Revisiting Artistic Style Transfer for Data Augmentation in A Real-Case Scenario

Stefano d'Angelo, Frédéric Precioso, Fabien Gandon

REVISITING ARTISTIC STYLE TRANSFER FOR DATA AUGMENTATION IN A REAL-CASE SCENARIO

Stefano D'Angelo[⋆]*, Frédéric Precioso*[⋆]*, Fabien Gandon*[†]

[⋆]Université Côte d'Azur, CNRS, Inria, I3S
[†]Université Côte d'Azur, Inria, CNRS, I3S

ABSTRACT

A tremendous number of techniques have been proposed to transfer artistic style from one image to another. In particular, techniques exploiting neural representation of data; from Convolutional Neural Networks to Generative Adversarial Networks. However, most of these techniques do not accurately account for the semantic information related to the objects present in both images or require a considerable training set. In this paper, we provide a data augmentation technique that is as faithful as possible to the style of the reference artist, while requiring as few training samples as possible, as artworks containing the same semantics of an artist are usually rare. Hence, this paper aims to improve the state-of-the-art by first applying semantic segmentation on both images to then transfer the style from the painting to a photo while preserving common semantic regions. The method is exemplified on Van Gogh's paintings, shown to be challenging to segment.

***Index Terms*—** Neural Style Transfer, Image Segmentation, Image-to-Image Translation, Artistic Style Transfer, Photorealism

1. INTRODUCTION

Witnessing the impressive results Deep Learning techniques have provided to other sciences, more and more non-AI-experts are considering these methods in their own field. Museum curators are progressively keen on benefiting from the potential of Deep Learning in the tedious task of analyzing artworks, maintaining thesaurus and databases, and combining unstructured content with metadata to bring analysis to another level [1]. In order to support the level of requirement from very precise analysis of artworks by curators, the need for a faithful data augmentation method is immediate. Indeed, if one wants to apply a deep network to analyze artwork semantic content, it is obviously better to get a network better fitted to the data [1], for instance by fine-tuning an existing architecture on the target artworks. A previous work [1] showed that data scarcity for some types of art pieces hindered transfer learning. Standard base CNN models (VGG, ResNet, ...) were not providing an accurate classification/detection of objects and the fine-tuning was not satisfactory either, due to the lack of training data. Thus, without data augmentation, transfer learning and fine-tuning could not provide the expected support to the work of curators - in particular the automation of cultural data quality maintenance.

When Vincent Van Gogh painted *Starry Night*, he saw none other than a beautiful landscape, and he captured his *impression* of the scene on the painting by applying his unique style. Luckily, his style is not completely lost, since Deep Learning has lately enabled one to be able to capture artistic style and to transfer it to a new image [2]. However, most of the current methods rely on (very) large training sets to build a relevant neural model for style transfer. Furthermore, the consistency with the original art style is most often global which leads to many visual artefacts in the generated artwork from a real photo. This article contributes a new method to enable semantic style transfer by first applying semantic segmentation on both images, and then transferring the style from the painting to a photo while preserving common semantic regions.

In Sec. 2, we position our work with respect to the state-of-the-art, while in Sec. 3, we detail our approach. The experimental results are presented in Sec. 3, and we then conclude the paper in Sec. 5.

2. NEURAL STYLE TRANSFER

Neural Style Transfer is an area of application of Image-to-Image translation. The goal is to transfer the style of an image, called *style image*, to another image, called *content image*. This field has been widely explored, and many methods have been proposed [2]. The seminal neural model in the state-of-the-art has been developed by Gatys et al. [3], and consists of: (i) first, providing the *content image* as input of a fixed pre-trained VGG19 network (on ImageNet), then to train another VGG19 network from scratch by providing a white noise image as input and aligning its feature maps with those of the former network; (ii) second, applying almost the same procedure to the *style image*, but aligning the Gram matrices of their feature maps rather than the feature maps themselves. A further improvement was proposed in [4] by combining the CNN architecture with a Markov Random Field,

a regularizer that maintains local patterns of the "style" exemplars. These models still struggle to correctly transfer the style, often resulting in an overlap between content and style images and thus, in many local artefacts.

One branch of style transfer architectures that has made some improvements is the one involving unpaired datasets containing samples from two different domains (the source domain and the target domain). This is a point of advantage over previous architectures, as it is always a tedious task to pair the samples. In this branch, the predominant models are Generative Adversarial Networks (GANs), whose properties have been exploited to generate "fake" images mapping together the two different domains. This is the case of GANilla [5], an architecture consisting of a CNN based on Resnet18 used in an autoencoder structure to learn a latent space and skip connections allowing to translate the input image from the source domain to the target domain. In CycleGAN [6] the network not only translates the input image from the source domain to the target domain but also translates the resulting image back to the source domain, enforcing the visual consistency between the two images in the source domain: the initial image and the image resulting from two consecutive translation steps.

However, the problem with these methods is that they are not semantic-aware. In practice, they make little distinction between objects in an image. Moreover, they do not exploit the specific mapping that exists between images of **paired datasets**.

Several other models have tried to include semantic information when transferring the style. Liao et al., for example, enforced mapping between source and target images at different feature map levels [7]. Their results are convincing when the source and domain contents are already fairly semantically aligned. In a more recent work [8], soft semantic masks of regions that should match between source and target, are extracted. Here again, the existing alignment between semantic regions that should match impacts the quality of the results greatly. One of the most impressive recent achievements in this field is based on advances in deep semantic segmentation which aim to identify the classes in both source and target images. The two two segmentation masks are then semantically aligned, after which they are integrated into the transfer [9]. However, as with the previous models, Park et al.'s model is not yet able to handle paintings in which the shape of the objects is very far from reality. In Van Gogh's paintings, for example, the line is not used to describe reality but has an *expressive* function - transfiguring reality itself.

3. PROPOSED APPROACH

Different to other methods, the approach presented here uses a paired dataset, which consists of pairs of images sharing a similar visual content. Its main strength is based on the fact that it can exploit the one-to-one mapping between images of each pair to guarantee results that are more relevant to the style contained in a specific painting. The starting point was hence the dataset used by Zhu et al. to train CycleGAN [6], where Van Gogh's paintings have been retrieved from WikiArt, while real photos have been downloaded from Flickr by using landscapes-related hashtags. Then, images were manually paired in order to match them as best as possible.

Since all the pre-trained segmentation models have been trained on real pictures, it is hard to directly segment paintings. Hence, a pre-processing step precedes the actual training phase and consists of first converting paintings to real images and then extracting segmentation masks in the real image domain according to the approach coined by Penhouët et al. [10]. The latter uses *image segmentation* and *semantic grouping* to merge minority classes in order for the masks of each pair of images to match. The resulting masks are then simply mapped back onto the painting image to provide a reliable semantic segmentation of both the painting and the target real image.

This pre-processing phase is also the main novelty brought by this paper to the current style transfer literature, since it solves the main issue of semantically segmenting paintings.

Results have been then evaluated according to how similar they look to Van Gogh's artworks, or, in other words, how "fake" they are. All the code is publicly available at **this** GitHub repository, including the one for results evaluation.

3.1. Pre-processing

Before applying style transfer, a pre-processing phase has been studied and applied. Since all the pre-trained models for image segmentation have been trained on real pictures, we cannot expect good results if the segmentation is directly applied to paintings. Hence, a strategy to overcome this issue consists of converting paintings into real photos. For this purpose, a CycleGAN has been trained on two sets: the set of Van Gogh's paintings, consisting of 400 samples (set **A**), and the set of photographs (set **B**), containing 6853 images. The model has been trained with the default parameters used in the original paper ([6]) for a total of 120 epochs. Results and benefits of this pre-processing phase are discussed in Sec. 4.

CycleGAN has been chosen for this task because it ensures **cycle consistency**: when we translate from one domain to the other and back again, we should arrive where we started. Its loss is composed of two terms: the **adversarial loss** and **cycle consistency loss**. The first is used to improve the quality of fake images generated from one domain to the other while the second loss, instead, incentivizes the cycle consistency. For further details on CycleGANs refer to [6].

Once all images have been converted, a subset of 21 Van Gogh's paintings have been selected and paired with real photographs. Then, all the selected images have been segmented using the approach of Penhouët et al. [10], which is composed

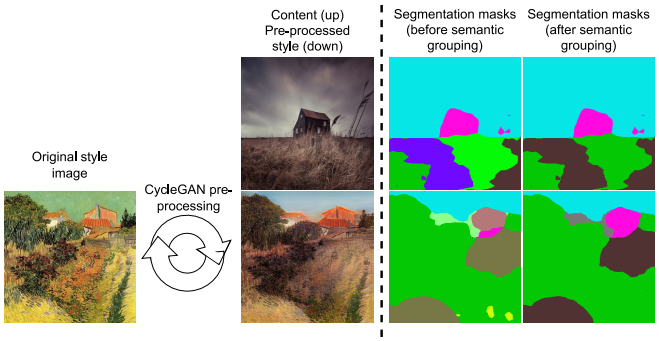

Fig. 1. Visual summary of the pre-processing phase.

of two parts: first, a pre-trained CNN called *Pyramid Scene Parsing Network (PSPNet)* creates a segmentation image; secondarily, a Knowledge Graph from a Python library, called *Sematch*, measures the similarity between two class words (e.g.: *sky* and *ground*). The purpose of this second part is to semantically group similar classes into a wider class, so that both the content and the style image share the same classes. Semantic grouping is regulated by a parameter $\theta \in [0, 1]$ called *semantic threshold*, which groups similar classes together, thereby allowing both the content and style masks to have the same number of classes. When $\theta = 1$ no grouping is applied, while with $\theta = 0$ all the classes are merged into a single class. Here, $\theta = 0.6$ as is in the original paper ([10]), which shows how quality of segmentation masks varies when θ deviates from this value. Fig. 1 illustrates what the pre-processing phase looks like for a single pair of images.
Note that in this paper, when segmentation masks are mentioned, we are referring to the semantically grouped masks.

3.2. Style Transfer

The architecture presented here is based on both Penhouët et al.'s *Automated Deep Photo Style Transfer* (ADPST) model [10], and the classical *Neural Style Transfer* (NST) model [3]. The former was introduced in Sub-Sec. 3.1, while the latter in Sec. 2. In brief, NST extracts features from a VGG19 for both style and content images, then jointly minimizes their losses $\mathcal{L}_c$ and $\mathcal{L}_s$:

$$\mathcal{L}_{total}(\vec{p}, \vec{a}, \vec{x}) = \alpha \mathcal{L}_c(\vec{p}, \vec{x}) + \beta \mathcal{L}_s(\vec{a}, \vec{x}), \quad (1)$$

where α and β are the weighting factors for content and style reconstruction, while $\vec{p}$, $\vec{a}$, and $\vec{x}$ are the photograph, the artwork, and the generated image, respectively [3].

ADPST model is based on *Deep Photo Style Transfer* ([11]), with the difference that in ADPST segmentation masks are created automatically. The objective is to minimize the following loss:

$$\mathcal{L} = \sum_{l=1}^{L} \alpha_l \mathcal{L}_c^l + \Gamma \sum_{l=1}^{L} \beta_l \mathcal{L}_s^l + \lambda \mathcal{L}_m + \eta \mathcal{L}_a, \quad (2)$$

where $\mathcal{L}_c$, $\mathcal{L}_s$, $\mathcal{L}_m$, $\mathcal{L}_a$ are the *content* loss of NST (Eq. 1), the *augmented style* loss, the *affine* loss, and the *image assessment* loss, respectively, which are all regulated by different parameters.

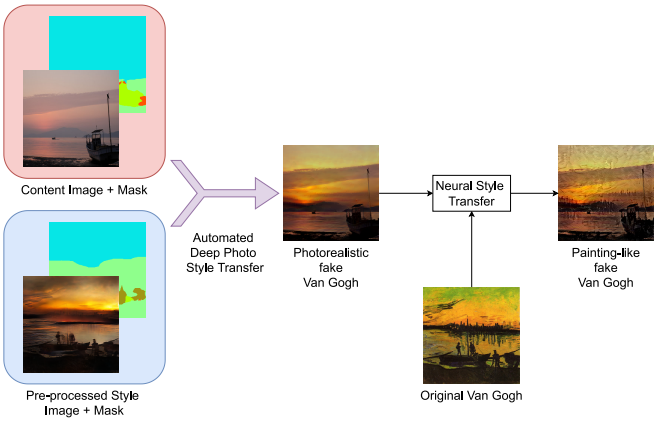

Fig. 2. Schema of the architecture.

The loss is minimized for a certain number of iterations, which, according to Penhouët et al., should be at least 1000. The authors also observed that good results are achieved with about 2000 iterations and improvements after 4000 iterations are most of the time negligible; therefore, the number of iterations was here set to 3000.
As mentioned above, the workflow presented in this Section combines both ADPST and NST models, and is shown in Fig. 2. Each pair of content and pre-processed style images is fed into the ADPST architecture, which transfers the style in the photorealistic domain. In this way, we are able to exploit the segmentation masks of both the images, mainly transferring the palette and the semantic content of the painting.
However, we still need to map the resulting image to the paintings domain. To accomplish this task, we resorted to the NST architecture due to its loss function which is the foundation of ADPST's loss, as evident from Equations (1) and (2). Hence, the final painting-like fake Van Gogh is the result of a style transfer from the original (not pre-processed) painting to the photorealistic fake Van Gogh obtained from ADPST.

4. EXPERIMENTAL RESULTS

Three baselines were chosen to compare the outcomes of the architecture presented in this paper with the state-of-the-art: (i) the classical NST model [3], (ii) CNNMRF architecture [4] due to it being similar to NST, but rather focusing on different patches of images, and (iii) CycleGAN [6], which involves the usage of an unpaired dataset.
In Fig. 3 the results of our model are, at first glance, those that balance content and style better. A closer look shows that style is more correctly transferred between two semantically similar areas of the images. Concerning the three baselines,

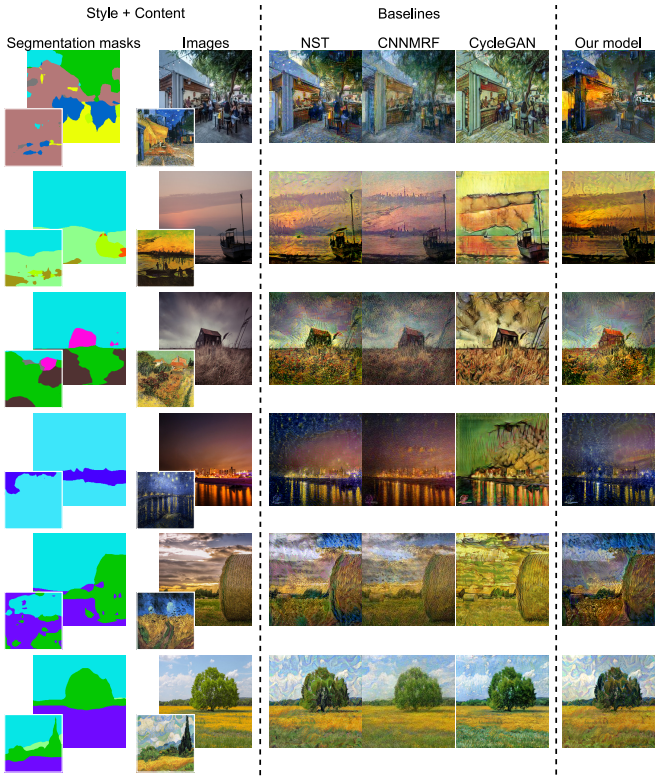

Fig. 3. Some of the results compared to three state-of-the-art models.

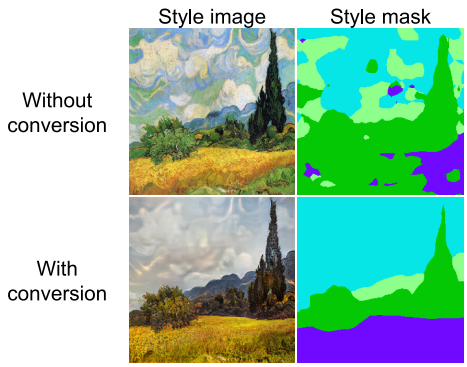

Fig. 4. Benefits of painting to photo conversion.

with a visual assessment we can easily observe that results of CNNMRF are too close to the content image, in the sense that the style is in general poorly transferred. CycleGAN generates the most differing images compared to all the others, because it takes elements from all photos in the dataset. Indeed, due to the nature of GANs, the style is not captured by a precise correlation between content and style image, but is rather dispersed throughout each image. Results of NST are, instead, more visually appealing than in the other models, which is especially impressive because it is the oldest model. Still, the images it generates are too simplistic. In fact, since the objective of this model is to jointly minimize only the content and style losses, the final image tends to be the result of a trade-off between content and style.

A numerical evaluation of the results further confirmed the effectiveness of our work. Indeed, we fine-tuned the classifier of a pre-trained Resnet18 to predict whether a painting was real or generated, and our model yielded the highest error, thus fooling the CNN more than the other three architectures.

As previously mentioned, the novelty this paper wants to bring to current literature is mostly inherent to the pre-processing phase, where the conversion of paintings to real photographs allows one to have a better mapping of semantic information between each pair of images. The intermediate results to be interpreted are derived from the pre-processing phase itself. The advantages of this are easily seen in the quality of the semantic masks generated by the style images, which depends on both the quality of the photorealistic conversion and the semantic grouping applied. In Fig. 4, the style mask extracted from the photorealistic painting is clearly closer to the actual content of the painting itself. Indeed, in the mask generated directly from the original artwork the sky is segmented in multiple instances. Furthermore, the ground is confused with the greenery, as we can see from the mixture of purple and green in the style mask. This is not the case when the mask is instead generated from the converted style image, whose result is more pertinent to the content of the painting.

5. DISCUSSION AND CONCLUSION

In this paper, we presented a new way of generating faux-realistic paintings of an artist, moving towards a precise augmentation of data. This could be used to pre-train a more faithful CNN and then use it to support the work of museum curators in the classification and analysis of artworks.

The simple novel idea we propose here is to first convert the style image into a real image such that a more precise semantic segmentation can be extracted. Mapping the resulting semantic regions back onto the style image is then straightforward. Thanks to this precise semantic segmentation of the style image, we can better map semantics between the painting and real photo, leading to a highly improved style transfer.

To improve the results, a future work might focus on improving the quality of semantic masks. Indeed, the segmentation in ADPST is optimized for "Scene Parsing", obtaining the best results for landscape paintings. Therefore, a first possible solution is to include photos containing common objects in our dataset. By doing so, CycleGAN will be aware of how to pre-process those kinds of images, and the results may be refined. We could also consider tuning the parameters of the ADPST model. A final option would be to use our model as a generator in a GAN architecture, fine-tuning the generation of images according to the performance of the discriminator.

References

[1] A. Bobasheva, F. Gandon, and F. Precioso, "Learning and Reasoning for Cultural Metadata Quality," *Journal on Computing and Cultural Heritage (ACM JOCCH)*, 2022. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03363442

[2] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, "Neural style transfer: A review," *IEEE Transactions on Visualization & Computer Graphics*, vol. 26, no. 11, pp. 3365–3385, nov 2020.

[3] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.

[4] C. Li and M. Wand, "Combining Markov Random Fields and Convolutional Neural Networks for Image Synthesis," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 2479–2486.

[5] S. Hicsonmez, N. Samet, E. Akbas, and P. Duygulu, "Ganilla: Generative adversarial networks for image to illustration translation," *Image and Vision Computing*, vol. 95, p. 103886, 2020.

[6] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[7] J. Liao, Y. Yao, L. Yuan, G. Hua, and S. B. Kang, "Visual attribute transfer through deep image analogy," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–15, 2017.

[8] H.-H. Zhao, P. L. Rosin, Y.-K. Lai, and Y.-N. Wang, "Automatic semantic style transfer using deep convolutional neural networks and soft masks," *The Visual Computer*, vol. 36, no. 7, pp. 1307–1324, 2020.

[9] J. H. Park *et al.*, "Semantic-aware neural style transfer," *Image and Vision Computing*, vol. 87, pp. 13–23, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885619300435

[10] S. Penhouët *et al.*, "Automated deep photo style transfer," *CoRR*, vol. abs/1901.03915, 2019. [Online]. Available: http://arxiv.org/abs/1901.03915

[11] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep Photo Style Transfer," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, Jul. 2017, pp. 6997–7005.