Computational Visual Media
https://doi.org/10.1007/s41095-0xx-xxxx-x

Research/Review Article

Style Transfer Dataset And Insights for Future Stylization Improvements

Victor Kitov[1,2](✉), Valentin Abramov[1], and Mikhail Akhtyrchenko[1]

Abstract We present a new dataset with the goal of advancing the state-of-the-art in image style transfer. The dataset covers various content and style images of different size and contains 10.000 stylizations (made by ArtFlow method) manually rated on a scale of one to ten by three annotators. The dataset allows to train models predicting quality of stylization, recommend best style for given content as well as its optimal size. After labeling 10.000 stylizations we provide our insights, which factors are mostly responsible for creating high-quality stylizations, that can direct and boost future progress in image style transfer.

Keywords dataset, benchmark, stylization quality, aestethic factors, artistic impact factors

1 Introduction

Style transfer is an exciting research area where given a content image (e.g. a family photo) and a style image (e.g. a painting of a famous artist) the task is to redraw the content in the style of the style image, resulting in target stylization, as shown on Fig. 1.

Research articles in style transfer concentrate on introducing new methods, the advantages of which are proven in custom user surveys, where users are asked to rate or to select the best stylizations from a number of variants. The set of content and style images is specific for each article which makes the experiments carried out irreproducible.

Another important aspect that is hidden in such studies is the size of content and style images. Since all stylization methods are based on convolutions with fixed kernels, image sizes is an important factor affecting the final result, as mentioned in [4] and shown on Fig.2.

Finally, different research groups have to reimplement software, rating different stylizations, from scratch.

To overcome these limitations, we made a public style transfer dataset with permissive license, containing content, style and stylization images together with their ratings by three independent respondents. To our knowledge this is the first dataset in this field. The dataset contains 10.000 stylizations and 30.000 ratings on the range from 1 (lowest quality) to 10 (highest quality). Stylizations were made with style images of various sizes and style image were recolored to colors of content to omit the bias, favoring particular style colors.

Contents cover diverse photos with people, animals, buildings, other common objects such as trains, cars, airplanes, trees and flowers. Styles were taken from github repositories of major style transfer methods that work well with them.

The stylizations are made by one of the latest methods [1] whose results do not differ much from more recent attention-based methods such as Stytr2 and AdaAttN [3].

Our dataset can be used to train models, solving the following tasks:

- Given content, style and style size, predict the quality of resulting stylization.
- Given content, list styles that are the most suited for it.
- Given content and style, predict best style size for stylization of highest quality.

After labeling 10.000 stylizations we obtained extensive experience in common successes and failures of style transfer. We summarize them in the paper and give various ideas for future research in this field. Since dataset is public, everyone can validate them and form his own set of recipes for improving style transfer.

1 Lomonosov Moscow State University, Moscow, Russia. E-mail: v.v.kitov@yandex.ru.

2 Plekhanov Russian University of Economics, Moscow, Russia.

Manuscript received: 2024-08-30; accepted: 2024-08-30

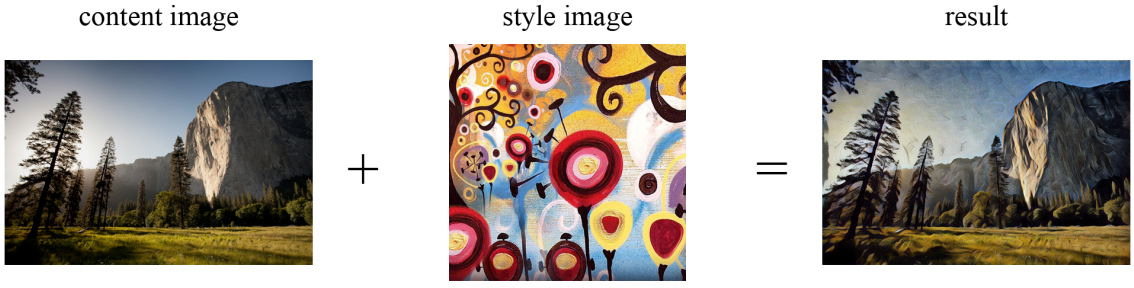

Fig. 1 Style transfer task: redraw content image in the style of the style image. Substitute with 3 images from dataset with high quality result

Fig. 2 Style transfer results using different scale of the style image. Substitute content and 2 examples of stylizations with style image in large and low scale

Overall the contribution of our work is the following:

- We provide a fixed diverse set of content and style images, that can be used as a standardized set for future user studies in style transfer.
- Public stylizations and their ratings are also made public, which can be used as reference baseline in future user studies.
- Our dataset can be used for training models, predicting ratings for different content, style and style size combinations.
- We summarize common pitfalls of style transfer and form a concise set of recipes, what makes style transfer successful, forming directions for future style transfer improvement.

The paper is structured as follows. Section 2 describes data collection and preparation process. Section 3 provides general analysis of obtained ratings. In section 4 we discuss and summarize the experience gained as a result of multiple stylizations and give our insights which factors contribute to good and bad stylizations

and form a concise set of recipes for future style transfer improvement. Section 5 concludes.

2 Dataset Creation Process

2.1 Content and Style Images

Most content images were downloaded from unsplash.com with permissive licence[①]. Since unsucessful style transfer degrades the quality of photographs, large photographs of people were generated using a generative adversarial network [5] (contents 14 and 17) and a diffusion model [6] (contents 8,13,22,26,48) and upscaled [②] without losing photorealism. Content images were selected to cover a broad range of common everyday photographs of people, landscapes, flowers, animals, buildings, rooms, cars, etc. They were also selected to cover panoramic and macro photography, day, evening and night shooting.

Style images were taken from github repositiories of style transfer metthods mainly ArtFlow[③], since these images were selected to suit especially well style transfer and can be freely used for research purposes.

All content images were rescaled, keeping aspect ratio, to contain 900^2 pixels.

2.2 Accounting for Style Size

Since style size strongly effects stylization result, as shown on Fig.2, style images were rescaled (keeping aspect ratio) to contain $150^2, 300^2, 500^2$ and 700^2 pixels.

2.3 Stylization Creation

After obtaining 50 content images, 50 style images of 4 different scales, 10.000 stylizations were generated using ArtFlow [1] algorithm using official implementation[④], which is relatively new, well-recognized in research community. It produces high-quality results on average, works with contents and styles of different sizes with arbitrary aspect ratio and economical in terms of computing resources (can be trained and is able to produce HD results on GPU even with 11GB memory, we used an official pretrained model). Implementations of more recent style transfer algorithms, namely AdaAttN [2][⑤] and Stytr2 [3][⑥] are also able to produce high quality results, however

(1) are much more memory hungry (11GB memory in GPU is not enough to produce HD results),
(2) work only when content and style have equal size (we mentioned that in case of style transfer style size is important and has a major effect on stylization),
(3) require content and style image to have square size (actual image sizes are rectangular).

Because of these limitations, we stick to ArtFlow method on our analysis. Besides, the comparative analysis of stylizations obtained by ArtFlow, AdaAttN and Stytr2 show qualitatively similar results as can be seen in [3]. We didn't consider generative adversarial nets and diffusion models, capable of performing style transfer, since they are much more demanding in terms of computational resources and we were interested in the specific field of style transfer models, which are easy to train and lightweight.

2.4 Style Recoloring

Different styles have different color distributons. Standard style transfer transfers not only drawing patterns (like brush strokes), but color distribution as well. Since this may lead to inconsistent results (e.g. redrawing a bright photo with dark colors) and to color bias in rating (some colors are more preferrable by respondents), before stylization style image was recolored to the colors of the content image, using the following algorithm:

(1) Convert content and style image from RGB to LAB color scheme.
(2) Calculate means $\mu_l^s, \mu_a^s, \mu_b^s$ for LAB color channels of the style image.
(3) Calculate means $\mu_l^c, \mu_a^c, \mu_b^c$ for LAB color channels of the content image.
(4) Calculate standard deviations $\sigma_l^s, \sigma_a^s, \sigma_b^s$ for LAB color channels of the style image.
(5) Calculate standard deviations $\sigma_l^c, \sigma_a^c, \sigma_b^c$ for LAB color channels of the content image.
(6) Rescale LAB color channels l^s, a^s, b^s of the style image to bring color distribution closer to the content image:

$$l^s := \frac{\sigma_l^c}{\sigma_l^s}(l^s - \mu_l^s) + \mu_l^c$$
$$a^s := \frac{\sigma_a^c}{\sigma_a^s}(a^s - \mu_a^s) + \mu_a^c$$
$$b^s := \frac{\sigma_b^c}{\sigma_b^s}(b^s - \mu_b^s) + \mu_b^c$$

① https://unsplash.com/license
② Upscaling with https://www.pixelcut.ai/image-upscaler
③ https://github.com/pkuanjie/ArtFlow/tree/main/data/style
④ https://github.com/pkuanjie/ArtFlow
⑤ https://github.com/Huage001/AdaAttN
⑥ https://github.com/diyiiyiii/StyTR-2

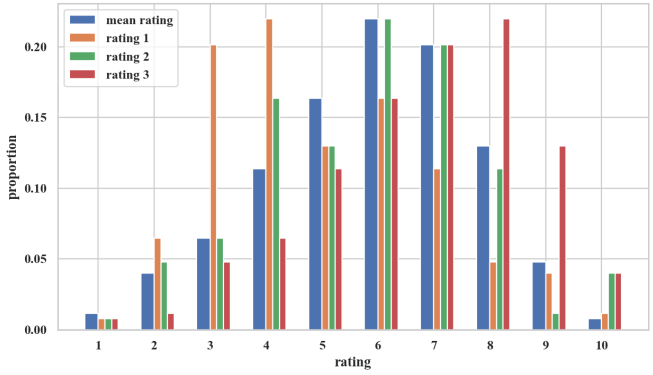

Fig. 3 Rating distribution

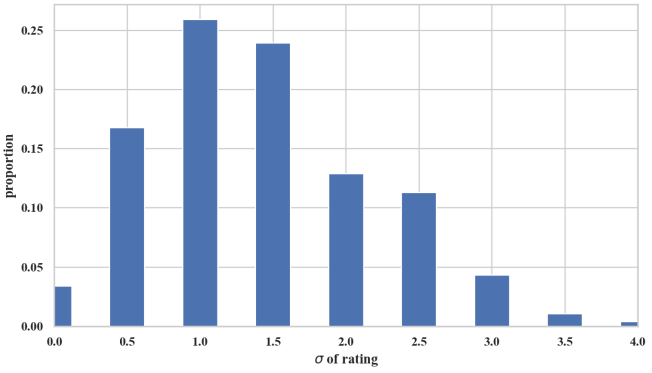

Fig. 4 Rating standard deviation distribution

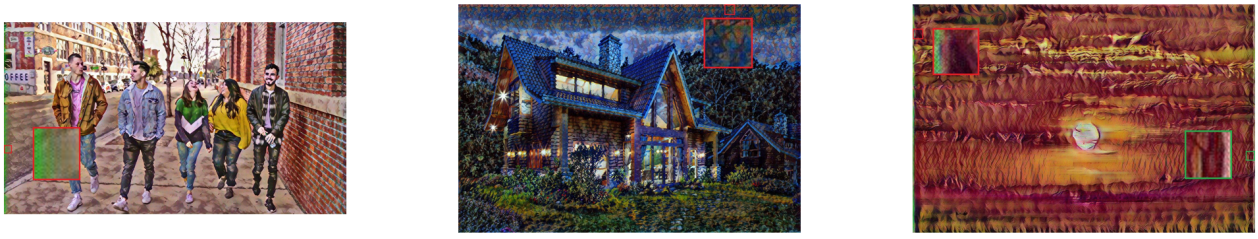

Fig. 5 Common border artifact present in stylizations by ArtFlow.

Recoloring was performed in LAB color space, since this color space is more natural: uniform changes of components in the LAB color correspond to uniform changes in perceived color by human eye[⑦]. We also experimented with histogram matching [⑧], using skimage package, but it performed worse.

2.5 Labeling Process

Each of 10.000 stylizations were labeled by three annotators - the authors of the article having expertise in various style transfer methods. Each stylization was rated on scale from 1 (worst) to 10 (best), giving 30.000 ratings.

We acknowledge that in the world of big data this dataset is relatively small, but since we distribute not only the dataset, but the specialized software[①] to rate stylizations, we encourage research community to extend it. Also, since we rated a diverse set of common content images, proposed dataset is versatile enough to come up with general conclusions about common cases when style transfer fails and succeeds to achive its task of rendering a photo in an artistic expressive way.

During labeling process, stylizations were shown to each annotator in full screen. The annotators were asked to cover the whole range of grades from 1 to 10 dateset-wide, however there were no restrictions to cover this range for particular content and style. Indeed, some contents and styles yielded good and some - bad results on average.

Only stylizations were shown, and annotators were asked to grade them according to their subjective aesthetic pleasure, by "willingness to use it as a picture on the wall or as an illustration on a website of particular topic". Annotators were asked not to take the actual content intro account - only artistic expressiveness and impact caused by style transfer. For particular styles stylization produced photorealistic result without any artistic effect. In such cases annotators were asked to grade stylization in the range 4-6, depending on color vividness and presence of style transfer artifacts.

It was noted, that ArtFlow frequently produces artifacts at the borders of the stylization image, as shown on Fig. 5.

Annotators were asked not to downgrade stylization for the presence of this artifact, since it can be easily removed by slight border cropping. All other artifacts were downgraded.

⑦ https://en.wikipedia.org/wiki/CIELAB_color_space
⑧ https://en.wikipedia.org/wiki/Histogram_matching
① Позже вставлю ссылку

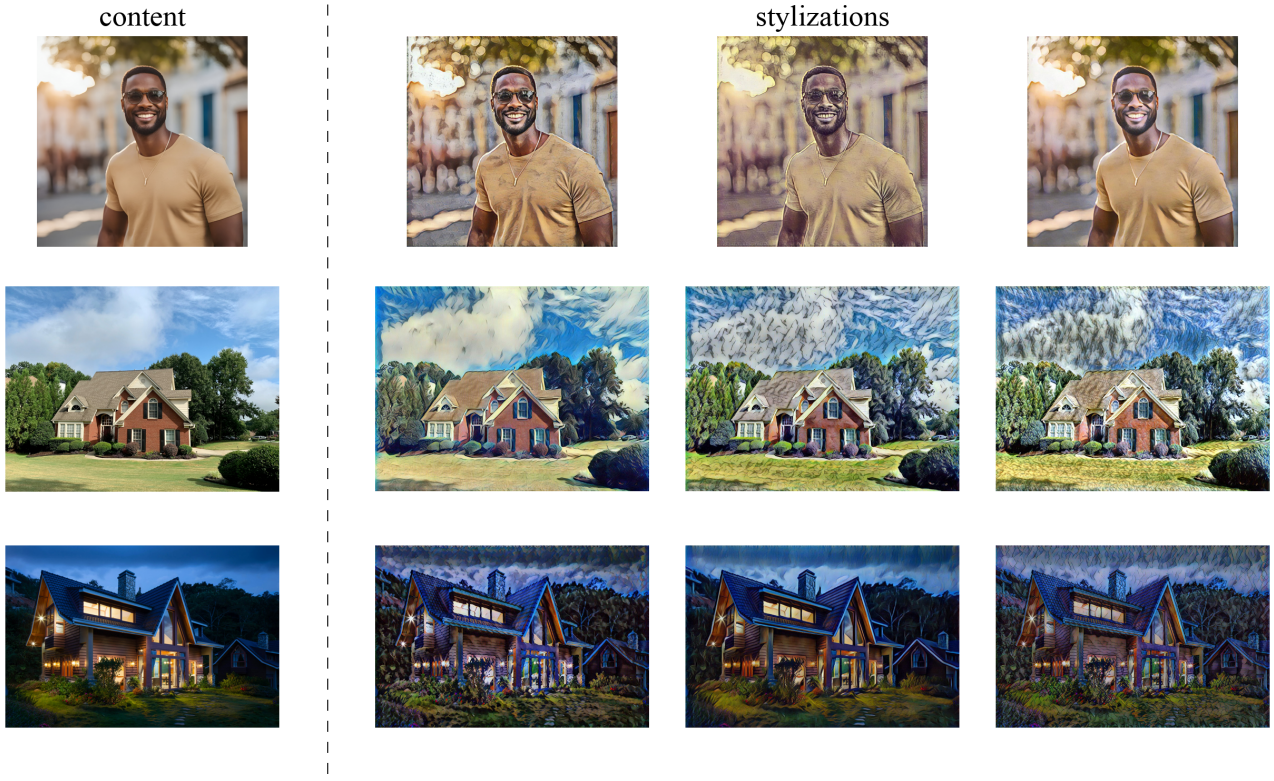

Fig. 6 Examples of contents that provide high quality stylizations

	worst					best				
content №	14	17	36	43	19	4	20	5	3	26
mean rating	4.6	4.64	4.9	4.91	5.06	6.44	6.46	6.6	6.62	6.65
std rating	1.41	1.5	1.48	2.2	1.51	1.34	1.26	1.42	1.22	1.07

Table 1 5 contents with the least rating and 5 contents with the greatest rating in the dataset

	rating 1	rating 2	rating 3
rating 1	1.0000	0.4030	0.3612
rating 2		1.0000	0.4314
rating 3			1.0000

Table 2 Kendall's Tau-B correlations between ratings, p-value = 0 (float64)

measure	$\sigma(L)$	$\sqrt{\sigma(L)^2 + \sigma(A)^2 + \sigma(B)^2}$	sharpness score
Kendall's Tau-B	0.0846	0.0941	-0.2258
p-value	3e-35	3e-43	2e-253

Table 3 Kendall's Tau-B correlations between mean rating and different measures

Я выложу потом на гитхаб датасет и программу разметки.

3 General Analysis of Ratings

As shown on Fig. 3, the mean rating has a skewness towards the higher ratings. Individual ratings have very different distributions, from which we conclude that the quality of stylizations highly depends on the taste of the person, labeling the data. Although the rating distributions differ, the standard deviation of the ratings mostly doesn't exceed 2 grades, as seen on Fig. 4. To analyze the correlations between ratings, we use Kendall's Tau-B measure implementation from scipy library with the two-sided alternative (null hypothesis – the rank correlation is zero) [8] [9]. The correlations in table 2 show that the ratings have a monotonous dependency between each other, and thus we can use the averaged ratings to get insights to the quality of the stylization results.

In Table 1 we can see that the quality of the stylization highly depends on the content image: some images have lower average rating than others.

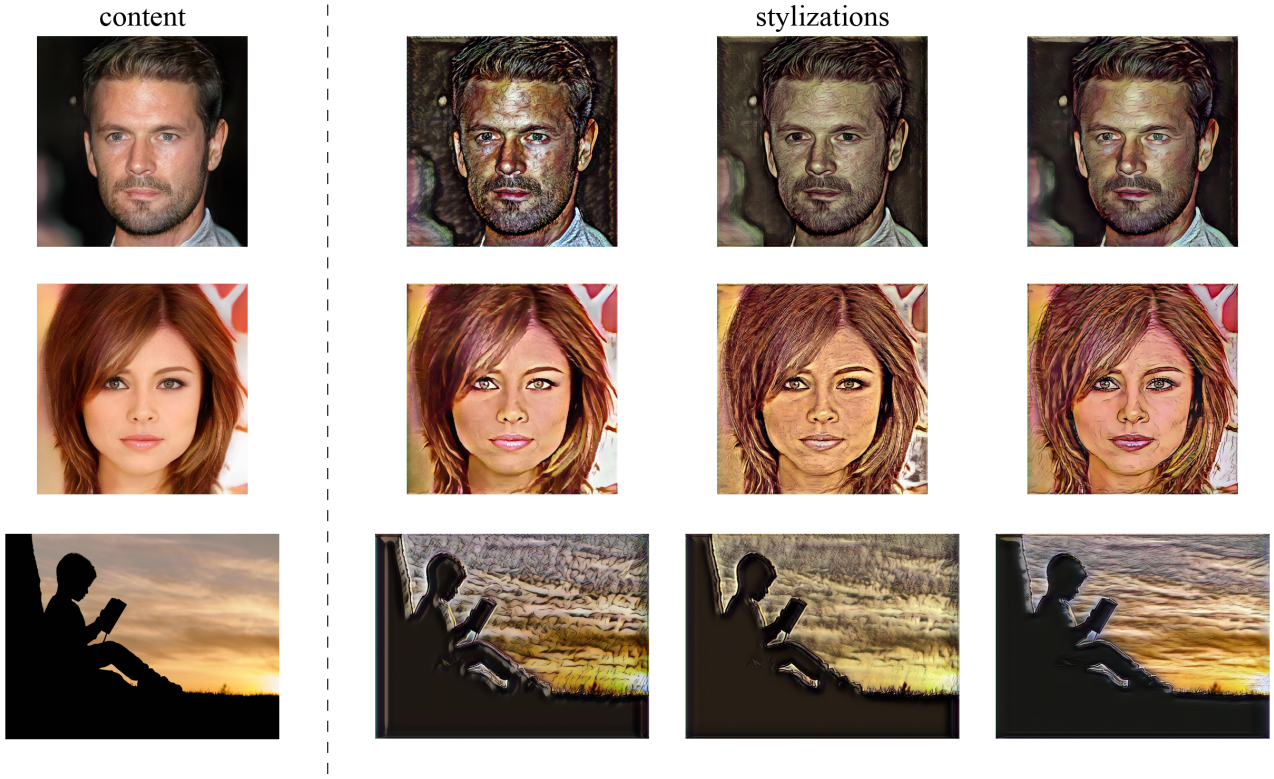

Fig. 7 Examples of contents that provide low quality stylizations

measure	$dist(I_{content}, I_{stylization})$
Kendall's Tau-B	-0.4696
p-value	1e-259

Table 4 Kendall's Tau-B correlations between mean rating and $dist(I_{content}, I_{stylization})$ with independence Null hypothesis. This analysis was restricted to pairs from dataset for which $Faces(I_{content}) \neq \emptyset$

style size (px)	150^2	300^2	500^2	700^2
mean rating	5.19	5.95	6.11	6.09
std rating	1.43	1.46	1.51	1.55

Table 5 Style sizes and their respective rating statistics

4 Discussion and Recommendations

4.1 Color diversity and sharpness

Highly rated stylizations have several common features: they do not heavily distort the shapes of the objects in the image, but add local textures from the style image to the background. The quality of stylizations also depends on the color and the brightness diversity of the resulting image. To compute the diversity, we convert the stylizations to the LAB color space and compute the brightness diversity as $\sigma(L)$ and color diversity as $\sqrt{\sigma(L)^2 + \sigma(A)^2 + \sigma(B)^2}$. The Kendall's Tau-B correlation coefficients in Table 3 show that a correlation between mean rating and these measures exists.

The sharpness of the stylization also affects the ratings. We compute the sharpness of the image as the variance of the image laplacian (the result of applying the discrete laplace filter to the image) as proposed in Pech-Pacheco et al. [10]. The correlation of this measure with the rating is negative, as shown in the Table 3: if the stylization is too sharp, the overall quality is lower. This can be explained by the intuition behind the proposed measure – sharper images have a higher variance, and if the image is too sharp, it may contain a number small-sized artifacts, which drastically decreases the rating.

4.2 Texture blur and edges reproduction

Stylizations with low ratings show that the style transfer process highly depends on the content images. The contents with the lowest average ratings are present on the Fig. 7. Full-face portraits are rated lower than other pictures. The reason for it is the distortion of

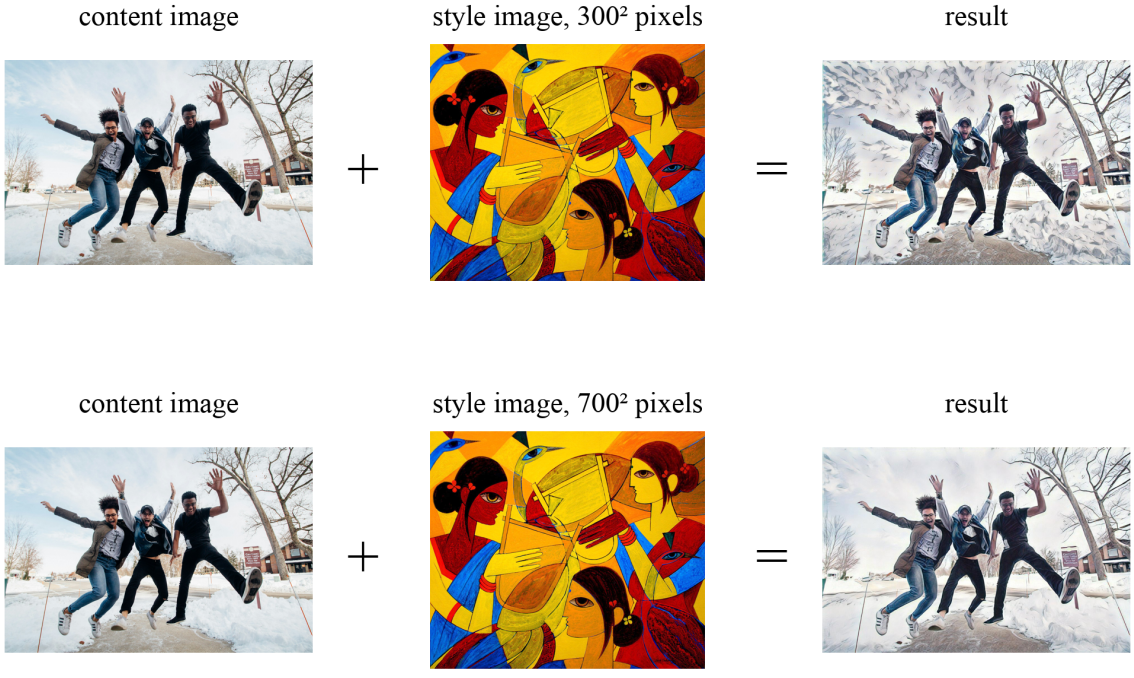

Fig. 8 Example: the style transfer algorithm fails to capture style features with styles of larger size

	worst					best				
style №	7	7	7	38	7	43	41	19	43	43
style size (px)	150^2	300^2	500^2	700^2	700^2	300^2	700^2	300^2	700^2	500^2
mean rating	1.87	2.02	2.1	2.24	2.28	7.85	7.86	7.92	7.98	8.28
std rating	1.06	1.25	1.28	0.84	1.25	1.28	1.38	1.26	1.27	1.11

Table 6 5 styles with the least rating and 5 styles with the greatest rating in the dataset

facial features – style transfer methods tend to apply local style textures to human skin and hair, which makes them blurry and unpleasant-looking.

Another common feature in contents with low average stylization rating is the presence of an area with low color diversity and clear edges, like shadows. They get distorted after style transfer and tend to contain visual artifacts.

4.3 Face reproduction

As mentioned above, content images with full face portraits are generally sensitive to the style transfer operation. We explain this by the fact that small changes in facial features lead to a decrease in the recognizability of a particular person's face in the image, which plays an important role in the case of subjective perception of stylization. We were able to find a strong anti-correlation of the distance metric between faces in the content and stylized image with the stylization quality scores. The metric used was the cosine distance between the embeddings of the face areas of content and stylizations. The embeddings were obtained using the VGG-face model proposed by [7] and implemented in the DeepFace library[①].

Define content image with height $H_{content}$ and width $W_{content}$ as $I_{content} \in \{x \in \mathbb{Z} \mid 0 \leqslant x \leqslant 255\}^{H_{content} \times W_{content}}$.

Define stylized content image as $I_{stylization} \in \{x \in \mathbb{Z} \mid 0 \leqslant x \leqslant 255\}^{H_{content} \times W_{content}}$.

Define for image $X \in \{x \in \mathbb{Z} \mid 0 \leqslant x \leqslant 255\}^{H \times W}$, $X_{xyhw} = \{X_{ij}\}_{i=x,j=y}^{x+h-1,y-w+1} \in \mathbb{R}^{h \times w}$.

For each content image, we manually identified the set of faces present. For each detected face, we determined a rectangular bounding box defined by its coordinates and dimensions.

Let $Faces(I_{content}) \in \mathbf{2}^{\mathbb{R}^4}$ denote the set of vectors

① https://github.com/serengil/deepface

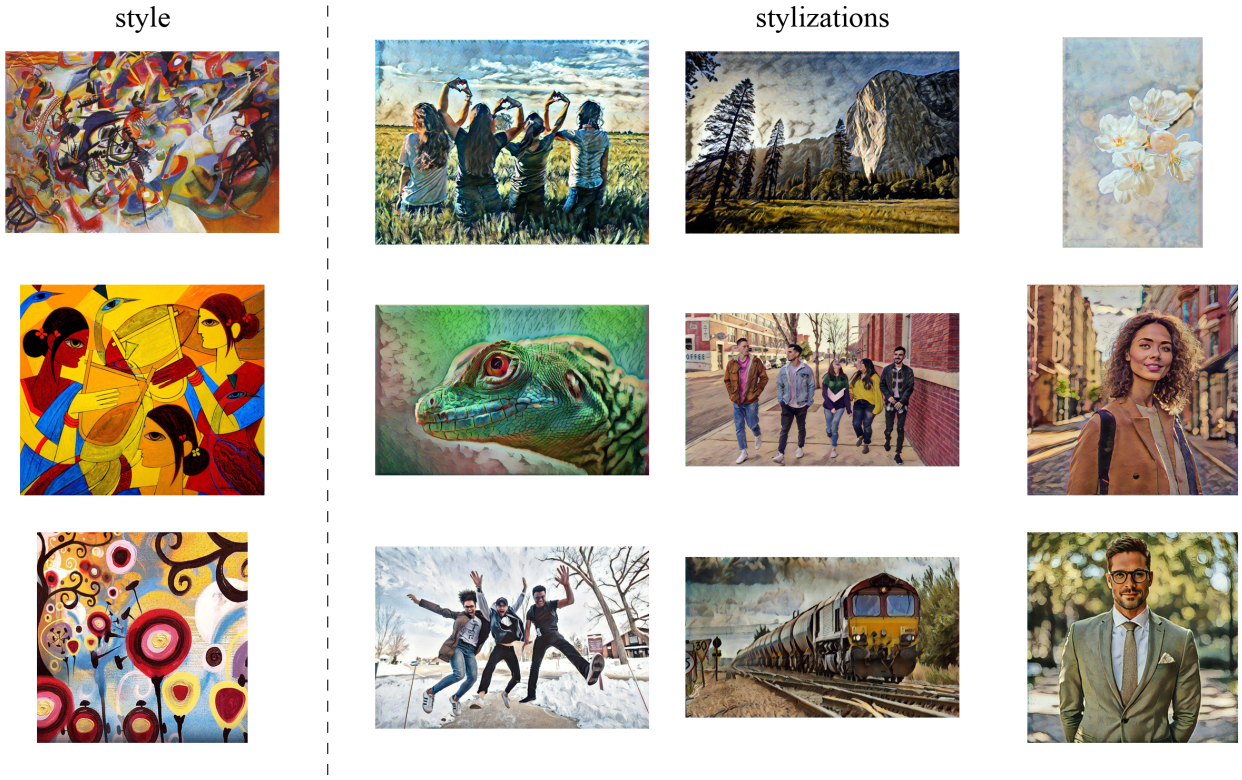

Fig. 9 Examples of styles that provide high quality stylizations. Styles respective sizes are 300^2, 500^2 and 700^2 pixels

(x, y, h, w), where (x, y) represents the coordinates of the top-left corner of the bounding box, and (h, w) specifies its height and width. By definition, for each identified face in $I_{content}$ $Faces(I_{content})$ includes exactly one element describing bounding box of that identified face.

Define last layer output of VGG-Face model applied to some image $X \in \{x \in \mathbb{Z} \mid 0 \leqslant x \leqslant 255\}^{H \times W}$ as $emb(X) \in \mathbb{R}^{2622}$,

$$dist(I_{content}, I_{stylization}) = \frac{1}{|Faces(I_{content})|} \sum_{(x,y,h,z) \in Faces(I_{content})} 1 - \frac{\langle emb(I_{content_{xyhw}}), emb(I_{stylization_{xyhw}}) \rangle}{\|emb(I_{content_{xyhw}})\|_2 \cdot \|emb(I_{stylization_{xyhw}})\|_2} \quad (1)$$

The results of the experiment in the Table 4 confirm the above hypothesis.

4.4 Style size and characteristics

The stylization is heavily dependent on the style images. In the Table 5 mean ratings aggregated by style image size show that styles with smaller resolution are graded worse than others. This can also be seen on the Fig. 2. In the small-sized style images the style features are too frequent, which results in noisy artifacts on the stylization. On the other hand, the difference between mean ratings in the higher sizes is not so evident, but the images of the highest resolution of 700×700 pixels are graded slightly worse than images containing 500×500 pixels. That can be explained by the architecture of encoder networks in style transfer algorithms: the receptive field of the convolution layers is fixed, and it may neglect local style features, as we can see on the Fig. 8.

Provided the style size is so important, we analyze the grades with respect to styles and their sizes at the same time. As we can see in the Table 6, the choice of the style-size combination affects the quality of the stylization much stronger than the choice of the content. The style-size combinations with the highest average ratings are present on the Fig. 9. They have common features: these images have a strong color diversity, different local textures and shapes and the smallest of them has the 300×300 resolution. Style transfer does not necessarily apply the artistic styles of these images to the contents, but it is able to transfer small local features to make the stylizations look pleasant.

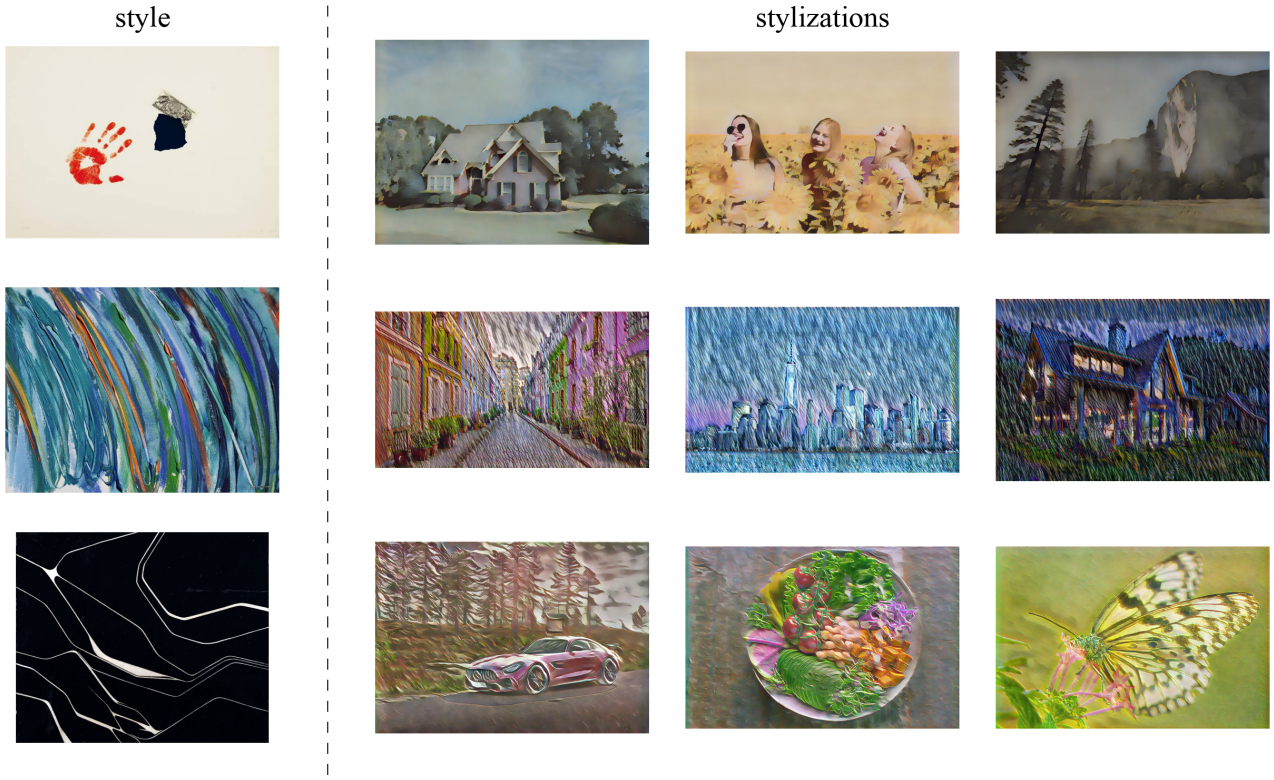

Fig. 10 Examples of styles that provide low quality stylizations. Styles respective sizes are 150^2, 300^2 and 500^2 pixels

On the other hand, styles with low average ratings, as seen on Fig. 10, can lack a variety of textures and color diversity (the first and the last style). This results in stylizations being bleached and also removes some contents of the image. That happens because the style image is too simple and the content distributions are too far from the style's. The second bad style image presents another problem: high color diversity in a combination with a complex high-frequency texture can make the style transfer algorithm apply randomized textures to the content images, effectively blurring them and creating many artifacts.

4.5 Recommendations

As the result of a thorough analysis of the dataset, we provide a list of recommendations for choosing a style image:

(1) The size should not be small.
(2) The style transfer algorithm should adapt to contents and styles of different sizes.
(3) The image should contain a variety of textures. Oversimplified styles result in bad stylizations.
(4) High-frequency textures affect the stylizations negatively.
(5) The style image should contain diverse colors.

This list of recommendations can also be interpreted as a list of future developments for the style transfer algorithms.

5 Conclusion

We introduced the first style transfer dataset, covering various content and style images and containing 10.000 stylizations, rated by three annotators. This dataset can be used to train models, predicting best content, style and style size combinations.

Given our experience in labeling a vast amount of stylizations, we provided an extensive analysis of what makes a good and bad stylization and summarized it in a concise list of recommendations for future style transfer improvement.

We hope that this work will drive the advancement of style transfer algorithms, making their results more vivid and appealing for the end user.

Список литературы

[1] An, Jie, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. Artflow: Unbiased image style transfer via reversible neural flows. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 862–871, 2021.

[2] Liu, Songhua, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. Proceedings of the IEEE/CVF international conference on computer vision, pp. 6649–6658, 2021.

[3] Deng, Yingying, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11326–11336, 2022.

[4] Gatys, Leon A, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3985–3993, 2017.

[5] Karras, Tero, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196, 2017.

[6] Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv preprint arXiv:2112.10752, 2021.

[7] Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman. Deep Face Recognition. Proceedings of the British Machine Vision Conference 2015, pp. 41.1–41.12, 2015.

[8] Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stefan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, Ilhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antonio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and the SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, vol. 17, pp. 261–272, 2020. https://rdcu.be/b08Wh, DOI: 10.1038/s41592-019-0686-2.

[9] Kendall, M. G. The Treatment of Ties in Ranking Problems. Biometrika, vol. 33, no. 3, pp. 239-251, 1945. https://doi.org/10.1093/biomet/33.3.239.

[10] Pech-Pacheco, J.L., G. Cristobal, J. Chamorro-Martinez, and J. Fernandez-Valdivia. Diatom autofocusing in brightfield microscopy: a comparative study. Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, vol. 3, pp. 314-317, 2000. DOI: 10.1109/ICPR.2000.903548.