CEIG – Spanish Computer Graphics Conference (2019)
A. Jarabo and D. Casas (Editors)

Structure-preserving style transfer

Santiago Calvo, Ana Serrano, Diego Gutierrez, and Belen Masia
Universidad de Zaragoza

Figure 1: *Our approach allows to perform style transfer in an image while preserving fine structures and details. Left: Input images, including a* style *image (*Starry Night*, top), and a* content *image (*Tübingen*, bottom), to which the style is to be transferred. Middle: Our result for style transfer, which successfully transfers the style (orange inset), while preserving details (blue inset) and preventing highly distorted image features. Right: Results by the seminal technique of Gatys et al. [GEB16], generated by tuning the weights of content and style in the final image: there is an inherent trade-off between preserving content features, and transferring the desired style.*

Abstract

*Transferring different artistic styles to images while preserving their content is a difficult image processing task. Since the seminal deep learning approach of Gatys et al. [GEB16], many recent works have proposed different approaches for performing this task. However, most of them share one major limitation: a trade-off between how much the target style is transferred, and how much the content of the original source image is preserved [GEB16, GEB*17, HB17, LPSB17]. In this work, we present a structure-preserving approach for style transfer that builds on top of the approach proposed by Gatys et al. Our approach allows to preserve regions of fine detail by lowering the intensity of the style transfer for such regions, while still conveying the desired style in the overall appearance of the image. We propose to use a quad-tree image subdivision, and then apply the style transfer operation differently for different subdivision levels. Effectively, this leads to a more intense style transfer in large flat regions, while the content is better preserved in areas with fine structure and details. Our approach can be easily applied to different style transfer approaches as a post-processing step.*

1. Introduction

Style transfer techniques aim to extract the distinctive texture or features of a source image (its style), and apply them to a target content image. While the quality of the final result is subjective, and can depend on the final goal or application, most commonly the resulting image should have the identifiable style of the source, while preserving the content features of the target so that it is still recognizable. In this work, following common denomination, we will refer to the source image, the one that provides the style, as the *style* image, and to the target image, the one that provides the content and to which the style is to be transferred, as the *content* image.

While there has been techniques addressing the problem of style transfer for decades, there was an inflection point in the field with the seminal work of Gatys et al. [GEB16]. In their work, they proposed a new method for style transfer based on deep neural networks (DNNs). They showed that DNNs can encode not only the content but also the style information of an image, and that the style and content components are somewhat separable: it is possible to change the style of an image while preserving its content. In their work, Gatys et al. use a DNN to extract the features of a source (*style*) image, and at the same time, the same neural network is used to extract the content features of the target (*content*) image. These features are used to generate a new image that pre-

DOI: 10.2312/ceig.20191200

serves the structure of the content image, while its style follows that of the source image, as enforced by each of the two terms of the loss function proposed to train the DNN (one term for style, and one term for content preservation). This method suffers from a major limitation: there is a content-style trade-off. The balance between content and style is given by the ratio between the weight of the two terms in the loss function. Varying this ratio allows to modulate the *intensity* of the style transfer process: relative low weights of the style term (lower intensities of the style transfer) do not properly convey the desired style; meanwhile, high weights of the style term (high intensities of the style transfer) cause distortions in regions with fine details (see Figure 1). Other novel approaches also suffer from the same content-style trade-off, both for images [GEB*17, HB17, LPSB17], and video [RDB18]. While some works have targeted this trade-off explicitly, they offer more complex solutions than ours, which are usually harder to generalize, as discussed in Section 2.

In this paper we propose a simple yet effective approach for this trade-off. Our method consists on performing a quad-tree subdivision on the content image: smaller patches will be made for regions of the content image with fine details, while larger patches will contain more uniform areas of the content images. Then, we apply different style transfer intensities to each of these patches: in smaller patches, containing fine details, we favor preservation of image content, while larger patches will receive a higher style transfer intensity, in order to convey the desired style. We tested our approach for the work proposed by Gatys et al. [GEB16], but it can be easily adapted for similar approaches. We believe that the simplicity of our technique is an advantage; to our knowledge, this approach has not been proposed before, and offers a fast, easy-to-implement solution for the limiting trade-off suffered by these techniques.

2. Related Work

Our method is built upon the work of Gatys et al. [GEB16], also known as neural style transfer. Their method is similar to previous proposals in the intent to extract the structure and content features of an image, and the style features of another image, to yield the final result. The key difference of Gatys et al.'s work with respect to previous methods is that they use a deep neural network (DNN) to extract these features from the images. Prior to this work of Gatys et al. [GEB16], the foundation for it was laid by introducing the idea of using filters and operations of DNNs trained for pattern recognition as a way to extract and synthesize the texture and style features of a source image [GEB15]. Later works will adapt this approach to perform style transfer in videos [RDB16].

A number of follow-up works stemmed from the work of Gatys et al., trying to alleviate its limitations, such as speed and memory consumption [ULVL16], or improve the results, both for images [GEB*17, HB17, LPSB17], and video [RDB18], but they all suffer from the aforementioned content-style trade-off. We thus focus here on those which have tried to address this trade-off. Rujie et al. [Yin16] also build on the system of Gatys et al.; in their case, they address the issue of having significantly different spatial resolutions between the style and content image, and they propose to segment the content image into regions semantically. This process requires heavy and non-trivial manual intervention, and results are only shown for cases in which the semnatics of the content and style images are the same (e.g., both illustrate a bird), and they are thus both segmented similarly and then merged. Frigo et al. [FSDH16] propose a technique for unsupervised style transfer, without any learning, in which they pose the problem of style transfer as a local texture transfer followed by a global color transfer; similar to us, they do an adaptive partition of the content image, although in their case it is used to do their local texture transfer step. The work of Chen et al. [CH16] generalizes the original neural style transfer to obtain what they term content-aware style transfer; with it, they can select what content information to include in the style transfer, either based on a semantic segmentation or on saliency information. In contrast, we do not need to rely on segmentation or saliency techniques, which may hinder robustness of the technique. Most recently, Sanakoyeu et al. [SKLO18] have significantly departed from the original neural style architecture by training a style-aware content loss jointly with a deep encoder-decoder network. While their system can naturally preserve fine details better than the original neural style transfer, our advantage is that we can offer a degree of control over the style-content trade-off that their method cannot. Finally, Li et al. [LTX18] focus on decomposing and analyzing the style, as opposed to the content, by investigating the style features produced by the CNN.

3. Background on neural style transfer

Our work takes as a starting point the work presented by Gatys et al. [GEB16]. This method introduces a convolutional neural network (VGG [SZ14]) to extract style features of a source image, and content features or structure from a target image. These features are combined to create a new image which will acquire the style of the source image, while keeping the image content of the target image. The key idea of this system is that it is possible to extract both the structure and style information from an image from the different layers of a convolutional neural network. For completeness, we include in this section a summary of their proposed method, for more details please refer to the original work of Gatys et al. [GEB16].

The target content image, **T**, is run through the network, and its structure features are extracted from the layer *conv*4_2. Similarly, the style source image, **S**, is run trough the network, and its style features are computed as the mean of the output of 5 different layers ($conv1\_1, conv2\_1, conv3\_1, conv4\_1, conv5\_1$). An additional image is provided as seed **X**. To transfer both style (from image **S**), and content (from image **T**) into the seed image **X**, the distances between the content features of the seed image and the content image, and between the style features of the seed image and the style image, are jointly minimized. This minimization is presented as a loss function in the form:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{content} + \beta \mathcal{L}_{style} \quad (1)$$

where α and β are the content and style weights, which allow to control the influence of the *content* term $\mathcal{L}_{content}$ (content structure of **T**), and the *style* term $\mathcal{L}_{style}$ (style features of **S**) terms, giving as a result different outputs according to the different ratios α/β applied.

The content term is defined as:

$$\mathcal{L}_{content} = \frac{1}{2}\sum_{i,j}(F^l_{i,j} - P^l_{i,j})^2 \quad (2)$$

where $P^l_{i,j}$ and $F^l_{i,j}$ are the activations of the i_{th} filter at a position j in layer l (in this case, $conv4\_2$) for the original image and the image being generated, respectively. The content term is formulated as the squared error loss between the feature representations of the two images.

The style term is defined as:

$$\mathcal{L}_{style} = \sum_{l=0}^{L}(w_l E_l) \quad (3)$$

where E_l is the contribution of layer l to the total loss term. In this case, as in Gatys et al., the style term takes into account five layers ($conv1\_1, conv2\_1, conv3\_1, conv4\_1, conv5\_1$) weighted by the term $w_l = \frac{1}{5}$, so all layers contribute equally. E_l can be expressed as:

$$E_l = \frac{1}{4N_l^2 M_l^2}\sum_{i,j}(G^l_{i,j} - A^l_{i,j}) \quad (4)$$

where N_l represents the different filters of size M_l that belong to layer l. A and G are the style representations for the original image and the image being generated, respectively. In particular, each of these terms is defined as the inner product between the vectorised feature maps i and j in layer l, which represent the correlations between the different filter responses. Intuitively, these correlations between multiple layers will capture the common features across different scales of the image (which can be interpreted as style), rather than capturing the global arrangement of the image.

Our goal is to achieve an output image where the structure content features are well preserved, including small details, lines and edges, but that is able to convey the style of the source image. Finding an α/β ratio that fulfills this purpose is challenging [GEB16, HB17, LPSB17], since there is a trade-off between preserving content and transferring style that is inherent to the loss function (Eq. 1). This trade-off can be seen in Figure 2: as the style weight β increases (and α remains constant), small details progressively fade.

Figure 2: *Content-style trade-off. Increasing the intensity of the style transfer (β), results in degradation of small details in the content image.*

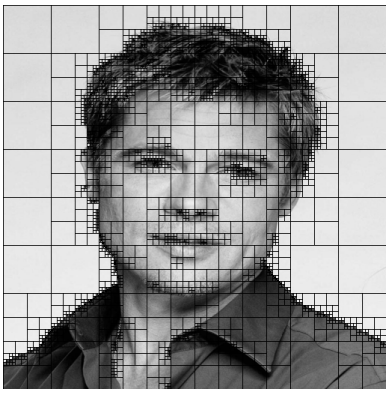

Figure 3: *Quad-tree decomposition of the content image. Flat regions result in larger patches, while areas with fine structure and details get further subdivided into smaller patches.*

Figure 4: *Left: Source style (Musicians) and content (Tubingen) images. Center: Resulting image after applying our method. Right: Enlarged area of the image showing our method with (bottom), and without (top) taking into account overlapping patches for blending.*

4. Structure-preserving image style transfer

We present a simple approach that strives to preserve the image details of the target image while still conveying the source image style. We propose a selective image style transfer, applying different values of the ratio α/β in different regions of the image, according to the amount of details present in such regions: regions with many details will be better preserved, while plain regions (such as empty planes, background, or skies) will be more stylized. In addition to this, we use the content image as seed image (as opposed to white noise used in the original neural style transfer), which leads to a faster convergence.

Our method is based on splitting the content image in patches of different sizes, with smaller patches corresponding to very detailed regions, and large patches corresponding to plain regions. We use a quad-tree decomposition of the content image that iteratively splits the image in squared patches with decreasing size in log_2 scale, with a subdivision criterion based on the homogeneity of the patches. In each step, we compute the difference between the maximum and minimum pixel values inside of each square, and we compare it with a previously determined threshold. If the difference is higher than this threshold, the patch is subdivided again (image patches can be divided until the blocks are as small as 1-by-1 pixels). For all our results, this threshold is set to 0.1, this implies that we will get very small squares for fine details as shown in Figure 3.

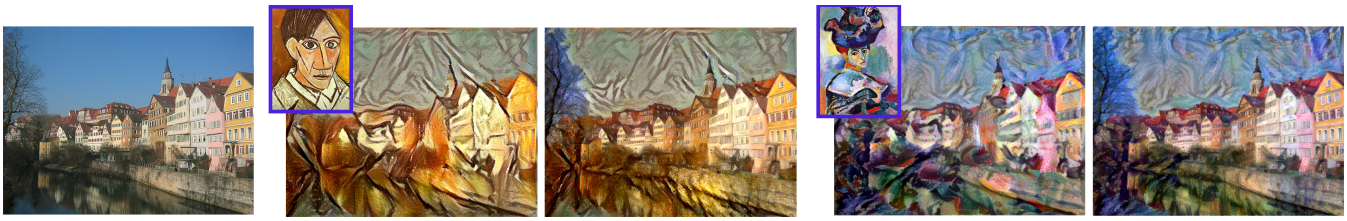

Figure 5: *Results using the* Tübingen *image as content image, and two different style images. For both results, the image on the left is the result from the original neural style transfer, and the image on the right is our result. Our ability to preserve fine structure can be seen, e.g., in the windows.*

Figure 6: *Results using the* Brad *image as content image, and two different style images. For both results, the image on the left is the result from the original neural style transfer, and the image on the right is our result. Our ability to preserve fine structure can be seen, e.g., in the eyes and other facial features.*

This method is very simple, but it has been proven to be very effective for several image processing tasks, including compression or segmentation tasks [SHB14]. We show an example of this decomposition in Figure 3: Once the process is completed, some very detailed regions, like the eyes, will be split in small patches. The α/β ratio applied to these regions will be higher (thus better preserving content) than the ratio applied to those image regions with large patches, such as the background of the image.

After applying the selective style transfer to all the patches, the next step is to merge this stylized patches. Due to the division in patches and the different ratios applied to them, simply joining the patches together causes blocking artifacts (Figure 4). We solve this problem by implementing a blending method that consists on expanding the boundaries of the stylized patches, leading to an overlapping area among neighboring patches. For our experiments, we have empirically found that expanding each of the boundaries of the image patch by 1 pixel is enough to produce satisfactory results. Then, for merging the overlapped regions, we compute the mean of the corresponding stylized patches (all overlapping patches are weighted equally for this operation), yielding smooth transitions while preserving details.

Resulting images obtained using this technique are shown in Figures 5, and 6, together with results from the original neural style transfer for comparison purposes. We show results for different style and content images, illustrating how we are able to generate a result that preserves fine structures while conveying the desired style. We additionally show in Figure 7 that our technique is amenable to texture transfer, enabling a better adaptation to the underlying content.

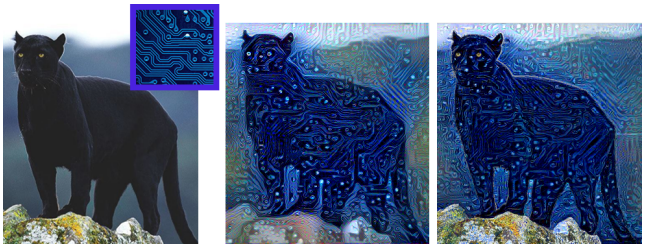

Figure 7: *From left to right: Input content and style images, result by the original neural style transfer, and our result. Although the line between texture and style transfer is fuzzy, this result aims to illustrate that our technique can also be beneficial in a texture transfer scenario.*

Extension to video. In order to handle video, we simply run our technique to perform style transfer in one out of every three frames. Then, we compute optical flow [CN*10] for the other two frames, and use it to propagate the style transfer results; Figure 9 illustrates the process. This procedure yields temporally plausible results at a reduced cost (only one in three frames is processed). We use bidirectional motion estimation for computing the motion vectors over each pair of stylized frames. Then, we reconstruct two motion compensated intermediate frames according to the estimated motion.

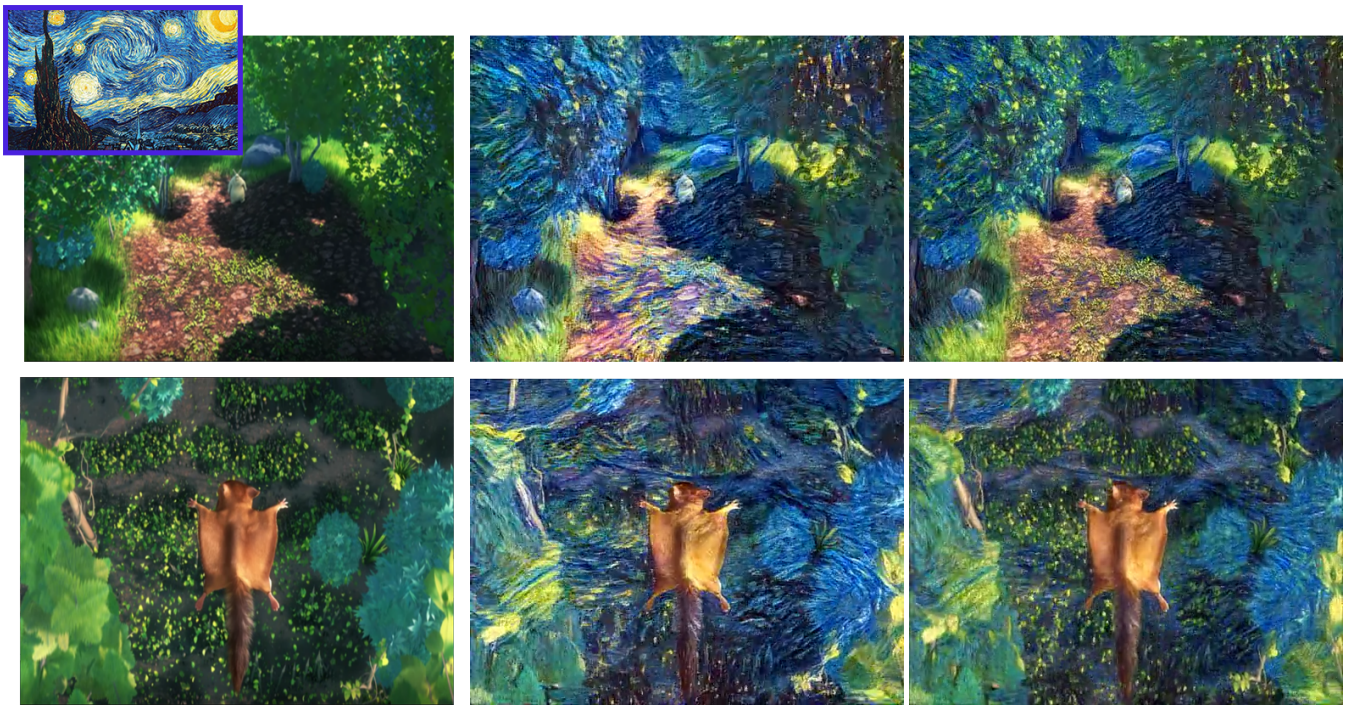

Figure 8: *Results using the* Starry Night *image as style image, and two different frames from the movie Big Buck Bunny as content images. From left to right: original content image, results from the original neural style transfer, and our result.*

Example frames of our technique applied to video are shown in Figure 8.

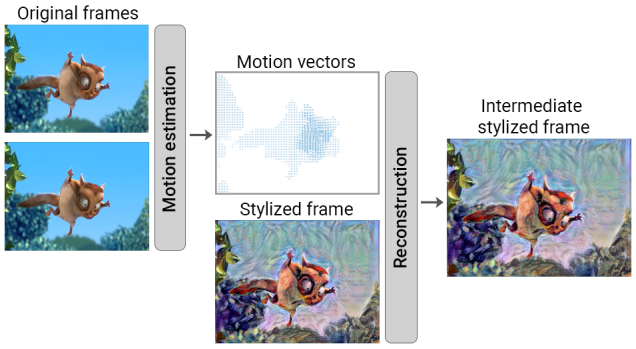

Figure 9: *We use the original frames (left) to estimate motion vectors via optical flow. Only one in three frames is stylized, we use that frame (center, bottom), together with the motion vectors (center, top) to generate the remaining intermediate frames in which stylization was not performed (right).*

5. Conclusion and Future Work

We have presented in this work a content aware style transfer method that introduces an alternative to alleviate one of the major limitations of the state-of-the-art style transfer methods, the content-style trade-off. We have also shown in this work that our method can be easily adapted to video style transfer, allowing to generate stylized videos where most of the details, edges, and content structure are preserved. Our results show that our method provides a reliable alternative for better preserving fine details in stylized images. However, lowering the intensity of the style transfer for some regions can cause that some results will not completely convey the desired style. Future work in this direction could include taking into account perceptual aspects in order to identify the regions or features of the image that maximize the impression of the style. Our video implementation with optical flow yields plausible results, however, more sophisticated techniques could be investigated in the future for obtaining better temporal consistency.

6. Acknowledgements

The authors would like to thank the members of the Graphics & Imaging Lab for fruitful insights and discussion. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (CHAMELEON project, grant agreement No 682080), and from the Spanish Ministry of Economy and Competitiveness (projects TIN2016-78753- P, and TIN2016- 79710-P). Ana Serrano was supported by an FPI grant from the Spanish Ministry of Economy and Competitiveness, and a Nvidia Graduate Fellowship.

References

[CH16] CHEN Y.-L., HSU C.-T.: Towards deep style transfer: A content-aware perspective. In *BMVC* (2016). 2

[CN*10] CHAN S. H., NGUYEN T. Q., ET AL.: Subpixel motion estimation without interpolation. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (2010), IEEE, pp. 722–725. 4

[FSDH16] FRIGO O., SABATER N., DELON J., HELLIER P.: Split and match: Example-based adaptive patch sampling for unsupervised style transfer. 2

[GEB15] GATYS L., ECKER A. S., BETHGE M.: Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems* (2015), pp. 262–270. 2

[GEB16] GATYS L. A., ECKER A. S., BETHGE M.: Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 2414–2423. 1, 2, 3

[GEB*17] GATYS L. A., ECKER A. S., BETHGE M., HERTZMANN A., SHECHTMAN E.: Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 3985–3993. 1, 2

[HB17] HUANG X., BELONGIE S.: Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 1501–1510. 1, 2, 3

[LPSB17] LUAN F., PARIS S., SHECHTMAN E., BALA K.: Deep photo style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 4990–4998. 1, 2, 3

[LTX18] LI M., TU S., XU L.: Computational decomposition of style for controllable and enhanced style transfer. *CoRR abs/1811.08668* (2018). 2

[RDB16] RUDER M., DOSOVITSKIY A., BROX T.: Artistic style transfer for videos. *arXiv preprint arXiv:1604.08610* (2016). 2

[RDB18] RUDER M., DOSOVITSKIY A., BROX T.: Artistic style transfer for videos and spherical images. *International Journal of Computer Vision 126*, 11 (2018), 1199–1219. 2

[SHB14] SONKA M., HLAVAC V., BOYLE R.: *Image processing, analysis, and machine vision*. Cengage Learning, 2014. 4

[SKLO18] SANAKOYEU A., KOTOVENKO D., LANG S., OMMER B.: A style-aware content loss for real-time hd style transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018). 2

[SZ14] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. *CoRR abs/1409.1556* (2014). 2

[ULVL16] ULYANOV D., LEBEDEV V., VEDALDI A., LEMPITSKY V.: Texture networks: Feed-forward synthesis of textures and stylized images. *arXiv preprint arXiv:1603.03417* (2016). 2

[Yin16] YIN R.: Content aware neural style transfer. *arXiv preprint arXiv:1601.04568* (2016). 2