Supplementary Material: Transformation Consistency Regularization– A Semi-Supervised Paradigm for Image-to-Image Translation

Aamir Mustafa and Rafał K. Mantiuk

Department of Computer Science and Technology, University of Cambridge, UK
{am2806,rafal.mantiuk}@cl.cam.ac.uk

1 Performance with Perceptual and L_1 Loss

In our experiments, we follow the exact same training protocol as reported in the original papers for image colorization, denoising ans super-resolution. We found that the addition of Transformation Consistency Regularization (TRC) in a semi-supervised fashion results in substantial performance boost when mean-squared error (L_2) is used as a loss function. In this section, we evaluate the efficacy of TRC for image super-resolution application when L_2 loss (Eq. 3 in the main paper) is replaced with perceptual loss [2, 3] or L_1 loss [7].

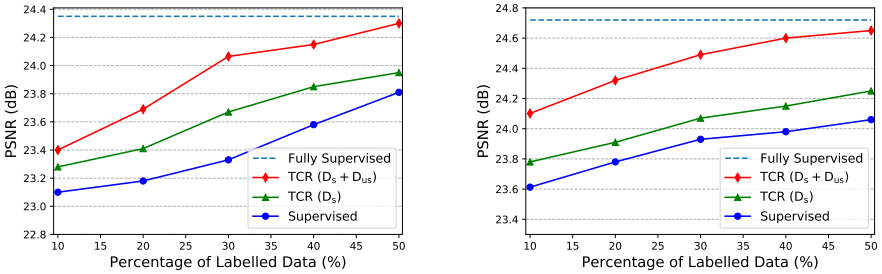
While pixel level L_2 loss is a commonly used SISR protocol, Ledig *et al.* in SRGAN designed a loss function based on the perceptually relevant characteristics of images [4] for image super-resolution. To show TCR is robust to the selection of the loss function, we perform additional experiments relying on using loss function that is closer to the perceptual similarity. For an unlabeled data sample u_i and its geometric transform $T(u_i)$, to compute the perceptual loss, we extract the feature maps from from a pre-trained VGG-19 network $\phi(\cdot)$ [6] and compute the euclidean distance between the two. Mathematically, the perceptual loss $\mathcal{L}_p(u)$ is given by:

$$\mathcal{L}_p(u) = \frac{1}{rB} \sum_{i=1}^{rB} \Big(\frac{1}{M} \sum_{m=1}^{M} \| T_m\big(\phi(f_\theta(u_i))\big) - \big(\phi(f_\theta(T_m(u_i))\big) \|_2^2 \Big) \tag{1}$$

Here $f_\theta(\cdot)$ is the super-resolution model used in our experiments, rB is the number of unsupervised samples fed to the network per batch and $T_m(\cdot)$ is the m-th geometric transformation. To train the model, we equivalently change the supervised loss in Eq. 2 (main paper) to the following:

$$\mathcal{L}_s(x, y) = \frac{1}{B} \sum_{i=1}^{B} \| \phi\big(f_\theta(x_i)\big) - \phi\big(y_i\big) \|_2^2 \tag{2}$$

The feature representations as extracted from deeper layers of the VGG network, which convey more information about the content of the images [5, 2, 4].



(a) Performance improvement using Perceptual loss for training.

(b) Performance improvement using L_1 loss for training.

Fig. 1: The plots shows the efficacy of our method for various types of loss functions for image super-resolution. We show the PSNR values for baseline models (using only supervised data), model trained with addition of our TCR but using only labeled data, and finally models trained using our semi-supervised paradigm. The addition of unsupervised data while training provides substantial improvement in image reconstruction.

Fig. 1a shows a comparison of our semi-supervised scheme using perceptual loss with a baseline model while incrementing the percentage of data used in supervised fashion. Our method consistency results in a significant performance boost compared to its supervised counterpart. Compared to the pixel-level MSE loss, we acheive a lower PSNR value using the perceptual loss, which is consistent with the findings of SR-GAN [4].

We further evaluate the performance of TRC using pixel-wise L_1 loss for training the image super-resolution model. Fig. 1b shows a comparison of our semi-supervised scheme using L_1 loss with a baseline model while incrementing the percentage of data used in supervised fashion.

This goes on to show that the proposed semi-supervised method effectively leverages information from unlabeled data performing equally well under various reconstruction loss terms.

2 Additional Experiments

We further show the potential of our semi-supervised learning method in applications like movie colorization, denoising and generation of high resolution of video clips.

2.1 Movie Colorization

Transformation Consistency Regularization (TRC) can go a long way in colorization of old movies/video clips requiring an artist to colorize only a few frames. To demonstrate the efficacy of our semi-supervised learning (SSL) algorithm for movie colorization we use Blender Foundation's open source short film 'Big Buck

Fig. 2: The figure shows a comparison between our semi-supervised method and the model trained in a completely supervised fashion for colorization. Both the models use only 1% of the training frames as labeled. Our method results in an average absolute gain of 6.0 dB compared to the supervised counterpart. Best viewed in color.

Fig. 3: A snapshot of the video showing comparison between our semi-supervised method and the model trained in a completely supervised fashion for super-resolution (upscale factor 4×). Both the models use only 1% of the training frames as labeled. Our method results in an average absolute gain of 5.7dB compared to the supervised counterpart. Best viewed when zoomed.

Bunny' [1]. We divide the movie into train and test set with each comprising of 510 and 43 seconds respectively. In our SSL settings for image colorization, we made use of only 1% of the total training frames in supervised fashion, while rest were fed into the TCR term. We achieve an absolute gain of 6.0 dB in PSNR using our semi-supervised method as compared to the supervised model trained with the same percentage of training data. Fig. 2 shows a snapshot from the video.

2.2 Movie Super Resolution

Following the same setting, we show how TRC can be used to enhance the resolution of movies by capturing only a few frames at a higher resolution. For this application we use the short film 'Elephants Dream' [1]. The movie frames are divided into train and test sets with each comprising of 600 and 54 seconds respectively. We again use only 1% of the total training frames in supervised fashion. Our method results in an absolute gain of 5.7 dB (for up-scaling 4×) and 2.9 dB (for up-scaling 3×) compared to the supervised baseline. Fig. 3 shows a snapshot from the video for upscale factor 4×.

Fig. 4: Qualitative results showing comparison between reconstructed images using our model and supervised baseline models for single image super-resolution. The column title indicates the percentage of data used for training the model. The last column shows our results where we use only 20% of the entire dataset as labeled and rest in an unsupervised fashion. Best viewed when zoomed.

References

1. Open source movies by blender foundation https://blender.org/
2. Gatys, L., Ecker, A.S., Bethge, M.: Texture synthesis using convolutional neural networks. In: Advances in neural information processing systems. pp. 262–270 (2015)
3. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)
4. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)
5. Mahendran, A., Vedaldi, A.: Visualizing deep convolutional neural networks using natural pre-images. International Journal of Computer Vision **120**(3), 233–255 (2016)
6. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
7. Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Loss Functions for Image Restoration With Neural Networks. IEEE Transactions on Computational

Imaging **3**(1), 47–57 (mar 2017). https://doi.org/10.1109/TCI.2016.2644865, http://ieeexplore.ieee.org/document/7797130/