Towards many-to-one neural style transfer method

Michal Bogacz, and Marcin Iwanowski
Warsaw University of Technology
Institute of Control and Industrial Electronics
ul.Koszykowa 75, 00-662 Warszawa, POLAND
m.bogacz1997@gmail.com, iwanowski@ee.pw.edu.pl

***Abstract*—Neural style transfer aims at transforming the artistic style of a painting into a photograph. The classic methods are based on the one-to-one principle, where stylization is performed based on a single style image. In the paper, an approach is proposed that allows transferring the style based on multiple images of the same reference painter multiplied, in addition, by data augmentation. Comparing to the original method, the proposed approach produces visibly better results, and allows for greater flexibility in the transferring process. The proposed many-to-one method is illustrated by an example of transferring the style of two Polish painters using the same content image**

***Index Terms*—style transfer, deep learning, convolutional neural network**

I. Introduction

Neural style transfer is nowadays one of the most impressive fields of computer vision. The style transfer algorithms require two input images: the style image and the content image. The first one is an artwork – w photograph of a painting created by a painter with all properties of hers/his characteristic style: colors, strokes, way of expressing imagination. The second, content image, is a photograph of a real-world scene. Starting from such input images, an algorithm based on the deep neural network transfers the style image to the content one, so the latter looks like painted by an artist, creator of the style image.

The classic approaches, as described above, perform within the frames of *one-to-one* approach: the style of a *single* style image is transferred to a single content one. However, the characteristic style of a particular painter is a set of techniques that are observable on many paintings of the same authorship. Often, to get a comprehensive feeling of the style, one should study the broader scope of the whole artwork. Thus, when discussing the transferring the style of a particular painter, one should consider multiple style images. Therefore the transferring should instead be done using a *many-to-one* process.

In this paper, we investigate the possibility of applying a set of style images instead of a single one. The reference set of images that are used in the neural style transfer process is created employing 1. using multiple paintings of the same painter as style images and 2. multiplying each image using augmentation techniques. The proposed approach is tested using the photographs of paintings of two Polish artists: Edward Dwurnik[1] (Fig. 2) and Tytus Brzozowski[2] (Fig. 1). The principal subject of paintings of both painters was cities. They were painted in wide frames with many details of buildings, objects on streets, etc. Each of both painters has his distinctive style. In the research, we'll use the proposed many-to-one transferring process. To illustrate the method, we'll apply it to stylize the photograph of one of the squares of the city of Warsaw, Poland. The stylization will be performed based on the set of style images shown in Fig. 2 and in Fig. 1. The final results might be compared to real images depicting the same city square painted by both artists that do not belong to the set of style images. One can easily notice that all three photos present the same location in Warsaw. The authors even decided to use similarly framed versions of the location. The content image[3] and both artworks are shown in Fig. 3. The defining features of every painter and artist's style can be divided into two main groups, micro, and macro characteristics. Fine-scale textures, such as little brush strokes and paint shapes, can allow identifying the artist immediately. Besides these small structures of material, the whole arrangements of brush strokes are also important and should be considered when describing the generalized style of the image.

Based on content and style it is possible to define an appropriate loss functions, yet these functions are not unanimous between different styles and contents. Unfortunately it is impossible to define a metrics, which would allow for quantitative measurement of style transfer quality. The results must be shown for each combination of style and content photo, these results must be evaluated individually [1].

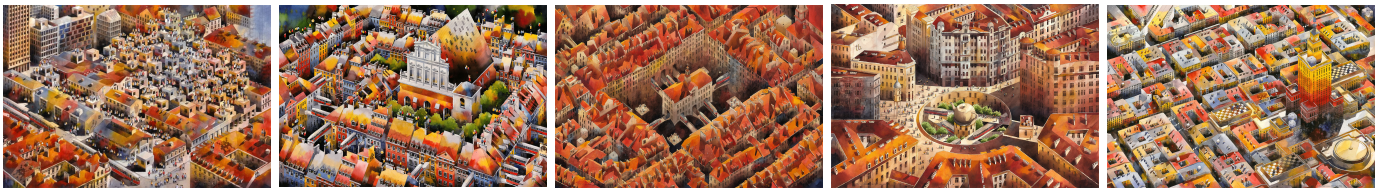
The structure of the paper is as follows. After mentioning the related work in Section 2, we will concentrate on describing the proposed approach. First, the primary image-to-image example transfer is going to be introduced. This approach will be treated as a benchmark method for assessing the changes introduced by the proposed methods. Next, in section 3, the method of augmenting the generalization of basic spatial factors will be introduced as well as the proposal of transferring features from multiple style images at the same time. In section 4, we will try to combine both proposed methods to create the generalized style representation of style image. Finally, section 5 concludes the paper.

[1]Source of paintings' images: https://desamodern.pl/artysci-edward-dwurnik

[2]Source of paintings' images: https://tytusbrzozowski.pl/pl/obrazy/

[3]Source of picture: https://polandonair.com/produkt/plac-zbawiciela/

978-1-6654-4112-4/21/$31.00 © 2021 IEEE



(a) 'Norblin' (b) 'Archikatedra w zieleni' (c) 'Rynek' (d) 'Plac Unii Lubelskiej' (e) 'Plac Defilad'

Fig. 1: Paintings created by Tytus Brzozowski (used as style images).

(a) 'Idealne miasto' (b) 'Poznań 2' (c) 'Pl. Zbawiciela' (d) 'Tarnów' (e) 'Wrocław'

Fig. 2: Paintings created by Edward Dwurnik (used as style images).

(a) (b) (c)

Fig. 3: Images showing same location, but in widely different styles: (a) real photo (content image), (b) 'Plac Zbawiciela' T. Brzozowski, (c) 'Pl. Zbawiciela' E. Dwurnik

II. Related works

When considering the style transferring problem, one can not forget about the precursors of this idea. Image Analogies [2], based on the several techniques of texture synthesis [3], [4] was the first-ever introduced example-based method of transferring not only the texture but also the exact semblance of highly defined style. The technique introduced color preservation and stylization of different regions independently. The following step in Style Transfer development was to use deep interlayer features of CNN networks [5] in patch-based methods instead of immediate actual features used up to this point.

Although the results were getting more precise and reliable with time, the next big leap was introduced with the use of pre-trained CNN classification structures [6], [7] in example-based Style Transfer. Compared to earlier approaches, the style is represented as the statistical sum of different deeply hidden features in CNN structure in this method. The images are not divided into separate patches during the stylization process and are not treated as disconnected objects - the images are stylized as a whole structure. These characteristics allow for a much more flexible optimization process and allow for creating objects that were not present on the original image. Unfortunately, the main advantage of the neural model is, at the same time, its greatest weakness. Due to the use of features that are not understandable to a human, it is not straightforward to control stylization. Some of these problems, like spatial and color matching, were addressed thanks to the control of perceptual losses [8], but many remain unsolved.

Besides the upper-mentioned problem of feature extraction, it is essential to understand that all previously introduced methods are computationally heavy iterative methods. To acquire a stylized image, one has to run a long optimization process in which two images are feed-forward through a pre-trained CNN network and later iteratively modified. The apparent response to this problem was to move the computations to the training phase using GAN, and auto-encoder structures [1], [9]–[11].

III. Training with single style image

The main objective of the texture synthesis with the use of example-based style transfer methods is to create an automated generation process, which takes two images x_s and x_c (resp. style image and content image), the process then creates new x' in which the style of x_s is applied to the image x_c. Both *style* and *content* are defined as abstract terms of image statistics. One can say that images are painted or created in the same style if they embody the correlation of those specific image features. Although we are able to define the process of mathematical optimization in which series of measurement matrices is used to perform the process of transformation, the main criterion for quality evaluation of performed transfer is the subjective human inspection. The synthesized image is considered well transferred if the observer is unable to differentiate between the original texture and the newly created texture. At the same time to approach the problem of enhancing the efficiency and the quality of methods, one must identify the underlying correlations of features between images.

The starting point of our research is a variant of the seminal Neural Style Transfer [6] method. The main advantage over previous methods described earlier is the usage of what is called Deep Image Representations. Stylized images were all created with the use of a pre-trained VGG-19 network [12], which was created and trained to perform image classification. During the process of adaptation, we used some features acquired from convolutional and pooling layers, as it was conducted in the original paper [6], a simplified process of extracting the deeply hidden image features is shown in Fig. 3. Contrary to what the paper said in our individual case, removing maximum pooling operations for average pooling did not yield any positive effects. Based on initial tests, we have decided to stay with the original VGG model. Another change, compared to the original method, was the usage of the original content photo as a starting point of the algorithm.

A. Content representation

When a convolutional neural network is prepared for satisfying the classification or detection problem, it develops a hierarchy of filters to convey information about a described object with growing distinctness. Therefore, to represent the content present on the image, it is possible to use the features produced on chosen convolutional layers directly. Additionally, features extracted from multiple layers are not constrained by pixel to pixel arrangement in the original image.

Based on the definition of features from the last paragraph, each given content image $\vec{x}$ is transformed by consecutive convolutional layers into filter response. Every layer can be described as N_l filters of N_l feature maps, each of size M_l. The size of the feature map can be denoted as the height times width of the feature map. So for the pursue of stylisation the response in layer l can be stored in matrix $F_l \in \mathbb{R}^{N_l \times M_l}$ and activated output of i^{th} filter at j^{th} position in layer l is denoted as F^l_{ij}.

Let $\vec{x}$ and $\vec{p}$ be the generated image and original image and F_l and P_l the upper described features in layer l. The content loss can be described as

$$\mathcal{L}_{content}(\vec{p}, \vec{x}, l) = 1/2 * \sum_{i,j} (F^l_{i,j} - P^l_{i,j})^2. \quad (1)$$

The gradient of this function can be later used with respect to an image being generated. Thus any optimization method can be used to transform it in such a way that it reproduces the same deep features as the original content image.

B. Style representation

Style is a much subtler element of the image or painting than bare content. Deep neural networks have no problem distinguishing between different objects like cats and dogs, even though these objects can be displayed in very different styles. Earlier, we have described the style mostly as paint strokes and brush strokes. Due to this observation, representation of style will be mostly concentrated on texture differences. Generating the difference in texture requires defining a special feature space [13] based on a correlation between different feature responses. Feature similarity can be presented with a Gramm matrix $G_l \in \mathbb{R}^{N_l \times N_l}$, this matrix represents the inner product of flattened feature maps i and j inside a given layer l:

$$G_{l_{i,j}} = \sum_k F^l_{i,k} F^l_{j,k}. \quad (2)$$

Combining correlation of multiple features, allows to capture a deep representation of texture, without taking into consideration their global arrangement. The loss of style between two images $\vec{a}$ - original image and $\vec{x}$ - generated image, where their Gramm matrices are respectively A_l and G_l on l^{th} layer, is

$$E_l = \frac{1}{4N_l^2 M_l^2} \sum_{i,j} (G^l_{i,j} - A^l_{i,j})^2, \quad (3)$$

and total loss on L layers is

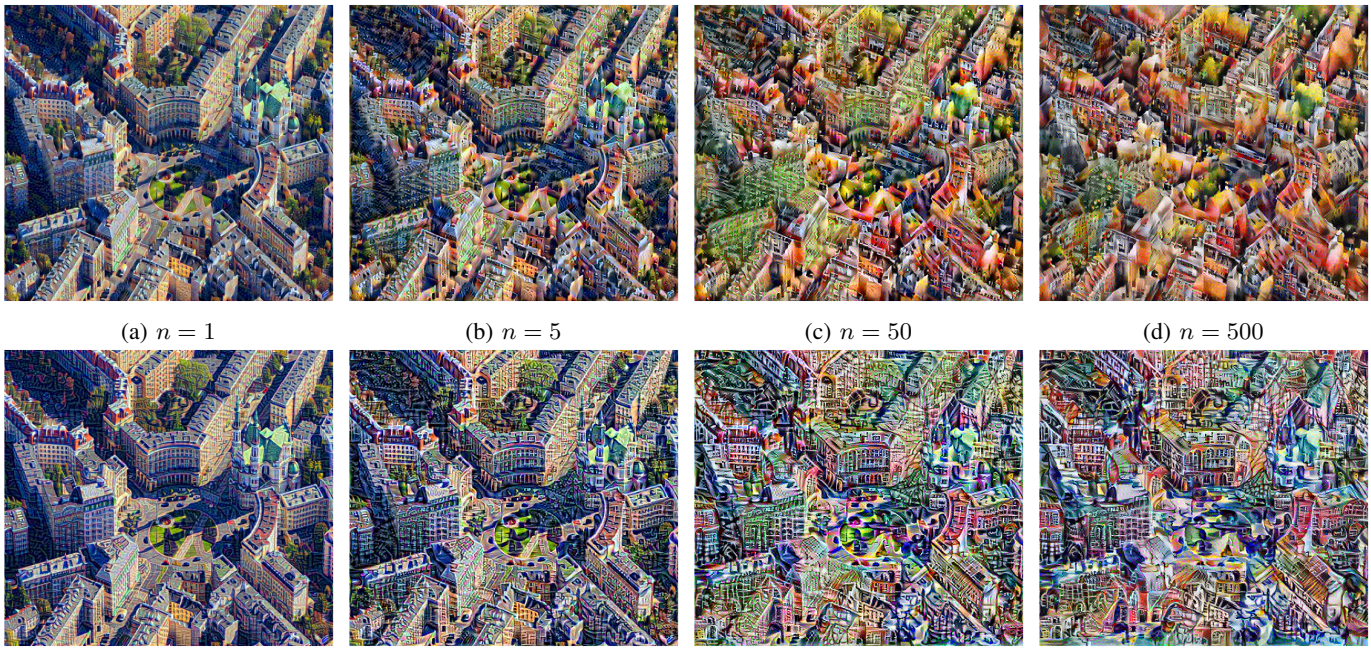
$$\mathcal{L}_{style}(\vec{a}, \vec{x}) = \sum_{l=0}^{L} E_l. \quad (4)$$

Analogically to content loss, the gradient of the style loss function can be used to create an image similar to the original one regarding abstract texture features.

To transfer the style between two images, for simplicity and clarity, the style photo or artwork will be denoted as $\vec{a}$ and the content image or photo will be denoted as $\vec{p}$. The transfer must be performed in such a way that content features will not be lost in the process. At the same time, the newly generated image $\vec{x}$ must be brought as closely as possible to style the image regarding the style texture features. Under such conditions, the loss function can be written as follows

$$\mathcal{L}_{total}(\vec{p}, \vec{a}, \vec{x}) = \alpha \mathcal{L}_{content}(\vec{p}, \vec{x}) + \beta \mathcal{L}_{style}(\vec{a}, \vec{x}). \quad (5)$$

α and β weights are responsible for defining a relationship between the importance of content and style in the optimization process. This implementation of the algorithm



(a) $n = 1$ (b) $n = 5$ (c) $n = 50$ (d) $n = 500$

(e) $n = 1$ (f) $n = 5$ (g) $n = 50$ (h) $n = 500$

Fig. 4: Results of the style transfer performed on n iterations of the algorithm. First row shows the transfer of style from Fig. 1b to Fig. 3a. Second row shows transfer from Fig. 2e to Fig. 3a

differs in two major ways from the original [6]. Instead of using the proposed L-BFGS optimization algorithm, we are using Adam [14]. Additionally, instead of defining the layers' importance factor in the style loss function, we have decided to continue the experiments with a static set of equally factored layers after performing comparative tests. All shown in this paper examples pull style features from 'conv1-4', 'conv2-4', 'conv3-4' and 'conv4-4' layers, choice of these layers is purely subjective. Still, it is important to notice that texture properties are extracted from all levels of convolutional abstraction. Content is pulled from the 'conv5-2' layer. Again the choice is subjective and based on analysis of multiple results. The only necessity was for content features to be pulled from deeper layers of convolutions so that highly abstract features would be compared in the loss function.

C. Baseline results

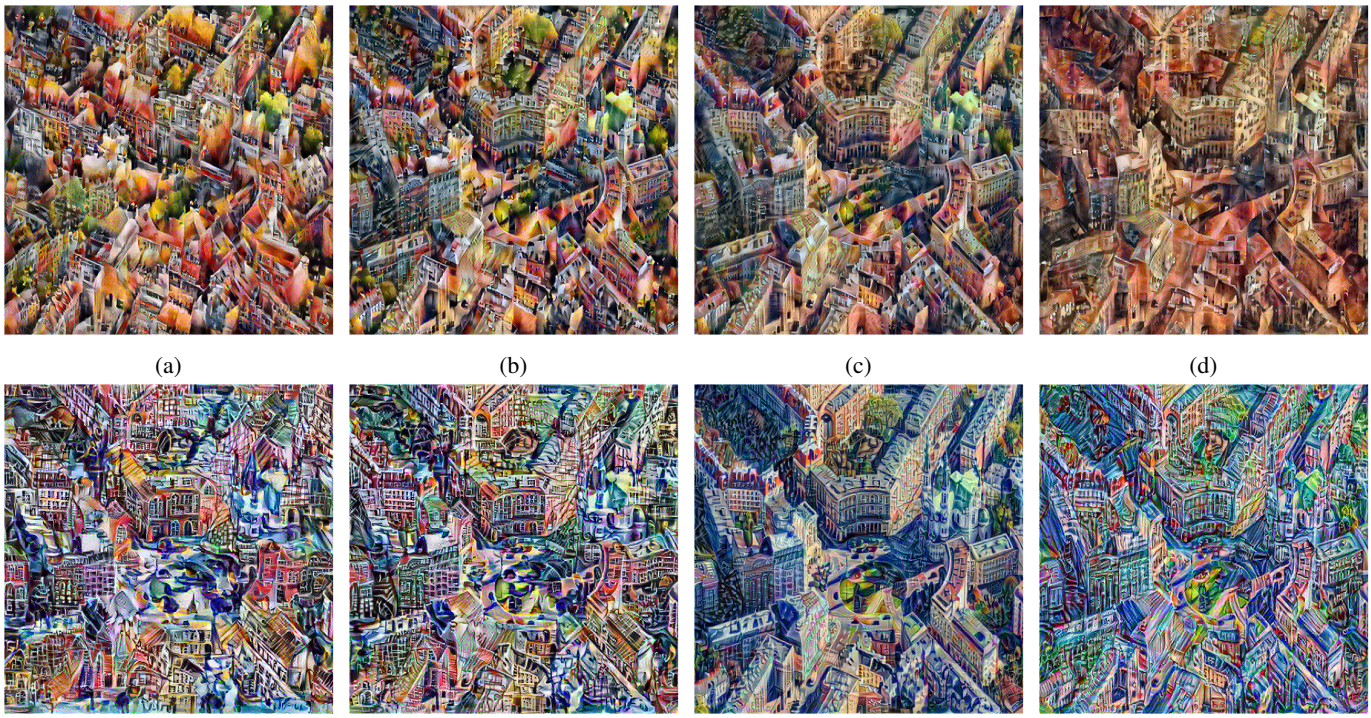
On Fig. 4 several chosen steps from the whole iterative process are shown. We have decided on these particular n values in order to show the whole process of optimisation. Two artists were chosen as the style sources, and the photo of the 'Plac Zbawiciela' will be used consistently as the content photo. This exact configuration was chosen since artists Brzozowski and Dwurnik represent two widely different styles of painting, and the content image contains many details, which can be used to evaluate the quality and precision of style transfer.

One can easily observe that the style of the painting was successfully translated to the content image. During the first ten to twenty iterations, not much happens. Mostly high-frequency components are added to the groups of pixels, at around iteration, forty to fifty transferred style gets recognizable. Later iterations are performed to get rid of noisy areas on the textured image. It is essential to notice that transfer is not perfect. A lot of contours are sharpened and crooked. At the same time, full objects are being transferred from one image to another, e.g., the rainbow, which is not present in the original photo.

IV. Many-to-one approach

The problem of achieving generalized artistic style representation can be approached from many different angles. It is essential to remember that the primary metrics of success are the impression and satisfaction of the viewer. The main goals of the methods presented in this section are to augment the performance of the earlier proposed algorithm and modify the algorithm to achieve a style representation of several artworks. This kind of transfer should be able to generate new paintings, where the style of several artists is blended or where several images painted by one artist are combined to generalize all paintings into one.

Before approaching the problem of multiple image generalization, it is vital to reduce the number of artifacts appearing on the transformed content images, like blocky and crooked buildings, the appearance of objects that were not originally on the content image distortion of details. This step is essential in generalization because all details of style images should



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 5: All images are the best possible chosen stylisation, chosen from iteration $n = 1000$. First row presents stylisations performed on Brzozowski's and second row on Dwurnik's artwork. Images (a) and (e) are single image transformations, images (b) and (f) are single augmented image transformations, in both cases Fig. 1b and Fig. 2e were transfered to Fig. 3a. In order to create images (c) and (g) set of not augmented paintings was used, and in case of (d) and (h) an augmented set of images was used. In this case all four were created with help of respectively Fig. 1 and Fig. 2.

be blended and transformed into context images to allow for mixing between multiple style images.

The methodology chosen to augment the images to increase the input style data variation was to augment the style photo. Style representation is highly identified with color and texture, so the chosen augmentations couldn't change the texture or color of the picture. Introduced augmentations should change the presented structure of objects without changing their composition on the image, so textures pixel to pixel presentation in both micro and macro scale should stay constant, but the whole image should change. Transformations that satisfy all those requirements are rotations and flips. Unfortunately, the use of rotations would require some padding or filling, which would result in a change of style of the image. In this experiment, we decided to use only two types of augmentations: horizontal and vertical.

The idea of transferring generalised artistic style has been already consider and implemented [9] [10] [11] [1], yet all this tries are performed with use of GAN or auto-encoder networks. To perform the generalized stylization process, either a new loss metric is required, one which would incorporate features from many input images, or a change in the process of iterative optimization and generation algorithm. In the proposed method, the second approach was chosen. We have decided that this approach should introduce fewer changes to the overall performance of the method. At the same time, the comparison of the results should be more reliable because no change to the loss metrics is introduced to the method. Introduced change to the algorithm was an iterative change of the style image so that the stylization algorithm is optimizing the output photo on all introduced style examples. This approach can be treated as a heuristic of combined loss values for all style images so that the loss function can be presented as

$$\mathcal{L}_{total}(\vec{p}, \vec{a}, \vec{x}) = \sum_{n} \left(\alpha \mathcal{L}_{content}(\vec{p}, \vec{x}) + \beta \mathcal{L}_{style}(\vec{a_n}, \vec{x})\right), \quad (6)$$

where n is the number of introduced images. In the case of the introduced heuristic, we denote all loss components as in the basic method, except for $\vec{a}$, here it is a vector of n style photos on which the new generalized texture will be based.

In Fig. 5b and Fig. 5f it is visible that when using the generalized metrics, the transfer is not as invasive as in the original method case. Yet, the transferred artists' style is recognizable, and the main elements are consistent with the paintings used in the optimization process. In the transferred Dwurnik manner, a small point of color and brush-like strokes are visible, which is typical of this abstract artist. Generated

Brzozowski style is also imposing. One can see single-colored planes, which are watered down as is typical of the artist's style.

Two previous sections were introducing the methods created to perform better and more generalized stylization of a photo. The first was centered around the better transfer of style from a single image to another. In contrast, the second one aimed to combine multiple images to create the unanimous style of all included paintings. We presume that combining these two methods should yield better results, presenting very individual style representation.

Fig. 5d and Fig. 5h present the results achieved by using the set of augmented photos for two styles. Compared to results shown in Fig. 5c and Fig. 5g, the style is much more prominent, at the same time compared to images from Fig. 5a, Fig. 5b, Fig. 5e and Fig. 5f we managed to reduce number of artefacts without any loss in the style representation. Both Fig. 5d and Fig. 5h greatly convey the most important features of artists style, without obstructing the original objects of content images. When comparing respectively Fig. 5d and Fig. 1b with Fig. 5h and Fig. 2e, we can see that artists vision of 'Plac Zbawiciela' in Warsaw is highly convergent with images produces by the method.

The final results of the proposed approach are shown in Fig. 5. Although the original method has already implemented several precautionary measures that do not use the immediate pixel to pixel relations, when using multiple change versions of the same image, we were able to archive much better transfer. Small elements of the image were visibly less noisy, and many parts of the image which disappeared when using the original method now remained recognizable. Results, as shown, vary based on the artists' style, Impressionists like Dwurnik use small spots of paint to create the illusion of a consistent plane, so the method of generalization is less efficient because there are no big uniform objects on the photos. The second example is contrary. Watercolors create large unanimous planes which create objects. Due to this fact, the generalization method managed to reduce the number of artifacts.

To present and identify the factors used in rating the quality of the transformation, the similarities and differences between these photos shown in Fig. 3 will be used. Arguably most important is the relative position of different buildings on the photos, although the most outstanding one is the church with two towers. On all three photos, it is located in the same position relative to other buildings. The same goes for all elements of architecture. It is very important for the style transfer algorithm to change the style of such elements without changing their relative positions on stylized images. The second most important factor is the continuity of textures, walls, roofs, and roads are continuously represented in one particular fashion, which should not be changed randomly but should remain consistent. This characteristic is also applicable to different styles of spatial regions like sky, ground, water or skin, and clothes.

V. Conclusions

The paper addresses the greatest conceptual issue in the neural style transfer method [6], which is a single image to single image approach presented in the original paper. We believe that we have managed to succeed in more profound style generalization and that our results are visibly better and allow for greater flexibility in the transferring process. The proposed solution allows for significant improvements in stylization by capturing the influence of the general style on the content. Although the results are very satisfying, the base method has an inherited problem: lack of real-time transformation. Based on this property, it would be advantageous if similarly good results would be achieved in conjunction with other style transfer methods.

References

[1] A. Sanakoyeu, D. Kotovenko, S. Lang, and B. Ommer, "A style-aware content loss for real-time HD style transfer," *CoRR*, vol. abs/1807.10201, 2018. [Online]. Available: http://arxiv.org/abs/1807.10201

[2] A. Hertzmann, C. Jacobs, N. Oliver, B. Curless, and D. H. a. Salesin, "Image analogies." Association for Computing Machinery, Inc., August 2001. [Online]. Available: https://www.microsoft.com/en-us/research/publication/image-analogies/

[3] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in *IEEE International Conference on Computer Vision*, Corfu, Greece, September 1999, pp. 1033–1038.

[4] L.-Y. Wei and M. Levoy, "Fast texture synthesis using tree-structured vector quantization," *Computer Graphics (Proceedings of SIGGRAPH'00)*, vol. 34, 05 2000.

[5] A. J. Champandard, "Semantic style transfer and turning two-bit doodles into fine artworks," 2016.

[6] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *CoRR*, vol. abs/1508.06576, 2015. [Online]. Available: http://arxiv.org/abs/1508.06576

[7] ——, "Texture synthesis and the controlled generation of natural stimuli using convolutional neural networks," *CoRR*, vol. abs/1505.07376, 2015. [Online]. Available: http://arxiv.org/abs/1505.07376

[8] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman, "Controlling perceptual factors in neural style transfer," *CoRR*, vol. abs/1611.07865, 2016. [Online]. Available: http://arxiv.org/abs/1611.07865

[9] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images," *CoRR*, vol. abs/1603.03417, 2016. [Online]. Available: http://arxiv.org/abs/1603.03417

[10] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," *CoRR*, vol. abs/1701.02096, 2017. [Online]. Available: http://arxiv.org/abs/1701.02096

[11] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," *CoRR*, vol. abs/1610.07629, 2016. [Online]. Available: http://arxiv.org/abs/1610.07629

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[13] L. A. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," 2015.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.