Content and Style Disentanglement for Artistic Style Transfer

Dmytro Kotovenko   Artsiom Sanakoyeu   Sabine Lang   Björn Ommer
Heidelberg Collaboratory for Image Processing, IWR, Heidelberg University

Abstract

Artists rarely paint in a single style throughout their career. More often they change styles or develop variations of it. In addition, artworks in different styles and even within one style depict real content differently: while Picasso's Blue Period displays a vase in a blueish tone but as a whole, his Cubist works deconstruct the object. To produce artistically convincing stylizations, style transfer models must be able to reflect these changes and variations. Recently many works have aimed to improve the style transfer task, but neglected to address the described observations. We present a novel approach which captures particularities of style and the variations within and separates style and content. This is achieved by introducing two novel losses: a fixpoint triplet style loss to learn subtle variations within one style or between different styles and a disentanglement loss to ensure that the stylization is not conditioned on the real input photo. In addition the paper proposes various evaluation methods to measure the importance of both losses on the validity, quality and variability of final stylizations. We provide qualitative results to demonstrate the performance of our approach.

1. Introduction

Style transfer models synthesize a real image in the style of a given artwork. To achieve a convincing stylization, models must preserve the content of the real image and closely resemble the chosen artistic style. This raises the following questions: "what does it mean to maintain the content" and "what characteristics define style".

Artworks show different renderings of content: While some styles disregard content, such as Jackson Pollock's Abstract Expressionism or Wassily Kandinsky's highly abstract style, others display content but alter it in a specific manner. The modern paintings of Marc Chagall or Henri Rousseau transform reality into staged almost dream like scenes. These observations lead to the conclusion that a more in-depth study of the relation between artistic style and content is required to obtain a better image stylization. However, there is no tool which measures the degree to which an artist altered content. Indeed this would require original content photos which display the exact content an artist has painted in a specific artwork. Let us assume the converse scenario: imagine we do have a set of photos and a simple way to approximate

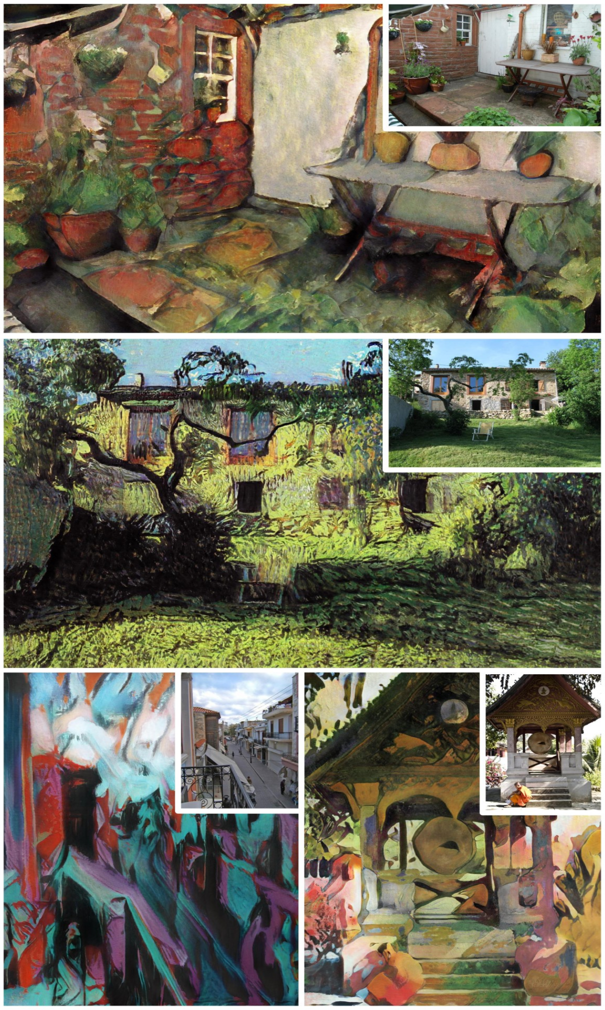

Figure 1. Stylization examples generated by our network. Paul Cezanne (top), Vincent van Gogh (middle) and Paul Gauguin and Ernst Ludwig Kirchner (bottom). Full sized images can be found in the supplementary material and on our project page[†].

[†] compvis.github.io/content-style-disentangled-ST/

the artist's stylization. Then we are given a collection of content-stylization-pairs, which can be used to solve the content alteration problem described above (and if we disregard the fact that the stylization is still only approximated and not yet optimized). Thus, if we stylize the same content photo in two different styles, the results should reflect the differences in style while displaying the same content. On the contrary if we use the same style but different content images, we should obtain stylizations in the same style but with different content. The latter constraint warrants an independence of style from content. We formulate this objective as a fixpoint disentanglement loss.

Recently there has been a great interest in the task of style transfer; existing works produce stylized images by extracting style features from a single artwork [7, 13, 28, 18, 10, 4, 31] or a collection of images [24, 33]. Although these approaches reproduce a given style, they lack sensitivity to subtle variations in style and a comprehensive understanding of style. Instead of learning all possible variations of a style, previous models only learn visual clues most dominant in the style and ignore the rest of the style manifold. However, artists rarely maintain a single style throughout their career, but more often change styles or develop variations of it. While still working in an Impressionist style, Monet's later works display a more loose and expressive brushstroke compared to earlier paintings due to declining health. To capture these small variations in style, we need a framework able to simulate this. We thus propose a novel method which learns the particular style of an artist as a single entity and adjusts the stylization to the particular example of style by introducing style similarities and discrepancies within a single style. This is achieved by stylizing the same content with two similar style samples and forcing stylizations which display identical content to still lie apart in the style space. We address this objective by introducing a fixpoint triplet style loss.

We propose the first approach which extracts style from a group of examples of the same overall style but with subtle variations therein, while still providing fine control over the stylization. We make the following contributions: (i) we propose two novel losses, namely a fixpoint disentanglement loss and a fixpoint triplet style loss to allow for a finer stylization of images and a better coverage of style distributions. (ii) Moreover we provide an approach to disentangle style and content of an artwork resulting in artistically compelling stylizations and a better content preservation as shown in the experiments section. (iii) Our model also provides a smooth style space and thus allows to interpolate within one style and across different styles. We also produce smooth video stylizations with our method; examples can be found on our project page.

2. Related Work

Style transfer Style transfer methods generate new images in the style of a specific artist by rendering an input content image utilizing style information extracted from an image of a real artwork. Gatys et al. [7] first proposed a neural style transfer to encode the style of an image using the pairwise correlation matrix between feature activations of a pretrained Convolutional Neural Network (CNN). Given a single content image and a single reference style image the stylization is then produced by an iterative optimization procedure which matches the style representation of a content image to a style image. Selim et al. [26] further extended the neural style transfer method [7] and applied it to portraits of faces. To enable faster stylization, other research works used neural networks [13, 10, 18, 30, 17] which approximated the slow iterative algorithm of [7]. To model multiple artistic styles within a single model Dumoulin et al. [4] proposed a conditional instance normalization, which enables to synthesize interpolations between different styles. [8, 12] introduced additional control over the results of stylization by altering color, scale and stroke size. [16] introduced a content transformation module between the encoder and decoder to achieve a content-and-style-aware stylization. They used similar content in photos and style to further learn an object-specific stylization.

Most of existing style transfer approaches extract style representations from a single artwork [7, 13, 28, 18, 10, 4, 17, 31] and treat each artwork as an independent style example. To the best of our knowledge, only [24, 33] learn style from a collection of related style examples. However, they cannot model multiple styles simultaneously, lack flexibility and do not have control over the stylization process. In contrast, our approach utilizes the rich information which is given in a group of very similar style samples taken from an image collection of one style, combines multiple styles in the same network and provides a more fine-grained control over the stylization process.

Latent space in generative models Learning an interpretable latent space representation has been a prevalent focus of deep learning research, especially in the field of generative models[3, 21, 1]. In recent years conditional image synthesis received much attention [11, 21]. Other research presents more theoretical approaches such as [20, 3] or state-of-the-art approaches, which show good results for image synthesis of natural images [2] and human faces[14, 15] but require immense computation power. Recently a lot of works have focused on the disentanglement of object shape and appearance [6, 19, 5].

3. Approach

Our initial task can be described as follows: given a collection of artworks $(y, s) \sim \mathbf{Y}$, where y is an art image

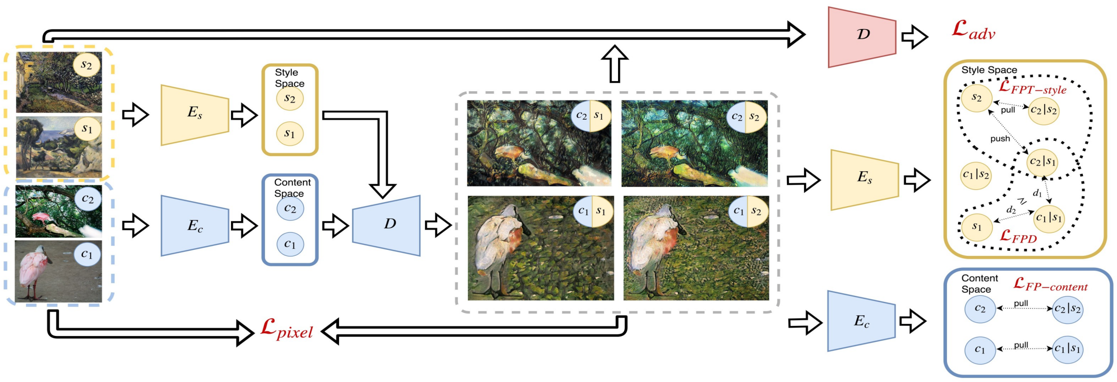

Figure 2. The training iteration is performed on a pair of content images with content representations c_1, c_2 and a pair of style images having style representations s_1, s_2. In a next step, image pairs are fed into the content encoder E_c and style encoder E_s respectively. Now we generate all possible pairs of content and style representations using the decoder D. The resulting images are fed into the style encoder E_s one more time to compute the $\mathcal{L}_{FPT-style}$ on two triplets that share $c_2|s_1$ by comparing the style representations of generated images with the styles $c_1|s_1, c_2|s_1, c_1|s_2, c_2|s_2$ to the styles s_1, s_2 of the input style images. The resulting images are given to the discriminator $\mathcal{D}$ to compute a conditional adversarial loss $\mathcal{L}_{adv}$ and to E_c to compute the discrepancy $\mathcal{L}_{FP-content}$ between the stylizations $c_2|s_2, c_1|c_1$ and the original c_1, c_2. Both depicted encoders E_s are shared as well as both encoders E_c.

and s is a style class label, and a collection of photos $x \sim \mathbf{X}$, we want to learn a transformation $\mathcal{G} : X \longrightarrow Y$. To measure how well the mapping $\mathcal{G}$ approximates the distribution $\mathbf{Y}$, we introduce a discriminator $\mathcal{D}$ whose task it is to distinguish between a real sample $y \in \mathbf{Y}$ and a generated sample $\mathcal{G}(x)$ for $x \in \mathbf{X}$. In our framework, this task is equivalent to learning an arbitrary mapping from the photo domain to the domain of artworks.

On its own this approach induces no constraints on the original content preservation and therefore can make the original content of the photo unrecognizable. To prevent this we force the generated image to be similar to the stylized image in the pixel domain, namely by minimizing the L_2 distance $\|\mathcal{G}(x) - x\|_2$.

As stated before, we want our image to be conditioned on the query style image $\mathcal{G}(x|y)$ allowing for finer style control. This requires a conditioning of the generated image on the input style image y. We propose to condition the output using the style encoder E_s by extracting style $E_s(y)$ from the style image y and then condition the generative network on this style vector.

Works on unsupervised and supervised domain translation[33, 11, 22] have shown that the task of image-to-image translation can be solved by exploiting the encoder-decoder architectures. We define our generator as a combination of three networks: content encoder E_c, decoder D and style encoder E_s. The former two are fully convolutional feed-forward neural networks responsible for the task of image generation, while the latter network infers a style vector $E_s(y)$ from the image y. The conditioning of the generator network is performed by substituting the offset and scale parameters of the instance normalization layers [29] of the decoder D. The decision which losses should be minimized depends on our defined goals. First, we aim to generate artistically convincing stylizations by preserving the style class of the given painting. Hence, we formulate the conditional adversarial loss as follows:

$$\begin{aligned}\mathcal{L}_{adv} := \underset{(y,s)\sim\mathbf{Y}}{\mathbb{E}}&[log(\mathcal{D}(y|s))]+\\ \underset{\substack{x\sim\mathbf{X}\\(y,s)\sim\mathbf{Y}}}{\mathbb{E}}&[log\left(1-\mathcal{D}\left(D(E_c(x),E_s(y))|s\right)\right)]\end{aligned} \tag{1}$$

Second, the stylization obtained from a style image (y, s) and the input content image x should resemble the input content image x. Thus we enforce a reconstruction loss between the input content image x and stylization result:

$$\mathcal{L}_{pixel} := \underset{\substack{x\sim\mathbf{X}\\(y,s)\sim\mathbf{Y}}}{\mathbb{E}} [\|(D(E_c(x), E_s(y))) - x\|_2^2]. \tag{2}$$

We do not, however, aim for a simple pixel-level similarity to the input content photo. Indeed such a loss is adversely to the style transfer task, because many artists tend to alter color and shape severely, thus a pixel-level loss might obstruct the stylization task. Considering this, we let the content encoder E_c to determine which features are relevant for content preservation and which can be neglected. This is achieved by using a fixpoint content loss:

$$\mathcal{L}_{FP-content} := \underset{\substack{x\sim\mathbf{X}\\(y,s)\sim\mathbf{Y}}}{\mathbb{E}} [\|E_c((D(E_c(x), E_s(y)))) - E_c(x)\|_2^2]. \tag{3}$$

Although these losses are sufficient to obtain convincing stylizations for one particular artist, they are not suitable to train a model capable to incorporate stylizations for multiple artists within a single network. Our ablations in Tab.2 show that these losses do not support the model to be susceptible to subtle style changes in the query style images, even if examples were taken from the same style.

Another issue is that if the model is only trained with these three losses, it inadvertently conditions the stylization on the input content. To overcome this, we introduce two additional losses which are novel to the task of style transfer: a fixpoint triplet style loss and fixpoint disentanglement loss.

3.1. Fixpoint Triplet Loss

If the objective is a weighted combination of the three losses 1, 2 and 3 defined above, we immediately observe that the style encoder E_s is only driven by the conditional adversarial loss $\mathcal{L}_{adv}$. This loss is minimized by learning to partition the domain of values of (E_s) into distinctive regions. Thus, we are not able to force the encoder to learn a smooth space of a style representation displaying continuous transitions between different styles and pronounced transitions within a single style. To alleviate this we introduce a fixpoint triplet loss:

$$\mathcal{L}_{FP-style} := \mathop{\mathbb{E}}_{\substack{x \sim \mathbf{X} \\ (y,s) \sim \mathbf{Y}}} [\|E_s(s) - E_s(D(E_c(x), E_s(s)))\|^2], \tag{4}$$

which is similar to $\mathcal{L}_{FP-content}$ defined in 3. The loss forces the network to preserve input style. However it shows a similar behavior as described above when visually very different examples of the same style $(y_1, s),\ (y_2, s) \sim \mathbf{Y}$ are mapped onto the same point, namely $E_s(y_1) \equiv E_s(y_2)$; resulting in identical stylizations $D(E_c(x), E_s(y_1)) \equiv D(E_c(x), E_s(y_2))$.

This reasoning could be formalized as follows: first, we want the stylization to be similar to the input style example in the style space. Secondly, a stylization obtained by a different style must also be distant in the style representation space. This resembles a triplet loss widely used in metric learning [25, 9]. In our case, for the style examples $(y_1, s_1), (y_2, s_2) \sim \mathbf{Y}$ and content photo $x \sim \mathbf{X}$, the anchor is the encoded style $E_s(y_1)$, the positive sample is $E_s(D(E_c(x), E_s(y_1)))$, the negative $E_s(D(E_c(x), E_s(y_2)))$ respectively. For a margin r we define a fixpoint triplet loss for a style:

$$\begin{aligned} \mathcal{L}_{FPT-style} := \mathop{\mathbb{E}}_{\substack{x \sim \mathbf{X} \\ (y_1,s_1),(y_2,s_2) \sim \mathbf{Y}}} &\max\Big(0, \\ r + \|E_s(y_1) - E_s(D(E_c(x), E_s(y_1)))\|^2 &- \\ \|E_s(y_1) - E_s(D(E_c(x), E_s(y_2)))\|^2\Big)&. \end{aligned} \tag{5}$$

3.2. Disentanglement Loss

The content within an image can be indicative for the style. For instance particular pieces of clothing may hint at the time and style of the painting. Thus, the content and style are entangled. The generated stylizations are also conditionally dependent on the content target of the photo and not only on the style target. To separate both characteristics, it is necessary to make the target style independent from the target content. This could be achieved by minimizing the following loss:

$$\mathbb{E}[\|E_s(D(E_c(x_1), E_s(y))) - E_s(D(E_c(x_2), E_s(y)))\|^2]. \tag{6}$$

However, this loss is too strict and obstructs a successful training of the model. Therefore we soften the constraint: instead of minimizing it we simply bind it from the top using the fixpoint style loss $\mathcal{L}_{FP-style}$. This loss is minimized by decreasing the $\mathcal{L}_{FPT-style}$ loss. Hence, we also minimize 6. In summary: for an input style sample $(y, s) \sim \mathbf{Y}$ and two random photos $x_1, x_2 \sim \mathbf{X}$ we define the fixpoint disentanglement loss L_{FPD}:

$$\begin{aligned} \mathcal{L}_{FPD} = \mathop{\mathbb{E}}_{\substack{x_1, x_2 \sim \mathbf{X} \\ (y,s) \sim \mathbf{Y}}} &\max\Big(0, \\ \|E_s(D(E_c(x_1), E_s(y))) - E_s(D(E_c(x_2), E_s(y)))\|^2 &- \\ \|E_s(D(E_c(x_1), E_s(y))) - E_s(y)\|^2\Big)&. \end{aligned} \tag{7}$$

The $\mathcal{L}_{FPD}$ penalizes the model for perturbations which are too large in the style representation: if given the style vector $s = E_s(y)$, then the style discrepancy of two stylizations is larger than the discrepancy between stylization and original style.

The main difference to the fixpoint triplet loss is that the latter prevents different stylizations from collapsing into the same style, while the fixpoint disentanglement loss alleviates the influence of the content image on the resulting stylization.

3.3. Training and Model Architecture

We summarize all aforementioned losses given the loss weights $\lambda_{adv},\ \lambda_{pixel},\ \lambda_{FP-content},\ \lambda_{FPT-style},\ \lambda_{FPD}$ to generate the compound loss $\mathcal{L}^*$. We use it as the final objective for the discriminator-generator minimax game: $\min_{\mathcal{G}} \max_{\mathcal{D}} \mathcal{L}^*$. The detailed model architecture and training step descriptions are provided in the supplementary material.

4. Experiments

4.1. Stylization Assessment

The quality of stylized images and thus the representation of artistic style can be measured in several ways. We assess the performance with four different methods, in 1:

Expert preference rate We first stylize various photos in the style of one artist using different methods listed in Tab.1. In a second step we cut out patches of the same size from all stylized images and create a batch. We then show different patches to experts from art history and have them select the patch which best represents the style of the respective artist.

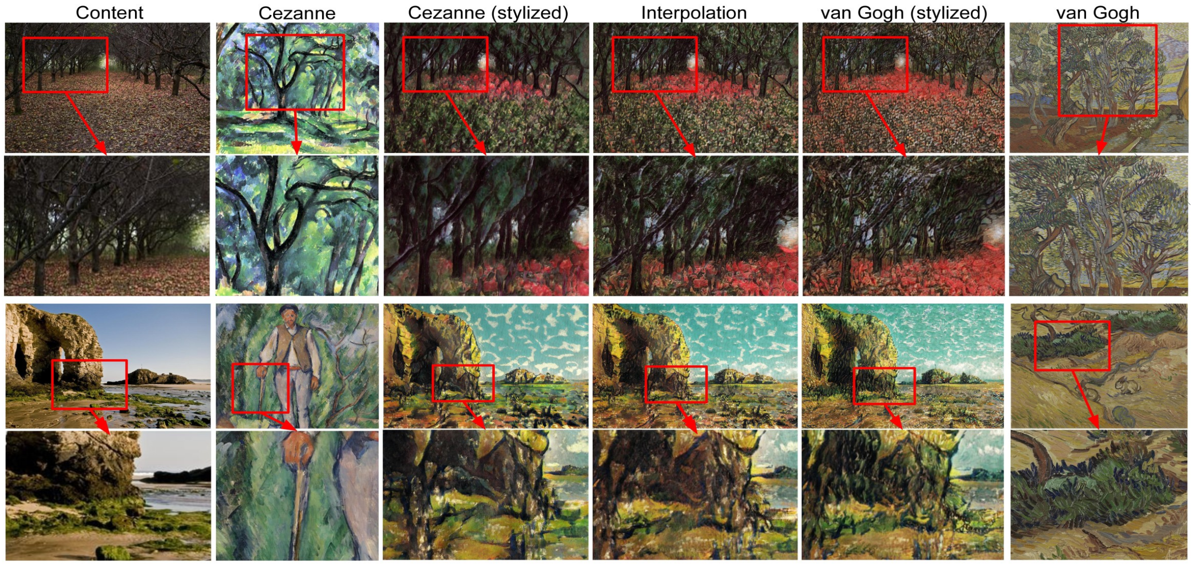

Figure 3. Interpolation between given style samples of Paul Cezanne (column 2) and Vincent van Gogh (column 6). Magnified regions show that our method mimics not only colors but also contours and textures specific to the style. Video interpolations are provided on our project page.

We then measure how frequently each approach is preferred.

Expert deception rate. An identical approach as described in the previous experiment is taken for expert deception rate. Again we show a group of patches, which have been cropped from the stylized images, to art historians. However this time we add a patch from a real artwork by an artist; we compute the number of times art historians identify a patch to be from a real artwork instead of a stylized image.

Non-expert deception rate. The same evaluation as described for the expert deception rate is performed with non-experts that have no prior training in art history.

Deception rate. This approach for assessing the quality of a stylized image was introduced by [24]: a stylized image is presented to a network, which was trained on artist classification. Given a stylized image, the deception rate is the frequency that a pretrained network predicts the artist used for stylization correctly.

The experiment is performed on patches and not full-size images for the following reason: content images are photos from the Places365 dataset[32]. Almost every image contains details which unambiguously refer to our time, i.e. a car, a train, sneakers or a cell phone. Thus humans can easily identify images as not being authentic paintings when spotting these objects. By cropping out patches from the stylized images we significantly mitigate this effect.

We run all experiments for ten different artists and summarize the averaged results in Tab.1. From the table we can conclude that our model significantly outperforms the state-of-the-art AST model [24]. Also note that the art history expert deception rate is higher than the non-expert deception rate since the latter group partially consists of computer vision students - therefore they were better at spotting artifacts typical for generative models. The supplementary material provides additional details on the evaluations.

Table 1. Measuring on image patches how compelling a stylization is (higher is better). The preference rate measures how often art historians prefer a particular stylization technique over others. Deception rates indicate how often stylized patches deceive the viewer, experts and non-experts respectively. Scores are averaged over 10 different styles. A Wikiart-test gives accuracy on real artworks from the test set.

Method	Deception rate	Non-Expert deception rate	Expert deception rate	Expert preference rate
CycleGan	0.130	0.025	0.032	0.012
WCT [18]	0.023	0.035	0.002	0.011
AdaIn	0.061	0.032	0.022	0.021
Johnson et al.	0.080	0.016	0.003	0.014
PatchBased	0.063	0.135	0.010	0.030
Gatys et al.	0.251	0.094	0.069	0.148
AST [24]	0.450	0.050	0.122	0.329
Ours	**0.562**	**0.182**	**0.240**	**0.486**
Wikiart test	0.626	0.497	0.599	-
Photos	0.003	-	-	-

4.2. Disentanglement of Style and Content

We introduce the fixpoint disentanglement loss to disentangle style and content. In order to measure the entanglement, we propose the following two experiments.

Style discrepancy. Our model is able to preserve fine style details independent of changes in the content target photo. To validate this we first measure the average style variation in real artworks for a selected style, which is represented by a collection of artworks $\mathcal{S}$. For measuring, we take a pretrained network used for artist classification $\widetilde{E}_s$ [24] and

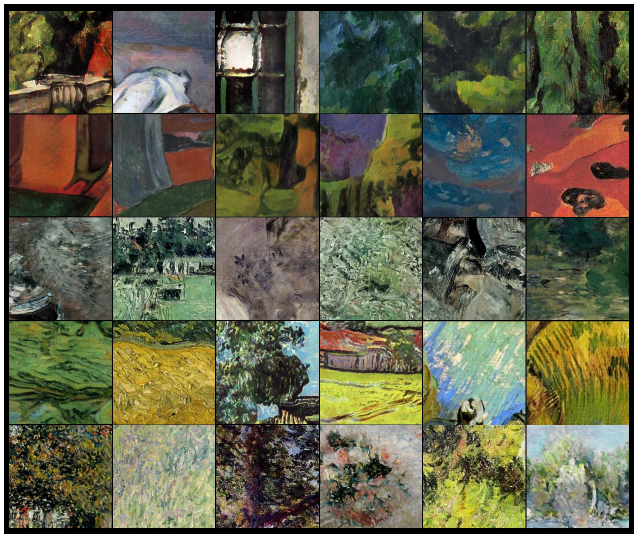

Figure 4. Which patch is taken from a real painting and which from a stylized image? Each row contains a few real patches. Styles (top to bottom): Cezanne, Gauguin, Morisot, van Gogh, Monet. The solution can be found on the last page.

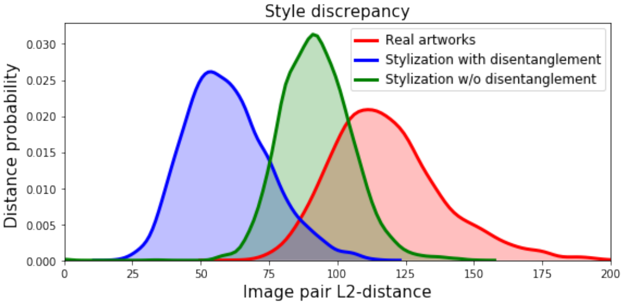

Figure 5. Impact of content on style: we take two different content images, stylize them using the same style example and measure euclidean distance of $\widetilde{E}_s^{fc}(\cdot)$ activations (see Sec.4.2). The experiment is repeated for all content pairs and different style examples to obtain a style distance distribution. This experiment is performed for a model with (blue) and without (green) disentanglement loss. As a reference we compute the distribution between activations of *different* artworks (red).

extract activations of the first fully connected layer from a real artwork $s \in \mathcal{S}$, denoted by $\widetilde{E}_s^{fc}(s)$. Eventually this allows us to approximate a distribution of style variation for a style $\mathcal{S}$ with

$$\{\|\widetilde{E}_s^{fc}(s_1) - \widetilde{E}_s^{fc}(s_2)\|_2 \mid s_1, s_2 \in \mathcal{S}\}. \tag{8}$$

Then we measure the variation in style for our stylized image given distinctive input photos x_1, x_2 and a fixed style sample s :

$$\{\|\widetilde{E}_s^{fc}(D(E_c(x_1), E_s(s))) - \widetilde{E}_s^{fc}(D(E_c(x_2), E_s(s)))\|_2 \mid \\ s \in \mathcal{S},\ x_1, x_2 \in \mathbf{X}\}. \tag{9}$$

In a final step we compute the same distribution 9 but for a model trained without the disentanglement loss.

The three distributions are summarized in Fig. 5 by visualizing their probability density function depicted in red, blue and green respectively. The plot indicates that the model with disentanglement loss produces stylizations which represent a selected style better than the model where the loss is missing. In addition we observe that different content affects stylization results for one style less than style variations within a collection of paintings by an artist. show that

Content discrepancy. In the second experiment we establish how much a change of style influences content preservation. The content similarity is formulated as a L_2 distance in the feature space of the first fully connected layer denoted by $\widetilde{E}_c^{fc}(\cdot)$ of the VGG16 network[27]; the network is pretrained on the ImageNet dataset[23].

First we require a baseline distribution representing subtle changes in content. Therefore we measure the content similarity between nearest neighbors in the $\widetilde{E}_c^{fc}(\cdot)$ space and plot the distribution. Let $\mathcal{C}_i$ denote a dataset of ImageNet photos of class i. Then the baseline distribution of content similarity between nearest neighbors from the ImageNet set in the $\widetilde{E}_c^{fc}(\cdot)$ space is:

$$\{\|\widetilde{E}_c^{fc}(x) - \widetilde{E}_c^{fc}(NN(x))\|_2 \mid x \in \mathcal{C}_i \forall i\} \tag{10}$$

where $NN(x)$ denotes the nearest neighbor of sample x in the $\widetilde{E}_c^{fc}(\cdot)$ space among all the ImageNet samples of the same class.

We now evaluate a change of content in images stylized with different art styles. For two style datasets $\mathcal{S}_1, \mathcal{S}_2$ we estimate the distribution:

$$\{\|\widetilde{E}_c^{fc}(D(E_c(x_1), E_s(s))) - \widetilde{E}_c^{fc}(D(E_c(x_2), E_s(s)))\|_2 \mid \\ s_1 \in \mathcal{S}_1, s_2 \in \mathcal{S}_2\ x \in \mathcal{C}_i \forall i\}. \tag{11}$$

We estimate an identical distribution as defined above for images stylized by a model without the fixpoint disentanglement loss. The probability density functions for all three distributions are plotted in Fig. 7.

This experiment indicates that a change of stylized images introduces less perturbation to the content than distance to the nearest neighbor in the same class.

Qualitative experiments. We provide qualitative results of our approach in Fig.3, 4 and 6. Fig.3 shows that our model captures subtle variations between two styles. In addition our approach learns finer artistic properties (i.e. variations in brushwork) (see Fig.4), reduces the number of artifacts and artificial structures and disentangles style and content (see Fig.6).

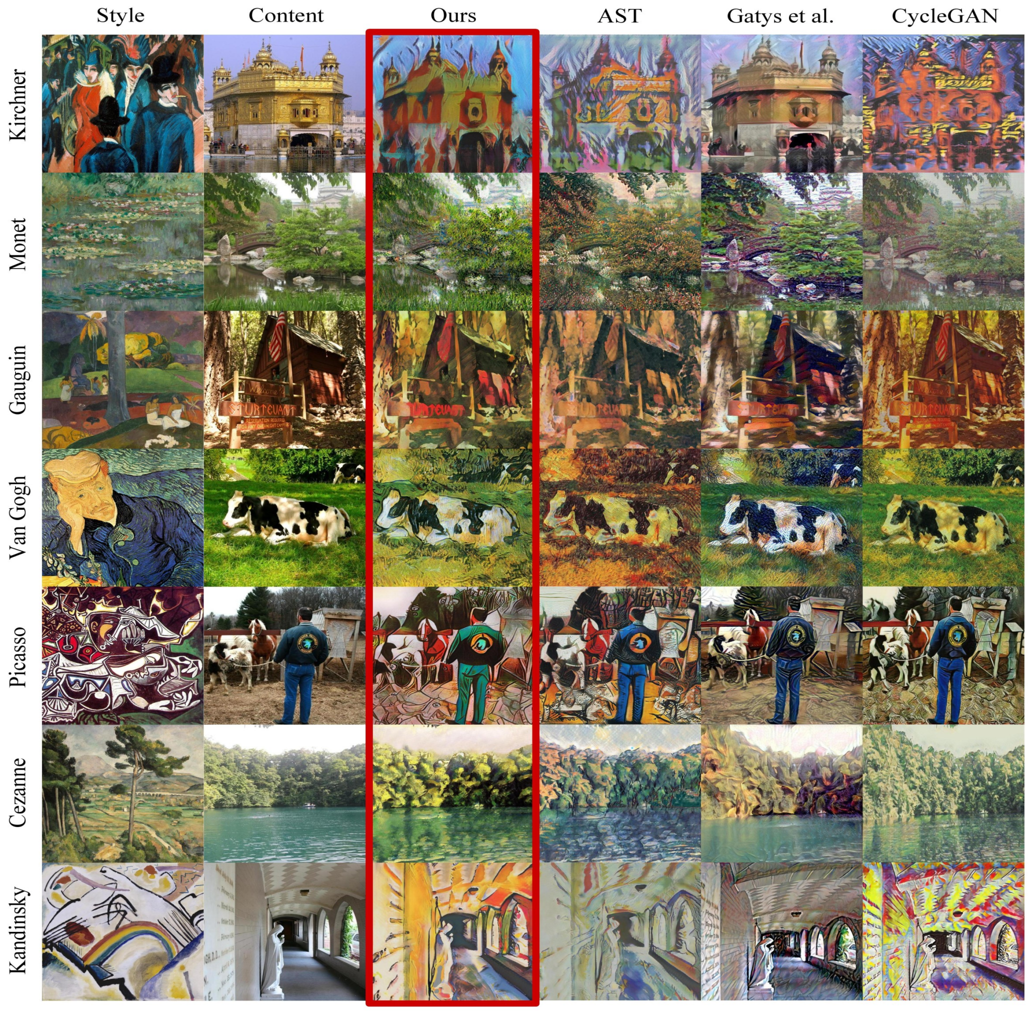

Figure 6. Stylized results from different models (from left to right): ours (in red), AST, Gatys et al. and CycleGAN. We provide style and content images in row one and two to allow a qualitative judgment of stylizations. The figure highlights improvements in quality for images obtained by our model. Images show less artificial structures (as seen in images of Picasso or Kirchner), contain no artifacts in homogeneous regions (see Cezanne) and most importantly highlight the successful disentanglement of style and content. This can be seen in the stylized example of Monet. In comparison the AST model produced ’flowers’, which are common in similar landscape paintings of the artist but not present in the content image. Results are best seen on screen and zoomed in. Full sized images are provided in the supplementary.

4.3. Distribution Divergence

Next we study how well our method covers the variability of the style distribution it aims to replicate. We compute the Kullback-Leibler divergence D_{KL} between the true style distribution and the style distribution of images we have stylized to measure how well our model represents the distribution of style it aims to represent.

We use a network $\widetilde{E}_s$ trained to classify the style of paintings to obtain the style distribution as approximated by the activations of the first fully connected layer, namely $\widetilde{E}_s^{fc}$. The true style distribution $\mathbb{P}_s^{art}$ is approximated by the $\widetilde{E}_s^{fc}$ activations on real artworks. Next we extract activations $\widetilde{E}_s^{fc}$ of the stylized images to approximate $\mathbb{P}_s^{stylized}$ and compute the divergence between the style distribution of real artworks and the style distribution of the stylized images $D_{KL}(\mathbb{P}_s^{stylized} \parallel \mathbb{P}_s^{art})$. We repeat this process for a model trained without the fixpoint triplet style loss $\mathcal{L}_{FPT-style}$ (4) to compute $D_{KL}(\mathbb{P}_s^{no\ \mathcal{L}_{FPT-style}} \parallel \mathbb{P}_s^{art})$. Tab.2 summarizes the style divergences.

Now we visualize style distributions for different stylization approaches. First we fix two artists and train one model with $\mathcal{L}_{FPT-style}$ loss and one without.Then we stylize an identical set of content images using both models and compute the activations of the network $\widetilde{E}_s^{fc}$[24]. As a reference we compute the distribution of style for the real artworks for the two selected artists. Next, we run a PCA on these activations and visualize projections on the first

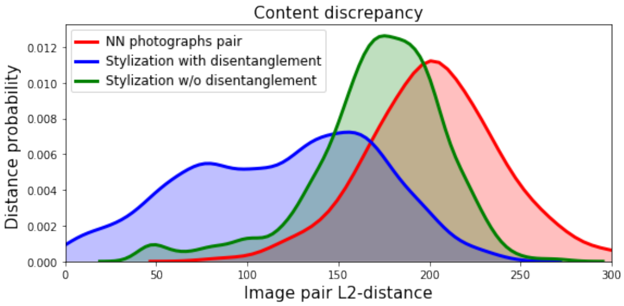

Figure 7. Disentanglement of content across different styles. The same content image is stylized using two different style examples; the difference in content of two stylzations is computed as L_2 norm between activations of the first FC-layer of the VGG-16 network[27]. All distances are accumulated and presented as a distribution(blue). The experiment is performed for the model with(blue) and without(green) disentanglement. Additionally, we compute the content distance from a photo to its nearest neighbor as a reference(red). See Sec.4.2 for further explanation.

Table 2. The deception rate indicates how close an obtained stylization is to the target style (higher is better). Classification accuracy shows how much content of an input photo is left after stylization (higher is better). Style divergence shows a divergence between the style distribution obtained by stylization and true style distribution (lower is better).

Method	Deception rate	Classification accuracy	True style divergence
AST [24]	0.45	0.09	1.12
Ours w/o $\mathcal{L}_{FPT-style}$	0.45	0.16	1.14
Ours w/o $\mathcal{L}_{FPD}$	0.52	0.08	0.32
Ours	0.562	0.17	0.21

principal component as a probability density function (see Fig.8). We observe that the model utilizing $\mathcal{L}_{FPT-style}$ can better match the target distribution of real artworks then the model without this loss.

4.4. Ablations

To summarize the influence of the proposed losses on the final model we use three metrics: deception rate, style divergence and classification accuracy. The latter corresponds to the classification accuracy of the VGG-16 network on the ImageNet stylized images.
We take the AST [24] model as a baseline, because it is trained to extract style from a collection of images. The ablation results are summarized in Tab. 2. They indicate that the $\mathcal{L}_{FPT-style}$ is crucial to incorporate style in its entirety. The $\mathcal{L}_{FPD}$ on the other hand is mostly responsible for a better content preservation but also improves the performance of the stylization task.

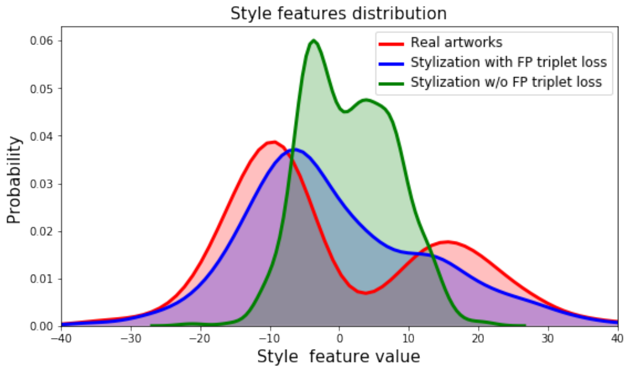

Figure 8. Projection of style features on the first principal component of the PCA decomposition. Style features are computed for real artworks (red), stylized images (blue) and images stylized by a model without the fixpoint triplet style loss (green). The stylized examples and artworks are taken from two artists only, hence the bimodal distribution. Evidently the model utilizing a fixpoint triplet style loss can approximate the distribution of style features of real artworks better.

5. Conclusion

Although previous works concentrated on improving the stylization task, they lack a formal investigation of the questions *How much variation do we find within one style or between different styles?* and *What is the relation between style and content?* – both are relevant to understand style. This paper presents a novel style transfer approach, which is able to capture subtle variations of style while also being able to distinguish different styles and disentangle content and style. We achieve the former by introducing a fixpoint triplet loss to the trained network. We further demonstrated that the introduction of a disentanglement loss makes stylization independent to changes in content. We studied the *influence* of content and style on final stylizations by measuring the preservation of content and representation of style in stylized images. Our approach offers control over the stylization process and enables art historians to study, for example, stylistic developments of an artist in detail.

Acknowledgements

This work has been funded by the German Research Foundation (DFG) - 421703927 and a hardware donation from NVIDIA Corporation.

Solution to Fig. 4:
Cezanne: fake, real, fake, real, real, fake
Gauguin: fake, real, fake, fake, real, real
Morisot: fake, fake, real, fake, fake, real
van Gogh: fake, real, fake, fake, real, real
Monet: fake, real, real, fake, real, fake

References

[1] Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. In *ICML*, 2018. 2

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *CoRR*, abs/1809.11096, 2018. 2

[3] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016. 2

[4] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *Proc. of ICLR*, 2017. 2

[5] Patrick Esser, Johannes Haux, and Björn Ommer. Unsupervised robust disentangling of latent characteristics for image synthesis. In *Proceedings of the Intl. Conf. on Computer Vision (ICCV)*, 2019. 2

[6] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2018. 2

[7] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 2414–2423. IEEE, 2016. 2

[8] Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[9] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017. 4

[10] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 2

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017. 2, 3

[12] Yongcheng Jing, Yang Liu, Yezhou Yang, Zunlei Feng, Yizhou Yu, Dacheng Tao, and Mingli Song. Stroke controllable fast style transfer with adaptive receptive fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 238–254, 2018. 2

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 2

[14] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017. 2

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018. 2

[16] Dmytro Kotovenko, Artsiom Sanakoyeu, Pingchuan Ma, Sabine Lang, and Bjorn Ommer. A content transformation block for image style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10032–10041, 2019. 2

[17] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016. 2

[18] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems*, pages 385–395, 2017. 2, 5

[19] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Björn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Oral + Best paper finalist: top 45 / 5160 submissions)*, 2019. 2

[20] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. 2

[21] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, 2017. 2

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 6

[24] Artsiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Björn Ommer. A style-aware content loss for real-time hd style transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2, 5, 7, 8

[25] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. 4

[26] Ahmed Selim, Mohamed Elgharib, and Linda Doyle. Painting style transfer for head portraits using convolutional neural networks. *ACM Transactions on Graphics (ToG)*, 35(4):129, 2016. 2

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6, 8

[28] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, pages 1349–1357, 2016. 2

[29] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 3

[30] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proc. CVPR*, 2017. 2

[31] Hongmin Xu, Qiang Li, Wenbo Zhang, and Wen Zheng. Styleremix: An interpretable representation for neural image style transfer. *arXiv preprint arXiv:1902.10425*, 2019. 2

[32] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014. 5

[33] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2017. 2, 3