Novel View Synthesis with Style Transfer via 3D feature embeddings

Shubham Agrawal Somendra Tripathi Najim Yaqubie

Abstract

The advent of adversarial models has given rise to very powerful image synthesis. However, despite the rich depth in 2D image models, mathematical models of 3D environments have not been explored thoroughly. We propose to apply a hybrid generational model for style transfer upon an underlying 3D scene structure. We apply a standardized dataset and quantitative metrics across multiple models in order for direct comparison.

1. Introduction

Recent years have seen considerable progress in applying machine learning techniques to develop a 3D representation of a particular scene from 2D images. With the introduction of efficient 3D representations such as DeepVoxels (Sitzmann et al., 2019a), Scene Representation Networks (SRNs) (Sitzmann et al., 2019b) and Neural Meshes (Kato et al., 2018a), these neural networks are able to generate novel views of an object learned from a set of 2D images.

Which type of 3D representation is most appropriate for novel view synthesis, transformations and style transfer? Polygon mesh approaches are promising because they are scalable and have surfaces for lighting and textures. Polygon meshes are sets of vertices and surfaces that are incredibly compact, can represent 3D shapes with a small number of parameters and are simple to transform.

In contrast, voxel grid approaches are memory and computationally expensive limiting their scalability. However, because they are regularly sampled from 3D space, they are more expressible representations of the objects. As natural extensions of 2D pixels, voxel representations can be processed by Convolutional Neural Networks (CNNs) and therefore take advantage of extensive progress made in understanding 2D scenes.

Little work has been done in transforming these underlying 3D structures in meaningful ways. Can we, for instance, apply a new floral pattern to a 3D model of shape? We aim to explore the style transfer of a 2D image upon latent 3D models for novel view synthesis.

2. Related Work

Our approach incorporates multiple active research areas such as novel view synthesis via 3D scene representations, style transfer through CNN's, and perceptual metrics.

2.1. Novel View Synthesis

There are many deep neural networks proposed to solve novel view synthesis that are multi-view consistent. Many techniques use various underlying structures such as voxel grids (Sitzmann et al., 2019a), continuous functions (Sitzmann et al., 2019b), or approximate gradients (Kato et al., 2018a). Approximate gradient methods using neural meshes address the large memory footprint of voxel grids at a cost of losing expressibility of certain features. We will measure the flexibility of these different approaches.

2.2. Style Transfer

Deep models for 2D image synthesis and manipulation have shown promising results in generating photo-realistic images (Karras et al., 2019). In other cases, deep neural networks separate and recombine content and style of arbitrary images to create artistic views (Gatys et al., 2016). Previous work has applied 2D style transfer to generated novel views from neural meshes to limited success.(Kato et al., 2018a) Through propagating style loss into the neural mesh network, content loss increased and resulting novel views show deformations. We will follow previous methods of style transfer to synthesize novel views and quantitatively measure content loss vs style loss in resulting views.

2.3. Perceptual Metrics

Previous works have shown that features of deep neural networks trained for object classification are a better perceptual metric then classic metrics (Zhang et al., 2018). Classic perceptual metrics such as PSNR, SSIM, MSSIM, FSIM, and HDR-VDP (Lin & Kuo, 2011) rely on shallow, hand-crafted functions and fail to capture many nuances of perception.

For evaluation of our model, keeping the style-network, optimization algorithm and the number of optimizations steps fixed, we will compare the average style loss of synthesized renderings over multiple viewpoints. We will also compare models with respect to the number of optimization steps and time required to reach a threshold style loss.

3. Methodology

We extend neural meshes to include perceptual metrics as guide to efficacy. We first recreate the style transfer using neural meshes and apply quantitative metrics to novel views (Kato et al., 2018a). Then, we use our ShapeNet dataset to establish a baseline of comparable inputs and outputs.

3.1. Data

We use four datasets. First, we use the partial dataset made available by Kato which consists of 2 3D objects, a teapot and a bunny (Kato et al., 2018b). Secondly, we use a group of 50 512x512 images for style transfer (Kato et al., 2018b). When working with DeepVoxels, we will reuse their dataset that consists of both 480 1920x1080 images captured with a DSLR and 480 synthetic 512x512 views of 3D objects (Sitzmann et al., 2018). All images have metadata associated with respect to perspective.

Finally, we extracted 15 objects from various categories from ShapeNetCore to use with both frameworks (ShapeNet, 2019a). ShapeNet has 3D models that can be converted into meshes as well as voxel grids. We will use ShapeNet-Viewer to generate a series of synthetic 512x512 2D images to be compatible with DeepVoxels (ShapeNet, 2019b). We will verify these poses and angles are sufficient for training. By using the same dataset from ShapeNetCore in both architectures, results will be directly comparable.

3.2. Neural 3D Mesh Renderer

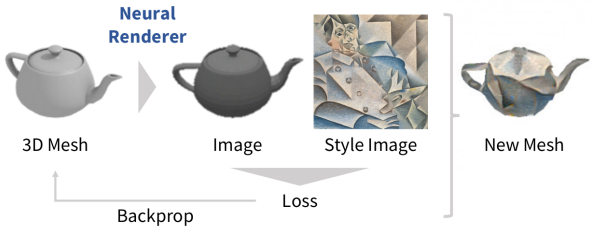

Figure 1. Overview of model components for 2D-to-3D style transfer using Neural Meshes (Kato et al., 2018a)

Using neural networks to generate novel views from meshes was not possible due to rasterization, a discrete non-differentiable operation that converts vectors into pixels. However, Neural 3D Mesh Renderer approximates rasterization gradients to develop a neural network capable of novel view synthesis (Kato et al., 2018a). Using this network allows for 3D mesh editing operations such as 2D-to-3D style transfer (see Figure 1). We use this network with trained meshes to backpropogate style loss onto the mesh and generate stylized novel views.

3.3. Style Loss Network

After training the 3D neural mesh, we then apply style transfer. First, we generate a set of novel views, compute a combined loss function using our style image as a reference, and backpropogate this loss back into the neural renderer. To ensure the underlying 3D mesh is not deformed during the style transfer, we use a weighted loss function consisting of content, style and regularization loss.

For calculating content loss, they assume that vertices-to-faces relationships $\{f_i\}$ are same for both meshes m and m_c (input mesh). Content loss is then calculated as $L_c(m|m^c) = \sum_{\{v_i, v_i^c \in (m,m^c)\}} |v_i - v_i^c|^2$. Similar to traditional image style transfer models, style loss is defined as $L_s(m|x^s, \phi) = |M(f_s(R(m,\phi))) - M(f_s(x_s))|_F^2$. Here $M(.)$ is gram matrix of feature vectors $f_s(.)$ from rendering $R(m,\phi)$ of mesh m at a particular camera position ϕ. $f_s(.)$, our style feature vector, is calculated using our style transfer network. A regularization loss of $L_t(m|\phi) = \sum_{\{\boldsymbol{p}_a, \boldsymbol{p}_b\} \in P} |\boldsymbol{p}_a - \boldsymbol{p}_b|_2^2$ is added to regularize adjacent pixel colors to be similar.

3.4. DeepVoxels with Style Transfer

We propose applying a similar style loss network to a DeepVoxels approach of 3D representation (Sitzmann et al., 2019a). Given that voxels based image generation is differentiable, we do not have to approximate similar to neural meshes. DeepVoxels has demonstrate high quality view synthesis that is encouraging towards applying style transfer with less deformation in the form of content loss. We aim to apply 2D-to-3D style transfer by generating novel views, applying style transfer, and backpropogating our combined style loss function (see figure 2). Since we do not have access to a mesh-based content loss function, we will explore using other 2D perceptual metrics or reusing the loss function within the DeepVoxels architecture.

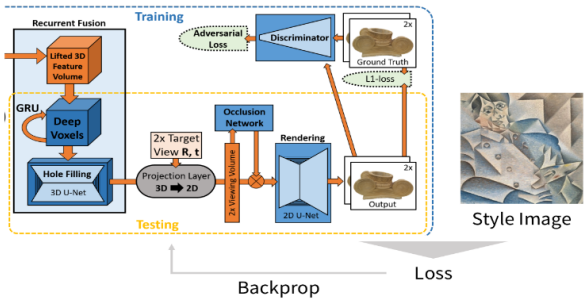

Figure 2. Overview of model components for 2D-to-3D style transfer using Neural Meshes (Kato et al., 2018a)

STYLE	OBJECT	CONTENT	STYLE	TIME
SIMPSONS	TEA POT	52.55	11.28×10^6	790
	BUNNY	42.81	70.79×10^6	1188
	URN	57.47	89.93×10^6	2498
FUNGAI	TEA POT	2.07	4.06×10^6	811
	BUNNY	1.45	3.02×10^6	1200
	URN	1.64	2.95×10^6	2480
PERRY	TEA POT	11.49	13.16×10^6	798
	BUNNY	5.29	7.78×10^6	1192
	URN	7.72	9.67×10^6	2467
NEW YORK	TEA POT	23.06	37.92×10^6	784
	BUNNY	12.35	24.91×10^6	1185
	URN	14.80	26.80×10^6	2488

Table 1. Content and Style loss for each style and object

Figure 3. Some styles are clearly harder than others. Here we use Simpsons vs Fungai to illustrate content loss and deformation of object during novel view synthesis.

4. Results

We have run the Neural 3D Mesh Renderer with a style transfer network on the original neural meshes dataset as well as objects from various categories from ShapeNet.

Models are trained on a virtual machine hosted by Google Cloud Platform with 2 virtual CPUs, 13GB of memory, 250GB of space and a NVIDIA Tesla K80.

4.1. Style Hardness

Not all style transfer is successful. We noticed, as can been seen in Table 4, that different styles have different difficulties. As style loss grows, content loss also follows given our linear loss function. We notice in Figure 3 that the urn is extremely deformed corresponding to a higher content loss. However, the urn is not as deformed when applying the fungai style corresponding to a lower content loss. Given this relationship, we can then grade our style dataset of varying difficulty levels for further research.

Another relationship we noticed is that objects also have varying difficulty. For example, the tea pot consistently has higher content loss values regardless of style compared to other objects. Despite having higher content loss, manually checking the synthesized novel view does not show as large deformations when compared to the urn. This is likely due to the overall size of smooth surfaces on the object compared to distinct features.

4.2. ShapeNet extension

When applying various new objects from ShapeNetCore such as the urn, relative loss was within range of previous examples(ShapeNet, 2019a) . We have successfully extended 3D style transfer using Neural Meshes to a much larger dataset than previously explored. By doing so, we can now directly compare results with other research.

Figure 4. Applying style transfer to original dataset vs ShapeNet dataset is comparable.

5. Future Work

We aim to apply similar methods to DeepVoxels and quantitatively compare style transfer results with standardized metrics using the same underlying objects.

6. Contributions

We summarize 3 main contributions of our work. First we defined a hardness scale for 3D Style Transfer for a dataset of 50 images. Then, we extended Neural Meshes to ShapeNet's objects for future standardized research. Lastly, we quantified loss with 3D style transfer.

All relevant code, gathered data, and results are accessible via https://github.com/najerama/3dstyletransfer.

References

Gatys, L. A., Ecker, A. S., and Bethge, M. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2414–2423, 2016.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Kato, H., Ushiku, Y., and Harada, T. Neural 3d mesh renderer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3907–3916, 2018a.

Kato, H., Ushiku, Y., and Harada, T. Neural mesh data, 2018b. URL https://github.com/hiroharu-kato/style_transfer_3d.

Lin, W. and Kuo, C.-C. J. Perceptual visual quality metrics: A survey. *Journal of visual communication and image representation*, 22(4):297–312, 2011.

ShapeNet. Shapenet core data, 2019a. URL https://www.shapenet.org/about#download.

ShapeNet. Shapenet model viewer, 2019b. URL https://github.com/ShapeNet/shapenet-viewer.

Sitzmann, V., Thies, J., Heide, F., Nießner, M., Wetzstein, G., and Zollhofer, M. Deepvoxel dataset, 2018. URL https://vsitzmann.github.io/deepvoxels/.

Sitzmann, V., Thies, J., Heide, F., Nießner, M., Wetzstein, G., and Zollhofer, M. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2437–2446, 2019a.

Sitzmann, V., Zollhöfer, M., and Wetzstein, G. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *arXiv preprint arXiv:1906.01618*, 2019b.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.