle of **(a)**

Figure 1 – The figure **(a)** and **(b)** are generated by the Gaussian noise. We then generated a circle and paste it to the middle of the figure **(a)** and **(b)**. We finally shuffled the patch of the figure **(a)**, and get **(c)**. Under the evaluation of the pixel-level L_2, **(a)** and **(c)** are closer than **(a)** and **(b)**, but under the evaluation of the genetic perceptual loss with a random weight network, **(a)** and **(b)** are closer.

ways are usually employed in the network design: max pooling, average pooling and convolutions with stride of 2. The VGG19 is employed as the basic perceptual network, and we change the down-sampling operators in each block. Following Hamed et al [**?**], the mapping ability of the random network is evaluated with the classification accuracy on a simple dataset. We fix all the weights except for the final linear classification layer of the random network, and train the final linear layer for 30 epochs on Cifar10. The classification accuracy can reflect the mapping ability of the random network. 'Mapping' means to transfer the input into a linearly separable space. In Figure 2, we show the correlation between the segmentation mIoU trained with different perceptual networks, and the mapping abilities. Note that the segmentation network is PSPNet with ResNet18 as the backbone, and the baseline without perceptual loss can only achieve 69.6% of mIoU. From Figure 2, we can see the classification accuracy on Cifar10 with fixed random network is higher than the chance rate 10%, which means the random network has the ability of mapping the input into a linearly separable space. Besides, with a max pooling layer, the random perceptual loss can achieve the highest performance. Meanwhile, with a max pooling layer, the mapping ability is also higher than other operators. The reason may because with the fix pattern of finding the max value in local region, the saliency of the structure output is easier to be encoded in the embedding space.

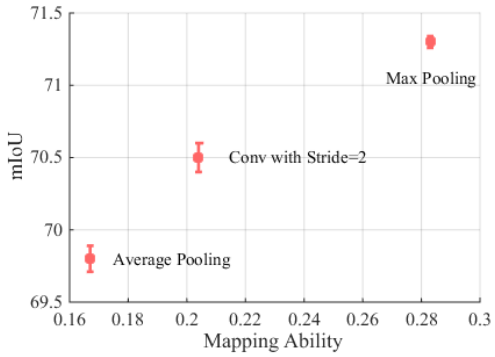


Figure 2 – Effect of different down-sampling operators. The segmentation network is PSPNet18, and the basic perceptual network is VGG19. Mapping ability means the classification accuracy of the random perceptual network on Cifar10.

References

[1] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Trans. Circuits Syst. Video Technol.*, 28(11):3174–3182, 2017.

[2] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. Blendmask: Top-down meets bottom-up for instance segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 8573–8581, 2020.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. Eur. Conf. Comp. Vis.*, pages 801–818, 2018.

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338, 2010.

[6] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2002–2011, 2018.

[7] Adam Gaier and David Ha. Weight agnostic neural networks. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 5364–5378, 2019.

[8] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.

[9] Kun He, Yan Wang, and John Hopcroft. A powerful generative model using random weights for the deep image representation. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 631–639, 2016. 1

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1026–1034, 2015.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 770–778, 2016.

[12] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 125–136, 2019.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei Efros. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1125–1134, 2017.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. Eur. Conf. Comp. Vis.*, pages 694–711, 2016. 1

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. *Comm. of the ACM*, 60(6):84–90, 2017.

[16] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4681–4690, 2017. 1

[17] Guosheng Lin, Chunhua Shen, Anton van den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3194–3203, 2016.

[18] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2604–2613, 2019.

[19] Mihir Mongia, Kundan Kumar, Akram Erraqabi, and Yoshua Bengio. On random weights for texture generation in one layer CNNs. In *Proc. IEEE Int. Conf. Acous., Speech & Signal Process.*, pages 2207–2211, 2017.

[20] Reiichiro Nakano. A discussion of 'adversarial examples are not bugs, they are features': Adversarially robust neural style transfer. *Distill*, 4(8), 2019.

[21] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *Proc. Eur. Conf. Comp. Vis.*, 2012.

[22] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018.

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Proc. Int. Conf. Learn. Representations*, 2015.

[24] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In *Proc. Eur. Conf. Comp. Vis.*, pages 631–648, 2018.

[25] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1–9, 2015.

[26] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *Proc. Eur. Conf. Comp. Vis.*, 2020.

[27] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 9627–9636, 2019.

[28] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018.

[29] Hu Wang, Guansong Pang, Chunhua Shen, and Congbo Ma. Unsupervised representation learning by predicting random distances. *Proc. Int. Joint Conf. Artificial Intell.*, 2019.

[30] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[31] Lijun Wang, Jianming Zhang, Yifan Wang, Huchuan Lu, and Xiang Ruan. Cliffnet for monocular depth estimation with hierarchical embedding loss. In *Proc. Eur. Conf. Comp. Vis.*, pages 316–331, 2020.

[32] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 8798–8807, 2018.

[33] Xinlong Wang, Wei Yin, Tao Kong, Yuning Jiang, Lei Li, and Chunhua Shen. Task-aware monocular depth estimation for 3d object detection. In *Proc. AAAI Conf. Artificial Intell.*, volume 34, pages 12257–12264, 2020.

[34] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic, faster and stronger. *Proc. Advances in Neural Inf. Process. Syst.*, 2020.

[35] Yin Wei, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. *Proc. IEEE Int. Conf. Comp. Vis.*, 2019.

[36] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2021.

[37] Changqian Yu, Yifan Liu, Changxin Gao, Chunhua Shen, and Nong Sang. Representative graph neural network. *Proc. Eur. Conf. Comp. Vis.*, 2020.

[38] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2881–2890, 2017.

[39] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.