Supplementary material for:
Style Transfer by Rigid Alignment in Neural Net Feature Space

Suryabhan Singh Hada Miguel Á. Carreira-Perpiñán
Dept. of Computer Science & Engineering, University of California, Merced
http://eecs.ucmerced.edu

November 7, 2020

1 Multi-level style transfer

As mentioned in the main paper (section 4.1), different layers provide different details during style transfer, and we achieve this by cascading the image through different auto-encoders. In figure 1, we show our complete pipeline for multi-level style transfer along with a comparison between single-level and multi-level stylization with the proposed approach.

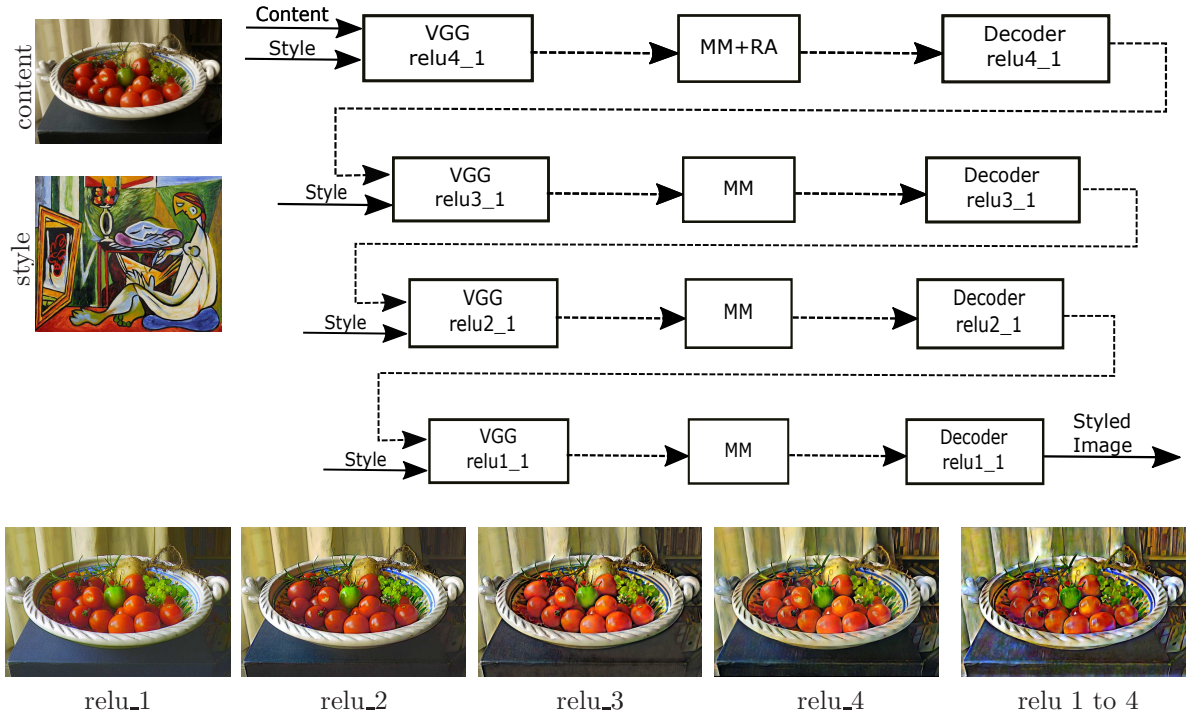

Figure 1: *Top:* Network pipeline of the proposed style transfer method that is similar to Li et al. (2017). The result obtained by matching higher-level statistics of the style is treated as the new content to continue to match lower-level information of the style. MM represents moment matching and RA represents rigid alignment. *Bottom:* Comparison between single-level and multi-level stylization with the proposed approach. The first four images show styled images created by applying moment matching and rigid alignment to individual VGG features. The last image shows stylization results by applying multi-level stylization, as shown in the above network pipeline.

2 Decoder training

As mentioned in the paper (section 5.1), we use a pre-trained auto-encoder network from Li et al. (2017). This auto-encoder network has been trained for general image reconstruction. The encoder part of the network is the pre-trained VGG-19 (Simonyan and Zisserman, 2015) that has been fixed, and the decoder network ($\mathbf{D}$) is trained to invert the VGG features to image space. As mentioned in Li et al. (2017), the decoder is designed as being symmetrical to that of the VGG-19 network, with the nearest neighbor up-sampling layer used as the inverse of max pool layers. Li et al. (2017) trained five decoders for reconstructing images from features extracted at different layers of the VGG-19 network. These layers are *relu5_1, relu4_1, relu3_1, relu2_1,* and *relu1_1*. The loss function for training involves pixel reconstruction loss and feature loss (Dosovitskiy and Brox, 2016):

$$\arg\min_{\theta} \|\mathbf{X} - \mathbf{D}_\theta(\mathbf{z})\|_2^2 + \lambda \|\Phi_l(\mathbf{X}) - \Phi_l(\mathbf{D}_\theta(\mathbf{z}))\|_2^2 \tag{1}$$

where θ are the weights of the decoder $\mathbf{D}$. $\mathbf{X}$, $\mathbf{z}$ are the original image and corresponding VGG features, respectively, and $\Phi_l(\mathbf{X})$ is a VGG-19 encoder that extracts features from layer l. In addition, λ is the weight to balance the two losses. The decoders have been trained on the Microsoft COCO dataset (Lin et al., 2014). However, unlike Li et al. (2017), we use only four decoders in our experiments for multi-level style transfer. These decoders correspond to *relu4_1, relu3_1, relu2_1,* and *relu1_1* layers of the VGG-19 network.

3 Spatial control

Spatial control is needed to apply different styles to different parts of the content image. A set of masks $\mathbf{M}$ are additionally required to control the regions of correspondence between style and content. By replacing the content feature $\mathbf{z}_c$ in section 4 of the main paper with $\mathbf{M} \odot \mathbf{z}_c$, where $\odot$ is a simple mask-out operation, we can stylize the specified region only, as shown in figure 2.

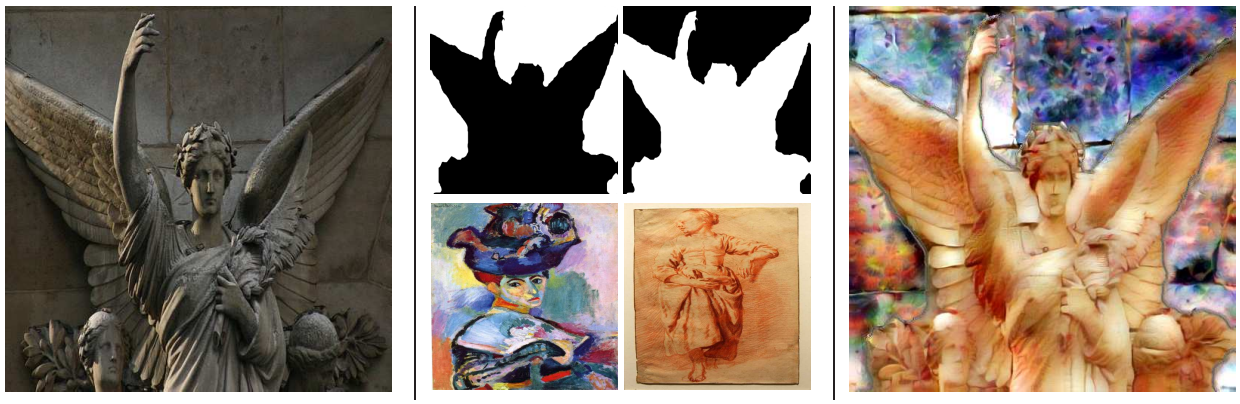

Figure 2: Spatial control in style transfer. *Middle column:* In the top row are the binary masks, and corresponding styles are in the bottom row.

4 Need to arrange features in $\mathbb{R}^{C\times HW}$ space

As mentioned in the paper (section 4), for alignment, we consider the deep neural network features ($\mathbf{z} \in \mathbb{R}^{C\times H\times W}$) as a point cloud which has C points each of dimension HW. We can also choose another configuration where each point is in $\mathbb{R}^C$ space, thus having HW points in the point cloud. In figure 3, we show a comparison of style transfer with the two configurations. As shown in the figure 3, having the later configuration results in complete distortion of content structure in the final styled image. The reason for that is deep neural network features (convolution layers) preserve some spatial structure, which is required for style transfer and successful image reconstruction. Therefore, we need to transform the features in a specific manner so that we do not lose the spatial structure after alignment. That is why, for alignment, we transform $\mathbf{z}$ such that the point cloud has C points each of dimension HW.

Figure 3: *Third column:* style transfer with features ($\mathbf{z}$) transformed as C cloud points and each in $\mathbb{R}^{HW}$ space. *Fourth column:* style transfer with HW cloud points and each in $\mathbb{R}^C$ space.

5 More styled Results

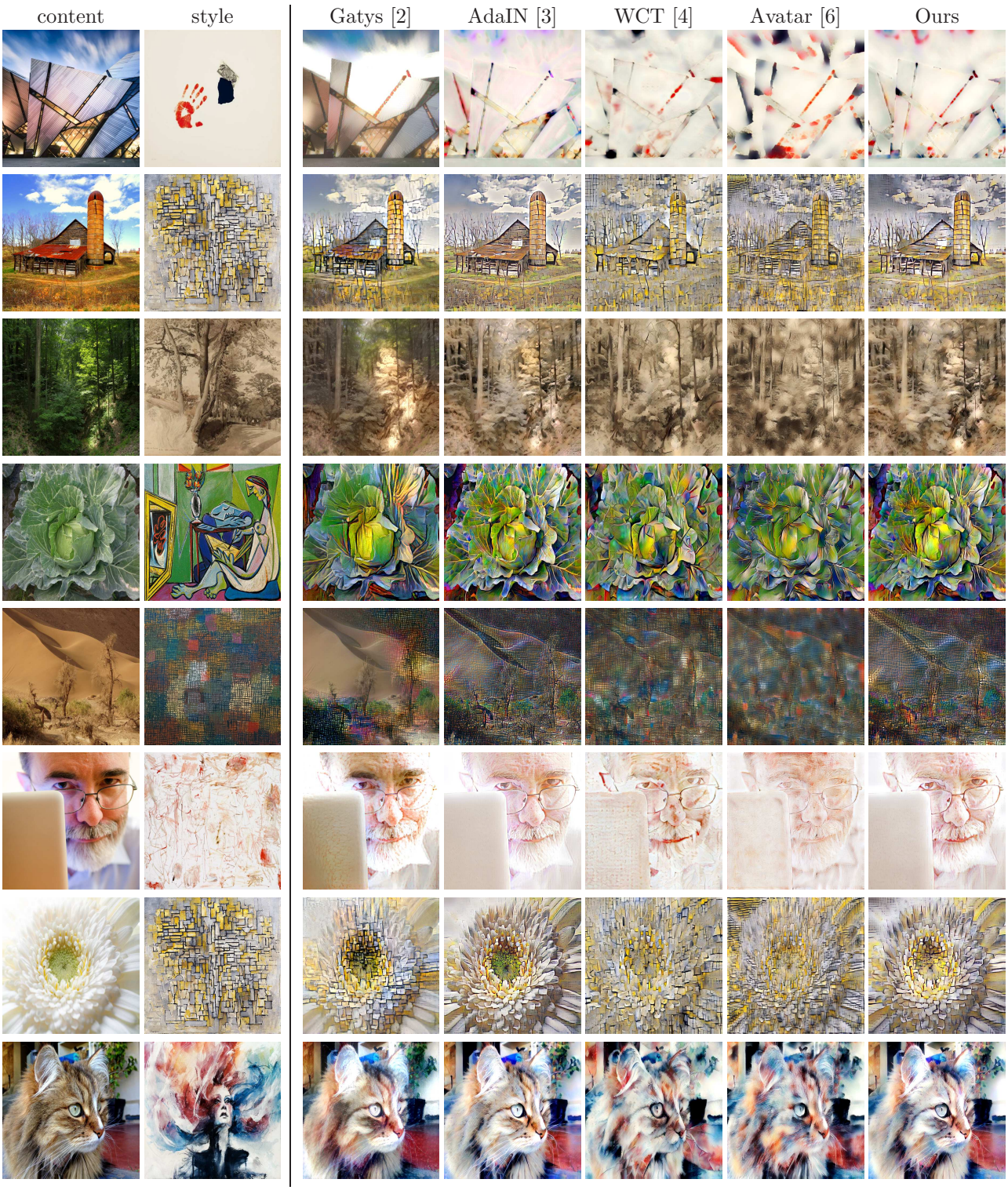

Figure 4: Figure shows comparison of our style transfer approach with existing work.

Figure 5: Figure shows comparison of our style transfer approach with existing work.

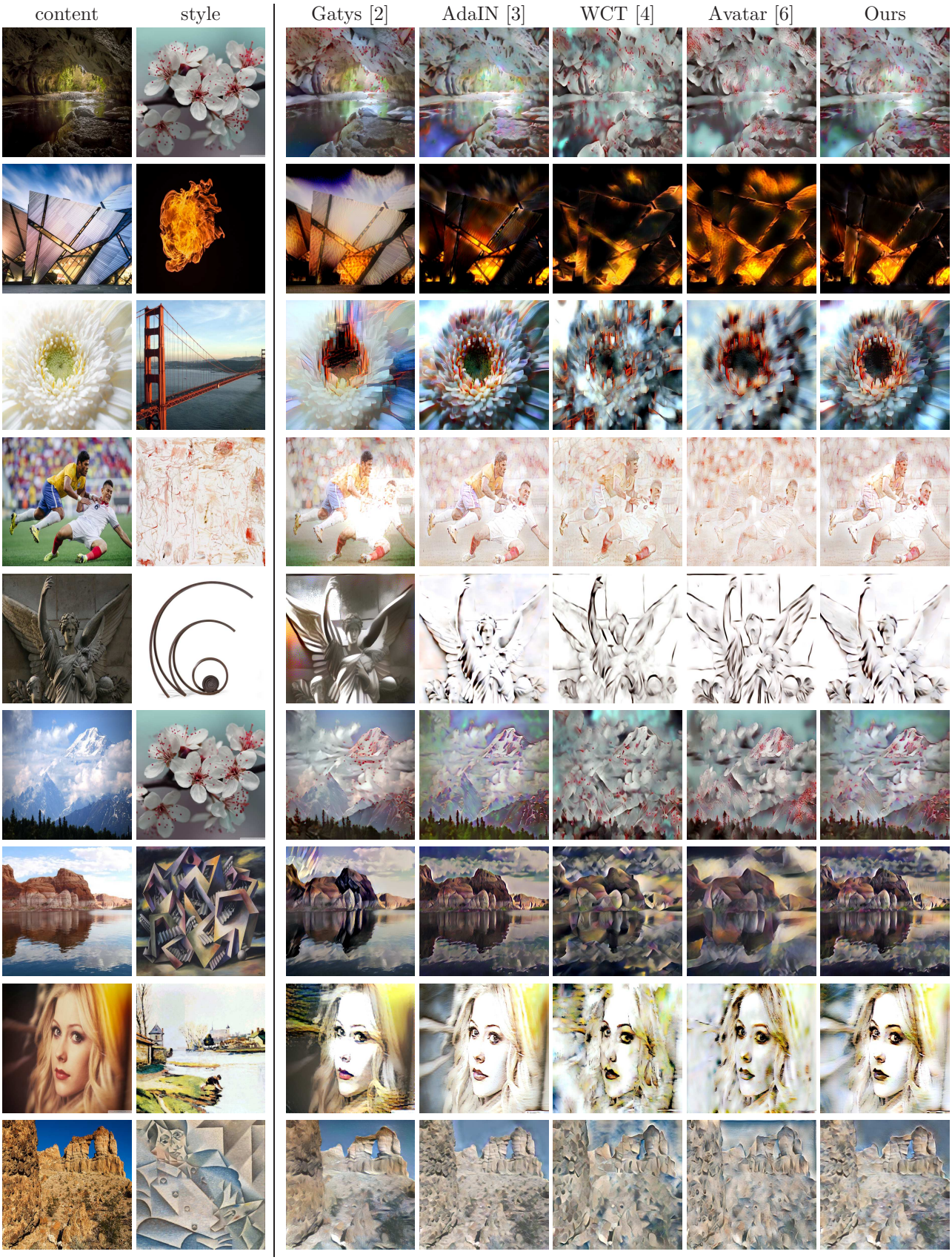

Figure 6: Figure shows comparison of our style transfer approach with existing work.

References

A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 29, pages 658–666. MIT Press, Cambridge, MA, 2016.

L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proc. of the 2016 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR'16)*, pages 2414–2423, Las Vegas, NV, June 26 – July 1 2016.

X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. 17th Int. Conf. Computer Vision (ICCV'17)*, Venice, Italy, Dec. 11–18 2017.

Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang. Universal style transfer via feature transforms. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 30, pages 386–396. MIT Press, Cambridge, MA, 2017.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *Proc. 13th European Conf. Computer Vision (ECCV'14)*, pages 740–755, Zürich, Switzerland, Sept. 6–12 2014.

L. Sheng, Z. Lina, J. Shao, and X. Wang. Avatar-Net: Multi-scale zero-shot style transfer by feature decoration. In *Proc. of the 2018 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR'18)*, Salt Lake City, UT, June 18–22 2018.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. of the 3rd Int. Conf. Learning Representations (ICLR 2015)*, San Diego, CA, May 7–9 2015.