The University Of Sheffield.

This is a repository copy of *Evaluation in neural style transfer: a review*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/215700/

Version: Published Version

Article:

White Rose
university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Evaluation in Neural Style Transfer: A Review

Eleftherios Ioannou and Steve Maddock

The University of Sheffield, Sheffield, UK
{eioannou1, s.maddock}@sheffield.ac.uk

Abstract

The field of neural style transfer (NST) has witnessed remarkable progress in the past few years, with approaches being able to synthesize artistic and photorealistic images and videos of exceptional quality. To evaluate such results, a diverse landscape of evaluation methods and metrics is used, including authors' opinions based on side-by-side comparisons, human evaluation studies that quantify the subjective judgements of participants, and a multitude of quantitative computational metrics which objectively assess the different aspects of an algorithm's performance. However, there is no consensus regarding the most suitable and effective evaluation procedure that can guarantee the reliability of the results. In this review, we provide an in-depth analysis of existing evaluation techniques, identify the inconsistencies and limitations of current evaluation methods, and give recommendations for standardized evaluation practices. We believe that the development of a robust evaluation framework will not only enable more meaningful and fairer comparisons among NST methods but will also enhance the comprehension and interpretation of research findings in the field.

Keywords: image and video processing, rendering; non-photorealistic rendering

CCS Concepts: • Computing methodologies → Image processing; Non-photorealistic rendering

1. Introduction

Neural style transfer (NST) refers to the class of methods that attempt to synthesize artistically stylized visual media. For images, where it is most predominantly applied, NST is the process of transferring the style of an artistic image onto an ordinary photograph, a process pioneered in the seminal work of Gatys et al. [GEB15]. Successive visual synthesis models have demonstrated remarkable improvements, and their capability to generate impressive outputs have attracted the attention of both academia and industry. Besides the significant advancements to the quality of the generated results, style transfer has also been extended to work for a range of media: images, videos, 3D meshes, point clouds and radiance fields.

To assess the quality of the results of style transfer methods, there exists an array of diverse evaluation techniques. Visual side-by-side comparisons are used to evaluate results qualitatively; user surveys are employed to quantitatively gauge the performance of the proposed models from collected subjective responses; and computational metrics attempt to quantify content retainment, style resemblance and overall efficiency. Despite the availability of a plethora of evaluation methodologies, the evaluation process differs amongst the NST algorithms. Depending on the medium, but also on the par-

ticular advancement or contribution a method aims to achieve, each style transfer approach resorts to different evaluation procedures. As a result, there are no established benchmarks or evaluation protocols.

The challenge in establishing a universally accepted evaluation protocol is, in part, rooted in the definition of NST. Artistic NST algorithms seek to generate outputs that resemble the reference artwork while retaining the content of the image that is stylized – by definition, the most significant aspects of the evaluation of NST results are *content preservation* and *style performance*. However, there is no agreement as to where the line between content fidelity (CF) and stylization intensity should be optimally drawn. This observation magnifies the difficulty of the task, which arguably encapsulates a degree of subjectivity. The subjectivity involved in the assessment of the outputs of NST methods, the particular intentions and goals of each approach and the perspective from which each method attempts to tackle the problem of stylization, render evaluation a complicated and demanding activity.

The existing evaluation approaches in NST research are typically classified into two categories: qualitative and quantitative [JYF*19]. *Qualitative evaluation* relies upon the subjective judgements of the
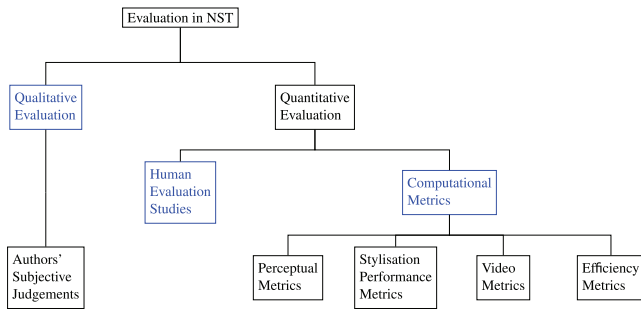
Figure 1: *Evaluation in NST: qualitative evaluation and quantitative evaluation. To align with current literature, we use three main categories for evaluation, highlighted in blue. NST, neural style transfer.*

corresponding authors. *Quantitative evaluation* includes numerous computational metrics to derive exact numerical estimations of particular aspects of the performance of the evaluated approach, as well as quantitative data from user studies that can be statistically interpreted.

In current literature, however, user survey results are often excluded from the qualitative or quantitative evaluation sections of the studies. The majority of the approaches provide the human evaluation studies' details and results in distinct sections, with minor exceptions [LYH21, SKLO18]. Notwithstanding, the user studies conducted in NST only gather quantitative data: the participants are either asked to choose their favourite stylization or rank a set of images, without any textual feedback or qualitative observations being collected. Consequently, it is justifiable to classify the data collected from the NST user studies as quantitative. While collecting qualitative data from user surveys could be a useful and insightful practice, in their current state, user surveys are solely utilized for the collection and analysis of quantitative data. The results of a user study can be depicted numerically and can allow for quantitative assessments; yet, other vital concerns arise, such as the necessity to estimate and explicitly indicate the statistical significance of those results. Reproducibility and repeatability are also crucial factors that should be taken into account to ensure the reliability of the reported results, similar to the results calculated using computational metrics.

An overview of the evaluation landscape in NST is given in Figure 1. Human evaluation studies and computational metrics are included in the quantitative evaluation category. However, to be consistent with current literature, we structure our review into three evaluation areas (highlighted in blue): *qualitative evaluation*, *human evaluation studies* and *quantitative evaluation metrics* (Figure 3 gives an in-depth illustration of these areas). All approaches provide a qualitative evaluation to some extent – visual side-by-side results are presented accompanied by subjective assessments of the authors. Whilst user studies are quite commonplace, not every method agrees on the structure of the user study that is conducted. Also, virtually all the methods quantitatively gauge the performance of their system selectively, drawing from

a pool of different computational metrics. This results in a variety of different metrics being present in each of the methods' evaluation sections, with no shared list of metrics being agreed upon.

Arguably, evaluation is an essential aspect of NST research and one that is catalytic to the success and progress of the field, as it can provide a solid understanding of the proposed approaches and a bedrock for novel research. This work aims to provide a comprehensive review of the current evaluation methods and identify the commonalities and differences between the state-of-the-art approaches. An extensive, in-depth survey of the evaluation techniques in NST would provide useful insights and potential solutions for the establishment of robust and reliable evaluation methodologies.

Previous review papers [JYF*19, SJJ*21] have analyzed the techniques and applicability of NST, but have not focused on evaluation methodologies. To our knowledge, this is the first work that directly considers the problem of evaluation in NST, reviewing all the approaches to evaluation and their limitations. As part of the review, we aim to highlight the most important evaluation techniques and propose suggestions for the design of a standardized evaluation procedure.

The rest of the paper is organized as follows. Section 2 provides an overview of the NST literature, distinguishing the different approaches by the visual medium they are applied to. Section 3 examines the evaluation techniques and metrics. Section 4 presents a review of the evaluation data and the qualitative and quantitative evaluation approaches in the current literature. It also provides an analysis of those methods, complemented with an array of recommendations.

2. Neural Style Transfer Methods

Jing et al. [JYF*19] categorize NST algorithms into Image-Optimization-Based Online Neural methods (which optimize a generative model online) and Model-Optimization-Based Offline Neural methods (which optimize a generative model offline, then produce a stylization with a single forward pass). The lower tiers in their hierarchical taxonomy split the methods into Non-Photorealistic and Photorealistic branches with the ending nodes further classifying the methods into Image and Video. For this paper, we follow a similar taxonomy by distinguishing the reviewed NST methods by the medium they are applied to images and videos.

We specifically concentrate on models that, once trained, are capable of synthesizing a stylized output given an input content image/video (and potentially a reference style image depending on the network's capability to reproduce multiple or arbitrary styles). We do not include text-to-image techniques, such as those utilizing *CLIP* [RKH*21]. Our review of evaluation techniques is based on the analysis of a single output. Consequently, we do not assess different aspects of stylization, such as interpolation between various styles. Moreover, although NST is also applied to 3D data, here we mainly focus on the evaluation techniques used by NST methods applied to images (Tables 1–6) and/or videos (Table 7).

Table 1: *Online image optimization methods.*

Year	Method	Mode		Evaluation		
		artistic	Photorealistic	qualitative	User study	Quantitative metrics
2016	Gatys et al. [GEB15]	✓	✓			
2017	Gatys et al. [GEB*17]	✓				
2017	Li et al. [LWLH17]	✓				
2017	Risser et al. [RWB17]	✓		✓		Speed
2017	Li et al. [LXNC17]	✓		✓		Content loss and style loss
2017	Li and Wand [LW16a]	✓	✓	✓	I	
2017	Luan et al. [LPSB17]		✓	✓	✓	
2017	Mechrez et al. [MSZM17]		✓	✓	✓	Speed
2019	Penhöuet and Sanzenbacher [PS19]		✓			

'I' under user study denotes 'Informal'.

Table 2: *Per-style-per-model offline model optimization methods.*

Year	Method	Dataset		Evaluation		
		MS COCO	ImageNet	qualitative	User Study	Quantitative metrics
2016	Johnson et al. [JAFF16]	✓		✓		Content loss and style loss, Speed
2016	Li and Wand [LW16b]		✓	✓		Speed, memory
2016	Ulyanov et al. [ULVL16]		✓	✓		Speed, memory
2017	Ulyanov et al. [UVL17a]	✓	✓	✓		
2017	Liu et al. [LCLR17]	✓		✓		Depth/edge/saliency map
2019	Cheng et al. [CLW*19]	✓		✓	✓	Speed, memory, SSIM, depth map, edge map
2022	Ioannou and Maddock [IM22]	✓		✓	✓	Depth map, saliency map, SSIM, hist, aHash, dHash

All methods are suitable for artistic style transfer.

Table 3: *Per-artist-per-model offline model optimization methods.*

Year	Method	Dataset	Evaluation			
			qualitative	User Study	User study w/experts	Quantitative metrics
2018	Sanakoyeu et al. [SKLO18]	Places365 and	✓		✓	Deception Rate, Speed, Memory
2019	Kotovenko et al. [KSLO19]	Wikiart	✓	✓	✓	Deception Rate, Style Divergence (D_{KL})

The methods are suitable for transferring a particular style, or the style of a specific art genre/artist.

2.1. Images

2.1.1. Online image optimization neural methods

Gatys et al. [GEB15, GEB16] noticed that the powerful capabilities of convolutional neural networks (CNNs) can be utilized for more than just object classification (Table 1). Observing that an input image can be transformed into feature maps that increasingly care about the content rather than any detail about the colour or texture as we go deeper into a CNN, they proposed an image-optimization technique that faithfully replicates artistic qualities of artworks on real photographs. Their technique falls into the Image-Optimization-Based Online neural methods and attempts to minimize an objective function that encompasses definitions of content and style loss. The content loss is defined by the squared Euclidean distance between the feature representations $F^l(l = conv4\_2)$ of the content image (I_c) and the target

image (I_t):

$$\mathcal{L}_{content} = \frac{1}{2} \sum_l \left(F^l(I_t) - F^l(I_c) \right)^2 \tag{1}$$

To represent the style of the input style image, a feature space is designed to capture the texture and colour information. This space looks at spatial correlations within a layer of a network. Features are extracted from multiple layers ($conv1\_1$, $conv2\_1$, $conv3\_1$, $conv4\_1$, $conv5\_1$) and the feature correlations are given by the Gram matrix which contains non-localized information about the image. The resulting style loss is defined by the squared Euclidean distance between the Gram-based style representations of the style image I_s and target image I_t:

$$\mathcal{L}_{style} = \gamma \sum_l w_l \left(\mathcal{G}(F^l(I_t)) - \mathcal{G}(F^l(I_s)) \right)^2 \tag{2}$$

E. Ioannou, S. Maddock / Evaluation in NST: A Review

Table 4: *Multiple-style-per-model offline model optimization methods.*

Year	Method	Dataset content	No. styles	Evaluation qualitative	User study	Quantitative metrics
2017	Dumoulin et al. [DSK17]	ImageNet	32			
2017	Chen et al. [CYL*17]	MS COCO	50	✓		
2017	Li et al. [LFY*17a]	DTD	1000	✓		
2018	Zhang and Dana [ZD18]	MS COCO	1000	✓		Speed, Memory

The methods are suitable for artistic style transfer.

where γ is a constant that accounts for the number of values in each layer and w_l are weighting factors for the contribution of each layer.

To generate the target image that renders the input content image artistically stylized based on the style of the input style image, the system then tries to minimise the loss function:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{content} + \beta \mathcal{L}_{style} \quad (3)$$

where α and β are used to control the weight of the content and style components in the stylized result, with β often being much larger. These values stay fixed over the stylization process. An extension of this algorithm was the introduction of style factorization that allows stylizing different regions of the image with different styles [GEB*17]. Other image-optimization-based online methods include the systems proposed by Risser et al. [RWB17] and Li et al. [LXNC17], whereas the work of Li et al. [LW16a] attempts to model the style in a non-parametric way utilizing a patch-based Markov random field prior in their style loss.

As these algorithms were implemented in the early period of NST research, no particular emphasis was given to evaluation. After the work of Gatys et al. [GEB15, GEB16], researchers started to provide qualitative evaluations in the form of visual side-by-side comparisons [RWB17, LXNC17, LW16a]. Risser et al. [RWB17] included speed comparisons, Li et al. [LXNC17] discussed content loss and style loss as a way of measuring stylization performance, and Li and Wand [LW16a] provided some insights from an informal user study.

2.1.2. Offline model optimization neural methods

Whilst online image optimization NST techniques produce visually appealing results, they are inefficient in terms of speed and computational cost. Model-optimization-based offline neural methods were introduced to mitigate the slow optimization process. These techniques attempt to optimize a generative model offline and produce a stylized image faster, with a single forward pass through the trained model. Formally, these models learn a style transfer function Φ_w which can be expressed by a feed-forward network with parameters w, and is optimized over a set of content images I_c for one or more style images I_s:

$$\mathrm{w}^* = \arg\min_{\mathrm{w}} \mathcal{L}_{total}(I_c, I_s, \Phi_{\mathrm{w}^*}(I_c)) \quad (4)$$

As described by Jing et al. [JYF*19], depending on the number of styles that the feed-forward network can reproduce, the Model Op-

timization methods can be distinguished into *per-style-per-model (PSPM)* methods, *multiple-style-per-model (MSPM)* methods, and *arbitrary-style-per-model (ASPM)*. In this section, we introduce another category, *Per-Artist-Per-Model*, in order to accommodate the methods that attempt to capture a particular artistic style or genre instead of emulating one particular input reference style image.

Per-style-per-model methods

Johnson et al. [JAFF16] was the first to address the inefficiency of the method of Gatys et al. [GEB16] (Table 2). Although the optimization problem remains the same, their work utilizes perceptual loss functions to train their network. Consequently, they generate stylized images three orders of magnitude faster. Further improvements to the inference speed of the trained image transformation network were introduced by Ulyanov et al. [ULVL16]. Later, they published an improvement to their algorithm [UVL17a] that achieves visually better results. This was based on the idea of replacing the batch normalization layers in the generator with instance normalization [UVL17b]. The work by Li and Wand [LW16b] addresses the speed issue by training a Markovian feed-forward network that also reproduces one reference style.

More recently, researchers have considered how to preserve the depth and structural information of the content images when performing stylization. Models have been developed that attempt to alleviate the undesired effects that occur when the input images include numerous objects at various depths. The system of Liu et al. [LCLR17] is built on the work of Johnson et al. [JAFF16] but introduces a depth reconstruction loss during training. Similarly, Cheng et al. [CLW*19] rely on depth maps and image edges to generate results of higher quality and retained structure. Other work combines the suggested depth loss [LCLR17] with an improved depth prediction network and with instance normalization layers for aesthetic and depth-preserving stylized images [IM22].

Per-artist-per-model methods

Since the seminal work of Gatys et al. [GEB16], researchers have attempted to view and tackle the style transfer problem from different perspectives (Table 3). Some approaches have tried to redefine what style is and how it can be captured and reproduced, while other methods address the aesthetics of the results. The system proposed by Sanakoyeu et al. [SKLO18] avoids fixed style representations and focuses on the details that are relevant to the

Table 5: *Arbitrary-style-per-model offline model optimization methods for artistic NST.*

| Year | Method | Dataset | | | Evaluation | | |
		MS COCO	Wikiart	other	Qualitative	User study	Quantitative metrics
2016	Chen and Schmidt [CS16]	✓	✓		✓		Speed
2017	Huang and Belongie [HB17]	✓	✓		✓		Content and style Loss, speed
	Ghiasi et al. [GLK*17]			ImageNet & PBN, TDT	✓		
	Li et al. [LFY*17b]	✓			✓	✓	Style loss, speed
2018	Gu et al. [GCLY18]			ImageNet	✓	✓	Speed
	Shen et al. [SYZ18]	✓	✓		✓		Speed, memory
	Xu et al. [XWF*18]			Behance, DTD, CelebA, UCSD birds, cars, Oxford building	✓	✓	
2019	Li et al. [LLKY19]	✓	✓		✓	✓	Speed
	Park and Lee [PL19]	✓	✓		✓	✓	Speed
2020	Hu et al. [HJL*20]	✓	✓		✓	✓	
	Svodoba et al. [SAOM20]		✓	Places365	✓		
2021	Liu et al. [LLH*21]	✓	✓		✓	✓	Speed
	An et al. [AHS*21]	✓	✓	Metfaces	✓	✓	SSIM, content and style error, content leak, speed
	Liu et al. [LYH21]	✓		PF-PASCAL	✓	✓	Speed
	Huo et al. [HJL*21]	✓	✓		✓	✓	Speed
	Liu and Zhu [LZ21]	✓	✓		✓	✓	Depth map, Edge map, saliency map, SSIM, hist, aHash, dHash
	Chen et al. [CWZ*21]	✓	✓		✓	✓	Speed
2022	Deng et al. [DTD*22]	✓	✓		✓	✓	SSIM, content and style error, content leak, speed
	Luo et al. [LHYZ22]	✓	✓		✓	✓	Speed
	Wang et al. [WZZ*22]	✓	✓		✓	✓	CF, GE, LP, speed
2023	Ruta et al. [RGC*23]	✓	✓		✓	✓	Chamfer, LPIPS, SFID, speed
2023	Huang et al. [HAW*23]	✓	✓	FFHQ MetFaces LandscapesHQ	✓	✓	LPIPS, style error FID, ArtFID, speed
2023	Xu et al. [XLN23]	✓	✓		✓	✓	LPIPS, style error speed
2023	Tang et al. [TLL*23]	✓	✓		✓	✓	Content and style error, $\mathcal{L}_{sim}$
2023	Hong et al. [HJL*23]	✓	✓		✓	✓	Style eError, CF, LP, speed
2023	Zhu et al. [ZHW*23]	✓	✓		✓	✓	Content and style error, LPIPS, Speed
2023	Gu et al. [GFZ23]	✓	✓		✓	✓	Content and style error, colour loss, speed
2023	Ma et al. [MZLB23]	✓	✓		✓	✓	Content and style error, LPIPS, speed
2023	Li et al. [LLW*23]	✓	✓		✓	✓	Content and style error, SSIM, speed
2023	Zhang et al. [ZTD*23]		✓	Places365	✓	✓	Content error, LPIPS deception rate

style when measuring the similarity in content between the input and the stylized image. This more generalized procedure is capable of transferring not only the style of one particular painting but also the distinctive style of *Cezanne* or *Van Gogh*. In a similar spirit, Kotovenko et al. [KSLO19] propose a fixpoint triplet style loss (to learn nuanced variations within a style) and a disentanglement loss (to prohibit the influence of the content on the stylization), and extract style from a group of examples of the same overall style (but with subtle variations) in order to generate stylizations in the style of *Picasso* or *Kandinsky*.

Multiple-style-per-model methods

An improvement to the style transfer algorithms was to create models capable of reproducing more than one style per trained network (Table 4). Dumoulin et al.'s method [DSK17] reduces style images into points in an embedding space and employs conditional instance normalization, allowing the style transfer network to learn multiple styles. Chen et al.'s [CYL*17] *StyleBank* decouples content and style representations utilizing multiple convolution filter banks with each filter bank representing one style, making the method capable

E. Ioannou, S. Maddock / Evaluation in NST: A Review

Table 6: *Arbitrary-style-per-model offline model optimization methods for photorealistic style transfer.*

Year	Method	Dataset	Evaluation		
			qualitative	User Study	Quantitative metrics
2018	Li et al. [LLL*18]	MS COCO	✓	✓	Speed
2019	Yu et al. [YUC*19]		✓	✓	Style loss, SSIM, speed, memory
2020	An et al. [AXHL20]		✓	✓	SSIM (edges, whole), speed
2023	Ke et al. [KLZ*23]		✓	✓	SSIM, style similarity speed

Table 7: *Video NST Methods.*

Year	Method	Dataset			Evaluation		
		MS COCO	Wikiart	Other	qualitative	User study	Quantitative metrics
Online Frame-By-Frame Optimization							
2016	Ruder et al. [RDB16]				✓		Speed, warping error
Per-Style-Per-Model (Offline Model Optimization)							
2017	Huang et al. [HWL*17]			Videvo.net	✓	✓	Speed, warping error
2018	Ruder et al. [RDB18]	✓		Hollywood2	✓	✓	Speed, warping error
2018	Sanakoyeu et al. [SKLO18]		✓	Places365	✓		Deception rate, speed, memory
2018	Gao et al. [GGZY18]			FlyingThings3D, Monkaa	✓	✓	Speed, warping error
2023	Ioannou and Maddock [IM23]	✓		FlyingThings3D, Monkaa	✓		Warping error, depth loss, LPIPS
Multiple-Style-Per-Model (Offline Model Optimization)							
2020	Gao et al. [GLYY20]	✓		Videvo.net	✓	✓	Speed, warping error
Arbitrary-Style-Per-Model (Offline Model Optimization)							
Artistic Style Transfer							
2019	Li et al. [LLKY19]	✓	✓		✓	✓	Speed, heatmaps
2020	Wang et al. [WYXL20]	✓	✓	PBN	✓	✓	Speed, warping error
2021	Liu et al. [LLH*21]	✓	✓		✓		Speed, warping error
2021	Deng et al. [DTD*21]	✓	✓		✓	✓	Content and style loss, temporal mean & variance
2021	Liu and Zhu [LZ21]	✓	✓		✓		Depth/edge/Saliency map, SSIM, hist, aHash, dHash
2021	Chen et al. [CWZ*21]	✓	✓		✓		LPIPS
2022	Lu and Wang [LW22]	✓	✓		✓		Content and style error, D^*, speed, heatmaps, $LPIPS_{vid}$
2022	Wu et al. [WZDB22]	✓	✓		✓	✓	Speed, warping error SFID, $LPIPS_{vid}$
2023	Gu et al. [GFZ23]		✓	MPI Sintel	✓	✓	Speed, warping error $LPIPS_{vid}$
2023	Zhang et al. [ZTD*23]		✓	Places265	✓		Warping error
Photorealistic Style Transfer							
2017	Xia et al. [XXL*21]			DAVIS 2017	✓	✓	Speed, warping error, TCC

The table includes image NST methods from Table 5 that have also been trained to work for videos. Some of these approaches have conducted a User Study but there are cases in which the study does not assess the performance of stylized videos, for example [LLH*21, SKLO18, CWZ*21].

of producing stylizations of more than one style or even new style fusion effects. Other notable work that achieves multi-style transfer includes Zhang and Dana [ZD18] and the technique of Li et al. [LFY*17a] on texture synthesis.

Arbitrary-style-per-model methods

These NST methods attempt to create one model capable of transferring arbitrary styles (Tables 5 & 6). For example, the algorithm of Chen and Schmidt [CS16] encompasses a procedure ('*Style Swap*') in which the content image is replaced patch-by-patch by the style image after a set of activation patches is extracted both from content and style activations. Later, the Adaptive Instance Normalization layer (AdaIN) [HB17] was suggested:

$$AdaIN(F(I_c), F(I_s)) = \sigma(F(I_s))\left(\frac{F(I_c) - \mu(F(I_c))}{\sigma(F(I_c))}\right) + \mu(F(I_s))$$

(5)

where *F* stands for the feature activations. This allows transferring the channel-wise mean and variance feature statistics between the content and style feature activations, achieving arbitrary style transfer. Their algorithm is also significantly faster than previous approaches. Following this, a myriad of approaches have been published that propose ASPM methods, improving upon performance [GLK*17, LFY*17b, XWF*18, HJL*21, PL19, SAOM20, AXHL20] or addressing a particular use case, for example aesthetics of the results [HJL*20, HJL*23] or geometric warping [LYH21]. Approaches have also emerged that utilize meta networks [SYZ18] or more recently transformers and attention [PL19, LLH*21, DTD*22, LHYZ22, ZHW*23]. Other more recent algorithms propose the use of contrastive learning [ZTD*23], quantization [HAW*23] and image restoration [MZLB23].

Arbitrary NST has also been exploited for the generation of photorealistic stylizations [LLL*18, YUC*19, AXHL20, KLZ*23]. These works attempt to stylize an input content photo using a reference photo as a style image. The aim is to avoid distortions and visual artifacts when transferring the photographic style to the input content photo. Some of the methods included in Table 5 (e.g. [HJL*21, HAW*23]) are also capable of performing photorealistic style transfer.

2.2. Videos

The main challenge in stylizing a video sequence instead of just a single still image is achieving temporal consistency. Temporal incoherence is observed visually as flickering between consecutive frames. This results in inconsistent stylization of moving objects (or idle objects but moving camera, or both moving) across subsequent frames. Ruder et al. [RDB16] proposed the first video style transfer algorithm that is based on the stylization algorithm by Gatys et al. [GEB16]. The algorithm utilizes a temporal constraint which penalizes deviations along point trajectories in order to preserve a smooth transition between adjacent frames. Optical flow is considered and smoothed stylized videos are produced. The technique is based on the idea of initializing the style optimization algorithm for frame $i + 1$ with the previous stylized frame warped:

$$x'^{(i+1)} = \omega_i^{(i+1)}\left(x^{(i)}\right) \tag{6}$$

where $x^{(i)}$ denotes the stylized frames to be generated, and $\omega_i^{(i+1)}$ is the function that takes as input an image and warps it using the optical flow field that was calculated between frames i and $i + 1$ of the original video. A stronger consistency is achieved by additionally detecting disoccluded regions and motion boundaries. The approach was later improved and extended to spherical images and videos [RDB18], hence being applicable for Virtual Reality applications.

Subsequent studies attempt to improve the speed of computation [HWL*17, GGZY18], whilst others account for the stylization quality [SKLO18] or the depth-preserving capabilities of the trained network [LZ21, IM23]. In addition, Gao et al. [GLYY20] proposed a fast model that incorporates multiple styles. The more recent approaches, though, are capable of arbitrary style transfer to video sequences. Some of them are extensions to image stylization with additional temporal considerations [LLKY19, LLH*21, LZ21, GFZ23] whilst others are explicitly focused on video stylization [WYXL20, DTD*21, LW22, WZDB22]. A photorealistic approach to video stylization has also been suggested [XXL*21]. Table 7 summarizes the video stylization approaches and the evaluation techniques each utilized. Although not included in Table 7, it is worth mentioning that other studies have also emerged that try to stylize different object classes with a different style [KVL19] or support user controllability [JvST*19].

In addition to the evaluation techniques used in image style transfer, maintaining temporal consistency and avoiding incongruities and undesired effects is also important when evaluating video NST algorithms. This is mainly measured with the help of optical flow (Warping error), or by looking at the perceptual (LPIPS) or depth temporal change consistency (TCC) differences between consecutive frames. Qualitative evaluations and user studies are also used for the evaluation of the temporal aspect of the stylized results.

2.3. Summary

The NST literature encompasses a diverse range of techniques applied to both images and videos. However, despite the success of the field, a standardized benchmark evaluation procedure is yet to be established. Tables 1–7 show some similarities in the evaluation approaches that have been used, but they also highlight the dissimilarities and divergences in the evaluation procedures employed across NST studies.

Figure 2 shows the frequency at which each metric is utilized by NST studies. Clearly, non-expert user studies dominate. The next section focuses on evaluation techniques. NST, neural style transfer.

3. Current Evaluation Techniques

Informed by Tables 1–7, this section examines the evaluation methodologies utilized by the NST methods. Figure 3 depicts the categorization used. *Qualitative Evaluation* involves subjective judgements based on side-by-side visual comparisons provided by the author(s) of the relevant paper, *Human Evaluation Studies* collect data from expert and/or non-expert participants, and *quantitative metrics* can be divided into perceptual metrics, stylization performance, video metrics and efficiency metrics, depending on the aspect an automated metric aims to assess. evaluation methodologies and respective datasets will be examined to shed light on the inconsistencies and limitations in evaluation. Ablation studies are also included. The final section presents and compares the results of state-of-the-art NST approaches.

3.1. Datasets

As for all deep learning approaches, data used both in training and evaluation of NST methods exerts a notable influence on the results. Although for some areas of NST research (Photorealistic Image NST, Videos) there is a level of agreement regarding the data utilized for evaluation, in general, there exists a lack of consensus regarding the precise dataset to be used for all the different forms of evaluation that are conducted. Here, we attempt to investigate what
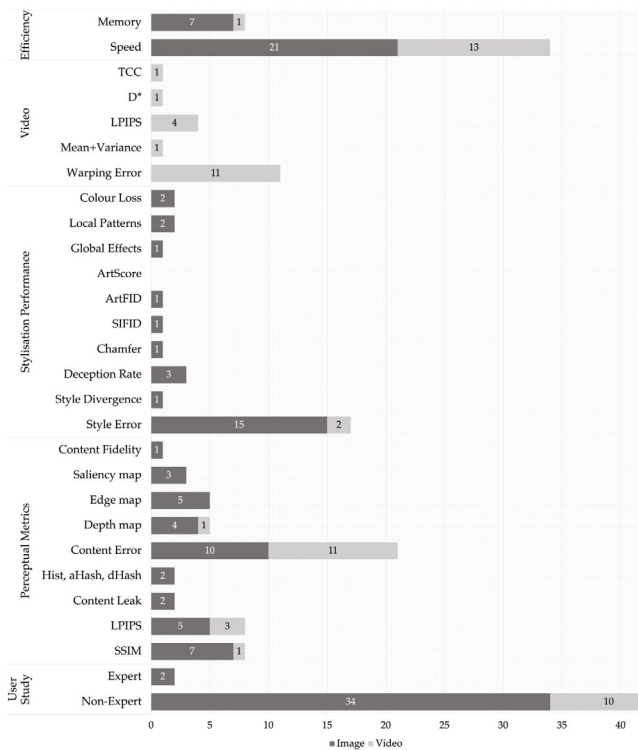
Figure 2: *The utilization of quantitative computational metrics and user studies by the NST methods. The graph shows the number of Image and Video NST methods that employed each of the quantitative metrics to evaluate their results and the amount of methods that conducted User Studies with Expert and Non-Expert participants. The graph shows that each method employs only a small selection of quantitative metrics for evaluative comparisons; some metrics are utilized more than others. At the time of this survey, the recently proposed* ArtScore *[2024] metric has not been utilized by any of the reviewed approaches. NST, neural style transfer.*

data is utilized to evaluate and present the results of the current techniques, distinguishing between Image and Video NST.

3.1.1. *Image NST evaluation data*

Gatys et al. [GEB16] mainly demonstrated their technique by performing the stylization procedure on the photograph of Tuebingen Neckarfront by Andreas Praefcke (2003). In their showcase of synthesized results, multiple well-known artworks were employed as style images, for example the shipwreck of the Minotaur by J.M.W. Turner (1805), and The Starry Night by Vincent van Gogh (1889). A similar selection of style images has been made by almost all NST researchers since then but with no particular agreement regarding the content images that are presented in the results.

Johnson et al. [JAFF16] stated that images from the MS COCO [LMB*14] validation set were used for the evaluation. For quantitative evaluation, they used 50 content images. Other PSPM (Table 3) and MSPM (Table 4) approaches have mostly considered images that were present in previous methods [ULVL16, LW16b,
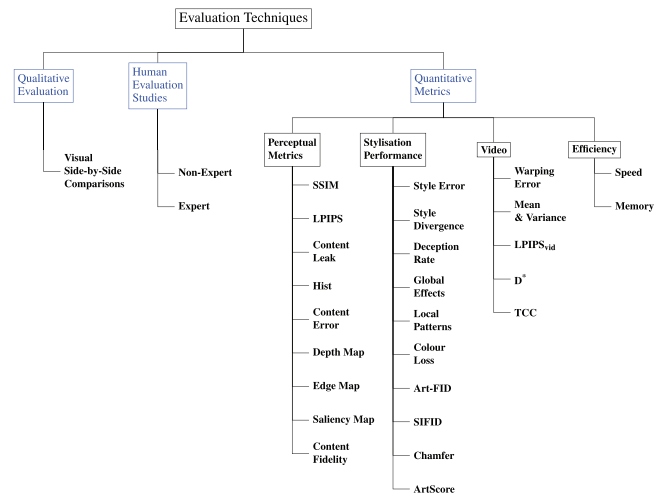

Figure 3: *Evaluation techniques utilized in the NST literature categorized into qualitative evaluation, human evaluation studies and quantitative metrics. The blue highlighting reflects the standard categorization in the literature (Figure 1). NST, neural style transfer.*

CYL*17, ZD18]. Similarly, most ASPM papers (Table 5) retrieve images from MS COCO for content and artworks from Wikiart [MK18] for style in their evaluation. Exceptions include the work of Gu et al. [GCLY18] which, in addition to content-style pairs from previous works, also used images from Ostragram [Ost], the work of Li et al. [LFY*17b] which utilized 40 images from [KTH*14], and the work of Xu et al. [XWF*18] which utilized the Behance test set [WFJ*17].

It is noteworthy that for Photorealistic Image NST (Table 1), the method of Luan et al. [LPSB17], which was the first one chronologically, published an evaluation dataset composed of 60 content images and 60 style images. Subsequent studies for Photorealistic NST utilized this dataset in their evaluation [MSZM17] and particularly for the user studies [LLL*18, YUC*19, AXHL20].

Nevertheless, for Artistic NST (Tables 1–5), the images present in both the qualitative and quantitative evaluations seem to be chosen at random with no particular agreement between the methods. Another point of disagreement is the number of stylized results (and the amount of content and styles that are used to produce them) that are utilized for all three evaluation types: *qualitative*, *user studies* and *quantitative*. The same applies to the characteristics/attributes of the content and style images used to derive the results demonstrated in the evaluation, that is portrait photographs, landscapes and lighting conditions.

An effort to establish a benchmark dataset was made by Mould and Rosin [MR17]. They proposed a set of 20 images (used as content) that represent a range of possible subject matter and image features. Their approach discusses a range of principles and image characteristics for developing a suitable benchmark. More recently, Ruta et al. [RGC*23] proposed a large-scale dataset, namely *BBST-4M*, composed of content and style images retrieved from Flickr and Behance.net, respectively. Utilizing a model capable of predicting if an image is stylistic or not, they filter out images from the two

subsets, resulting in a final dataset of 2.2 million stylistic images and 2 million content images.

3.1.2. Video NST evaluation data

The MPI Sintel [BWSB12] dataset is the most common source of data (for content videos) used in the evaluations of Video NST approaches (Table 7). As the dataset offers several video sequences and ground truth optical flow information, it is suitable for gauging the performance of the models in terms of temporal coherence and smoothness. For style images, there is no common practice, yet there is a presence of well-known artworks in the results. In some cases, frames from real-world videos (Videvo.net [Vid19]) or computer games (GTA [RVRK16]) are also used [HWL*17, LLKY19, GLYY20, IM23].

The work of Xia et al. [XXL*21] on Photorealistic Video NST (Table 7) utilized the DAVIS 2017 [PTPC*17] validation dataset for evaluation. In contrast to the abundance of methods available for Photorealistic NST on still images, there is a scarcity of adequate approaches within the domain of Photorealistic Video NST. Consequently, it becomes challenging to agree on a standard dataset for their evaluation purposes.

3.2. Qualitative evaluation

Qualitative evaluation in NST involves the visual inspection of the synthesized results. Researchers attempt to examine the generated images and derive judgements on their aesthetic quality and their ability to transfer the style efficiently while maintaining the contents.

Typically, qualitative evaluation is performed by demonstrating the results of the suggested method side-by-side with the results of similar or state-of-the-art approaches. The stylized images are also presented with the content and style images that were used to synthesize the results. Arguably the visual side-by-side comparison provided by all the NST approaches is not free from subjectivity. The authors attempt to elevate their results by comparing them to the results of state-of-the-art methods, commenting on visible artifacts of other methods and highlighting the effectiveness of their technique.

Image quality, perceptual quality and aesthetic appeal are some of the aspects that a standard qualitative evaluation addresses. Although the exact wording differs among the studies, a common practice has been the characterization of the results with comparative terms such as 'better', 'comparable', 'competitive' and 'slightly behind' when referring to the quality of the stylizations [JAFF16, UVL17a, HB17]. Often, zoomed-in cut-outs from the stylized outputs are accompanied by comments that highlight the superiority of the suggested method in preserving fine-grained details or in applying stylization more effectively in detailed structures compared to the failure to do so of the methods in comparison [SKLO18, HJL*20].

In other cases, the particular problem an approach attempts to solve is pointed out through these visual side-by-side comparisons. For example, structure-preserving [CLW*19, LZ21] and depth-
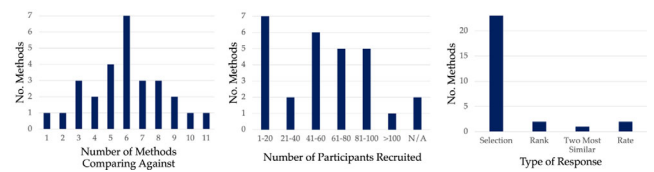

Figure 4: *The variation in methods comparing against, number of participants recruited and response format in user studies of artistic image NST approaches. 'N/A' in the second graph denotes that a study has not provided the relevant information. NST, neural style transfer.*

aware methods [LCLR17, IM22] use the illustrations to emphasize the capabilities of the technique in retaining the 3D spatial layout and local structures of the content images, as well as the distance relationships between the different objects in the scene and the background-foreground contrast.

For Video NST, consecutive stylized frames are visually inspected. Qualitative evaluation is employed to visually identify incongruities and undesired artifacts and highlight how the proposed approach facilitates temporal stability.

As stressed before, a significant point of disagreement lies in the selection of the style and content images that are used to generate the evaluation results. Every study showcases a diverse array of comparisons and due to the absence of a universally agreed-upon set of content and style images, it is possible that each method selectively employs specific content and style images that could potentially grant an advantage to their respective system. Examples of the most common content and style images used in qualitative evaluations are included in the supplementary material. Despite their extensive utilization, a standardized and universally accepted dataset has not been established. We argue that NST research and the community would benefit if studies used the same content and style images to derive their results for their qualitative comparisons. Such a practice could result in fairer comparisons and important insights for a better understanding of each method and its efficiency and quality.

3.3. Human evaluation studies

Due to the subjectivity in assessing stylizations, NST researchers commonly use Human Evaluation Studies to assess the performance of artistic/photorealistic image/video generation methods. As depicted in Tables 1–7, the majority of Image and Video NST approaches opt to conduct user studies to verify the effectiveness of the proposed system and further compare it with baselines. Despite the similarities in the way user studies are conducted, disparities arise in their design and formulation and subsequently in the presentation of the results. A review of the NST papers reveals an array of factors that methods approach in varied ways, as summarized in Table 8.

To give an estimation of the variation in the settings of the user studies conducted in NST, Figure 4 shows three graphs depicting the state of the literature. The graphs demonstrate how many methods (y-axis) choose to compare against different numbers of other state-of-the-art methods (Graph 1), the varied number of participants recruited by each NST study (Graph 2), and the different response

Table 8: *The different factors in human evaluation studies.*

Factor	Description
Number of participants	The number of participants, the number of responses collected per participant and/or the total votes.
Participants' profiles	The age and sex of the participants, and whether the users are experts.
Data	The number of content and style images/videos used to derive the results shown to the participants and if samples are drawn from a larger evaluation set of content-style pairs (and how many).
Presentation	Whether the content and the style images are revealed in each question and whether all the stylizations of the methods in comparison are shown in each question or only two of them (the proposed method and one other selected randomly).
Number of methods	The number of other methods that the proposed approach is compared against.
Number of questions	The overall number of questions and the number of questions asked to each participant.
Question formulation	What the participants are actually asked to answer and the exact factors they are asked to base their selection on.

formats that are followed (Graph 3). Although these graphs cannot be regarded as indicative of the whole NST literature (they only represent the artistic Image NST methods reviewed in this paper), they provide a good approximation of the differences in the practices followed when designing and implementing user studies to assess stylized imagery.

3.3.1. Number of participants

The number of participants recruited in the reviewed papers (Tables 1–6) varies from 10 to 220 people, even though on some occasions the number of subjects involved is omitted and only the number of votes is reported [AHS*21, AXHL20]. In addition, it is essential to distinguish between expert (e.g. Art Historians [SKLO18, KSLO19]) and non-expert participants. Sporadically, the age and sex of participants are provided [CLW*19, DTD*22, GFZ23].

3.3.2. Data generation/utilization

Different practices are followed for the images/videos that are utilized in the user studies. Typically, results are generated for an amount of content-style pairs, and for each method in comparison. In some cases, the number of content-style pairs chosen is not too large, and all the images are revealed to each participant. In other cases, a sample is drawn at random from the synthesized images, for example, 20 content-style pairs sampled from the total of 225 generated results (from 15 content to 15 style images) [LLH*21]. Another practice is to divide the repository of content-style pairs with the generated results into groups (e.g. a total of 240 outputs – from 12 content and 20 style images – divided into five groups, with each group containing 100 pairs [HJL*20]) of which one is shown to each participant.

3.3.3. Number of methods in comparison

Depending on the particular use case and applicability of the proposed technique, or the specific problem it aims to solve, the number of methods that it is compared against is also varied. For Image NST, it can be from 1 to 11 approaches that are considered as baselines. The number of questions asked to each participant and the overall questions asked is another important parameter.

3.3.4. Task presentation and question formulation

An essential part of the user study is the task presentation, and how each question is shown to the participant. This element in the design of the user study is varied amongst the NST research mainly by the following two factors: (a) Number of synthesized images presented side-by-side, and (b) Content and style images display. Multiple authors choose to compare the results of their approach with only one baseline each time [WZZ*22, TLL*23, HJL*23], a design also referred to as *two-alternative forced choice (2AFC), or A/B testing* [KLSH09]. In most cases, though, participants are presented with three or more options simultaneously – one output from each of the in-comparison methods. A few authors decided to reveal each generation result individually, asking for a rating [LPSB17, HJL*21]. Another important consideration in NST user surveys is the decision to reveal to the user the content and style images that were used to derive the results. Even for a small group sample (20), it has been demonstrated that omitting the content and style images significantly influences the responses of the participants [IM22].

The presentation of the questions is closely related to the formulation of the questions. The way the question is phrased and the type of response expected are impactful to the results. In most of the studies, the participants are asked to select their 'favourite' stylization (Figure 4, Graph 3). Examples of wordings used include 'most preferable', 'favourite' and 'the best' [LFY*17b, XWF*18, PL19, KKM19, AHS*21]. Occasionally, the participants are instructed to take into account indicators such as content presentation, style quality and overall quality [LHYZ22, GFZ23, ZHW*23]. To reflect the particular contribution of the technique and the problem that is addressed, questions are formulated distinctively in some cases: 'Choose an image that best and most realistically reflects the style' [SKLO18], or "Identify the patch to be real artwork from a stylized image" [KSLO19].

Measuring the 'realism' of fake paintings (stylizations) in comparison to human-created paintings has been in general a practice used in user surveys. Studies ask participants if the shown image is a real artwork or not, reporting the results as 'Deception score' [CWZ*21, WZZ*22]. A similar question ('Select the real artwork from a pair of a real artwork and a stylized image') was used by Huang et al. [HAW*23]. Zhang et al. [ZTD*23] designed a user study in which participants were shown 10 artworks of which 2–4 were synthetically generated by the same method (one of the methods in comparison). The participants were asked to select the synthetically generated images, with precision and recall reported for each method.

Hu et al. [HJL*20] asked the participants to vote for three results that are most similar to the style image in terms of colour, texture

and overall feeling. Another practice is to use three or four questions per task whereby the user is asked to answer each question individually, with each question considering a different aspect of the stylization, such as content preservation/integrity, style quality/level and overall quality. In rarer cases, user studies collect ratings for the aforementioned aspects [HJL*21] or they ask the users to rank the displayed images in order from 'Best' to 'Worst' [JAFF16].

3.3.5. *Photorealistic and video NST user studies*

The user surveys in Photorealistic Image NST seem to follow a more standardized procedure. Put forward by the work of Luan et al. [LPSB17], who also provided the most predominantly used test dataset, a study to evaluate the results of photorealistic approaches is two-fold. Firstly, the authors attempt to evaluate the realism of the results by asking the participants to rate a stylized image from 'Definitely photorealistic' to 'Definitely not photorealistic'. Secondly, the participants are asked to choose the output image that is more similar to the presented style image. Although not identical, similar task presentation and question phrasing are followed by subsequent Photorealistic NST methods [MSZM17, LLL*18, YUC*19]. An et al. [AXHL20] opted for a simpler approach, as the participants were only asked to choose the best image in terms of less artifact, less distortion and more details.

The user studies in Video NST generally follow the same design as in Image NST, with some obvious differences. First, instead of images, the participants are shown content and stylized videos. Additionally, the question formulation is revised to encompass a consideration for temporal coherence. More precisely, in addition to the resemblance to the style image [RDB18, GGZY18, GLYY20], participants are asked to select the synthesized video that is the most temporally stable [LLKY19, GGZY18, WYXL20, DTD*21, WZDB22] and which avoids visual artifacts and flickering effects [HWL*17, RDB18, GLYY20].

3.3.6. *Other considerations*

For both Image and Video NST, a multitude of factors regarding the design and reporting of user studies exist that can potentially influence the responses of the participants and the concluded outcomes [BHH*23]. Infrequently, the authors undertake measures to facilitate the attainment of more reliable results. For example, Mechrez et al. [MSZM17], when testing for realism, validated the responses of the participants by asking them to rate the original (realistic) image; if an average rating was less than a threshold, they excluded the responses of the respective participants to filter out poor-quality data. Similarly, Xu et al. [XWF*18] ruled out untrustworthy results labelled too soon.

Bylinskii et al. [BHH*23] provide a detailed and thorough evaluation of the current trends regarding the design and reporting of user studies in computer graphics and vision, combined with robust recommendations on how to ameliorate inconsistencies and misleading results. Other factors that researchers might need to take into account are:

- The means used to conduct the survey: Online questionnaire or face-to-face in a controlled lab environment?

- The image quality and resolution of the presented images.
- The screen specifications or the quality of the screen each participant has viewed to complete the survey.
- The exact placement of the images on the screen: Are all the images placed side-by-side, or are there images placed below or on top of others? Where are the content and style images placed relative to the results?

Given the discrepancies in the design of user studies in NST, it is important for future studies to establish a standardized human evaluation approach. This would allow for more accurate and reliable comparisons between different NST methods and their performance. Additionally, it would provide a clearer understanding of the potential of NST in various applications. Further analysis and recommendations regarding the user studies in NST are provided in Section 4.

3.4. Quantitative evaluation metrics

Depending on the problem a method attempts to solve, only a small selection of quantitative metrics are employed for evaluative comparisons (Figure 2). We classify the metrics into four distinct categories depending on the particular aspect of the stylization performance they quantify:

- **Perceptual metrics**: Metrics that mostly focus on the content and structure preservation performance.
- **Stylization performance**: Metrics that assess how well the stylized image resembles the style image.
- **Video**: Metrics that gauge temporal coherence and video stability performance.
- **Efficiency**: Metrics for benchmarking characteristics regarding the system's performance.

3.4.1. *Perceptual metrics*

To assess the content preservation performance of NST models, various widely used perceptual metrics are utilized. Typically, the perceptual metrics in computer vision research measure the similarity between two images. In NST, these metrics are computed between the original content images and the stylized results. Most of the metrics do not simply compare pixel values; instead, the majority of those take into account higher-level features such as edges, textures and colours.

One of the most frequently adopted metrics is the Structural Similarity Index (SSIM) [WBSS04] defined as:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (7)$$

where x and y are the two images being compared, μ_x and μ_y are the pixel sample means of the two images, σ_x^2 and σ_y^2 are the variance of x and y respectively, σ_{xy} is the covariance of x and y and c_1, c_2 are constants used to prevent the denominator from becoming zero. SSIM is based on the degradation of structural information and for its computation it takes into account luminance, contrast and structure [WBSS04]. Although SSIM can provide valuable insights,

it can be limited in modelling the human visual system, especially when working with colour images [NAM20].

With the advent of deep learning, other perceptual metrics have been proposed (such as LPIPS, Content Error and CF that are considered to be more compatible with the human visual system and capable of more effectively capturing the perceptual characterisitcs of images. The Learned Perceptual Similarity metric (LPIPS) works by extracting and comparing features from pre-trained neural networks, shown to model low-level perceptual similarity particularly well [ZIE*18]. As demonstrated by Zhang et al. [ZIE*18], LPIPS outperforms SSIM when evaluated on their proposed BAPPS dataset.

Another commonly used metric is Content Error [GEB16], which computes the mean square error of feature activations between the stylization and the content image, similar to the content loss function employed during the training of most NST methods. Unlike LPIPS, Content Error utilizes features from a pre-trained VGG in object recognition without exposing it to distorted images. Zhang et al. [ZIE*18] shows that training a network on distorted image patches contributes to capturing perceptual behaviour better.

CF was suggested as an improvement to the Content Error by Wang et al. [WZC*21]. CF can measure the faithfulness of the stylized result (y) to the original content image (x) at multiple scales. Unlike Content Error, which works in Euclidean space, CF utilizes cosine similarity to measure the differences in deep feature activations (f_l):

$$CF(x, y) = \frac{1}{N} \sum_{l=1}^{N} \frac{f_l(x) \cdot f_l(y)}{\|f_l(x)\| \cdot \|f_l(y)\|} \qquad (8)$$

where N is the number of different layers.

Contrary to the rest of the computational metrics that calculate an exact numerical value, Content Leak is observed by only counting the number of stylization rounds required to reach loss of content information. Proposed by An et al. [AHS*21], Content Leak is used to analyze the effectiveness of the NST systems in retaining content information. To gauge the amount of Content Leak, a content-style pair is used to perform a stylization. Then, using the stylized result as the new content image, the style transfer process is performed repeatedly multiple times. It is then easy to notice whether content information is lost/retained in the resulting outputs.

Depending on the specific problem addressed by an NST method, the evaluation of perceptual similarity may diverge from common metrics such as LPIPS and SSIM, with alternate metrics being employed. For example, approaches that are more focused on retaining global structure and depth information [LCLR17, CLW*19, IM22] use histogram and histogram-based methods that consider intensity and tone information within the image. The histogram is a depiction of the distribution of pixel values in the image. It is useful for detecting tonal and colour differences between the content image and the stylized result. Similarly, the image hash algorithms [Buc21] – average hash (aHash) and difference hash (dHash) – analyze the image structure on luminance and are suitable for identifying similarities in the input images. These depth and structure-aware approaches [LCLR17, CLW*19, IM22] also attempt to quantify the depth and content preservation capabilities of their model

using depth, edge and saliency map comparisons. This information is inferred using state-of-the-art approaches (e.g. depth prediction [CFYD16, RLH*20]; edge detection [XT15, LCH*17]). Saliency detection, which is considered an instance of image segmentation, can also be computed with models such as the one by Jiang et al. [JWY*13]. Comparisons between the derived depth/edge/saliency map of the content image and the corresponding map of the stylization image are performed using standard similarity measures, such as mean square error or SSIM.

3.4.2. *Stylization performance*

Assessing how well the stylized result resembles the reference style images is a challenging task, and maybe a subjective one. Although multiple metrics have been proposed for the evaluation of the stylization performance of NST methods, no metric is currently used as the gold standard. The stylization performance metrics compare the outputs of the NST methods with the corresponding style images that are used to generate them.

While a few approaches involve training models to assess how well the style is reproduced in the synthesized result, most metrics rely on features or other information (e.g. histogram) extracted from the images in comparison. For example, Style Error [GEB16] (analogous to the Content Error used to measure content preservation) uses features extracted from a pre-trained VGG and is defined identically to the style loss used in training – the mean square error of Gram matrices between the feature activations of the style image and the stylized result.

To improve upon the Style Error, Wang et al. [WZC*21] proposed two effects to assess stylization performance in addition to the CF CF defined in Section 3.4.1. The Global Effects (GE) metric combines Global colours (GC) that measure colour histogram differences ($hist_c$):

$$GC(y, s) = \frac{1}{3} \sum_{c=1}^{3} \frac{hist_c(y) \cdot hist_c(s)}{\|hist_c(y)\| \cdot \|hist_c(s)\|} \qquad (9)$$

where y and s are the stylized result and style image respectively; and Holistic Textures (HT) which is similar to the style error measurement using Gram matrices G_l, but uses cosine similarity:

$$HT(y, s) = \frac{1}{N} \sum_{l=1}^{N} \frac{G_l(y) \cdot G_l(s)}{\|G_l(y)\| \cdot \|G_l(s)\|} \qquad (10)$$

GE is then defined as: $GE(y, s) = \frac{1}{2}(GC(y, s) + HT(y, s))$. A different factor was also suggested that measures the quality of local style patterns. The Local Patterns (LP) factor is defined as $LP(y, s) = \frac{1}{2}(LP_1(y, s) + LP_2(y, s))$, with LP_1 measuring differences of LP counterparts directly:

$$LP_1(y, s) = \frac{1}{Z} \sum_{l=1}^{N} \sum_{i=1}^{n_y} \frac{\phi_i^l(y) \cdot \phi_{CM(i)}^l(s)}{\|\phi_i^l(y)\| \cdot \|\phi_{CM(i)}^l(s)\|} \qquad (11)$$

where $\phi_i^l(y)$ and $\phi_j^l(s)$ are used to denote the neural patches for multi-scale features, $CM(i) := \arg \max_{j=1,...,n_s} \frac{\phi_j^l(y) \cdot \phi_j^l(s)}{\|\phi_j^l(y)\| \cdot \|\phi_j^l(s)\|}$, and LP_2

compares the diversity of the pattern categories:

$$LP_2(y, s) = \frac{1}{N} \sum_{l=1}^{N} \frac{t_{CM}^l}{t_s^l} \qquad (12)$$

where t_{cm}^l and t_s^l are the numbers of $\phi_{CM(i)}^l(s)$ and $\phi_j^l(s)$, respectively.

Other metrics that only rely on information from the stylized image and the style image include the metrics employed by Ruta et al. [RGC*23] and Gu et al. [GFZ23]. Ruta et al. [RGC*23] uses Chamfer distance to measure colour consistency. Here, similarly to CF [WZC*21], which is based on colour histograms, Chamfer distance is used to measure dissimilarities between the stylization and the reference style image in image colour space. Formally, Chamfer distance calculates the minimum distance between each point in one set X (of pixels) and the nearest point in the other set Y (of pixels), and sums up these distances:

$$CD(X, Y) = \sum_{x \in X} \min_{y \in Y} ||x - y||_2^2 + \sum_{y \in Y} \min_{x \in X} ||x - y||_2^2 \qquad (13)$$

To assess colour differences, Gu et al. [GFZ23] employed the colour loss proposed in [IKT*17]. This computes the Euclidean distance between the (Gaussian) blurred versions of the style image (s_b) and the stylized result (y_b):

$$\mathcal{L}_{\text{colour}}(s, y) = ||s_b - y_b||_2^2 \qquad (14)$$

More sophisticated approaches propose training models on paintings to gauge stylization performance. Kotovenko et al. [KSLO19] proposed a method for content and style disentanglement. As part of their evaluation, they employed a statistical distance metric to measure how well their system is capable of covering the style distribution it attempts to reproduce. Using a trained network on painting classification, they extract activations on real artworks to define the true style distribution $\mathbb{P}_s^{art}$, and activations from the stylizations derived from the proposed style transfer model to approximate $\mathbb{P}_s^{stylized}$. The Kullback-Leibler Divergence, depicting how well the style distribution is represented, is then computed as:

$$D_{KL}\left(\mathbb{P}_s^{stylized}, \mathbb{P}_s^{art}\right) \qquad (15)$$

Style Transfer Deception rate, suggested by Sanakoyeu et al. [SKLO18], is based on a similar idea. A *VGG-16* network trained to classify artists from the Wikiart dataset, is used to classify multiple stylizations. The deception rate is then calculated as the fraction of synthesized images classified to belong to the artist whose style they attempted to replicate.

Similarly to the D_{KL} metric proposed in [KSLO19], that is related to the idea of measuring the style performance in accordance to how well the image distribution is captured, Fréchet Inception Rate (FID) [HRU*17] measures the deviation between the distribution of deep features of generated images and that of real images. FID is commonly used for the evaluation of GAN methods but it can also be suitable for evaluating stylizations by adjusting it to work for a single image and internal patch statistics (SIFID) [SDM19]. This can thus be utilized to measure style consistency between the style images and the stylized results [RGC*23].

More recently, an enhancement to the FID metric has been proposed that attempts to evaluate both content preservation and style matching [WO22]. The *ArtFID* metric uses the Inception network trained on a large-scale art classification dataset and computes feature distribution differences between image features extracted from the stylizations and image features extracted from the style images. The distance between the two feature distributions is calculated using the Fréchet distance. *ArtFID* is thus formulated as:

$$\text{ArtFID}(X_g, X_c, X_s) = \left(1 + \frac{1}{N} \sum_{i=1}^{N} d(X_c^{(i)}, X_g^{(i)})\right)$$
$$\cdot \left(1 + \text{FID}(X_s, X_g)\right) \qquad (16)$$

where X_g, X_c and X_s are the stylized images, the content images and the style images respectively, $d(X_c^{(i)}, X_g^{(i)})$ is measured using the LPIPS metric to account for content preservation and FID is defined as:

$$\text{FID}(X_s, X_g) = ||\mu_s - \mu_g||_2^2 + \text{Tr}\left(\Sigma_s + \Sigma_g - 2(\Sigma_s \Sigma_g)^{\frac{1}{2}}\right) \qquad (17)$$

where (μ_s, Σ_s) and (μ_g, Σ_g) correspond to the mean and covariance of the extracted Inception features of the style images X_s and stylized images X_g respectively, and Tr refers to the trace linear algebra operation. *ArtFID* is shown to be compatible with human judgement [WO22].

Another metric that is also shown to strongly coincide with human judgement is *ArtScore* [CALL23]. This measures how well a synthesized image resembles an authentic artwork. The framework proposed by Chen et al. [CALL23] applies transfer learning using StyleGAN models [KLA*20], employs interpolation image generation and trains a neural network (*ArtScore*) with ResNet-50 pretrained on ImageNet as the backbone and with an effective learn-to-rank objective.

3.4.3. *Video*

Except for the perceptual and style performance metrics, methods suitable for stylizing video sequences are required to quantify the effectiveness of their system in achieving temporal stability. As most of the approaches work in real-time – generating a fast stylization for a single frame through a forward pass – the derived stylizations are prone to flickering effects and undesired instabilities. The most reliable way to quantify temporal consistency is the warping error which calculates the difference between a warped next frame and a ground truth next frame [LHW*18].

$$WarpingError = \sqrt{\frac{1}{T-1} \sum_{t=1}^{T} ||O_t - W_t(O_{t-1})||^2} \qquad (18)$$

where O_t and O_{t-1} are the corresponding output stylized frames of the current and previous input frames, W_t is the ground truth optical flow and T is the number of time steps or the number of frames in a sequence. Methods that do not evaluate videos with available optic flow masks might compute those using state-of-the-art approaches, such as *FlowNet* [DFI*15]. Some approaches may also include the ground truth occlusion mask in the computation to account only for traceable pixels.

More straightforward approaches to measuring temporal stability include calculating the LPIPS (Section 3.4.1) between adjacent frames or the mean and variance of the subsequent video frames. LPIPS provides a way to compute the average perceptual distances between adjacent frames. As discussed in [ZIE*18], LPIPS performs better in computing perceptual differences in images compared to low-level approaches (e.g. MSE, SSIM), nevertheless, these have been used to quantify differences between subsequent frames. Deng et al. [DTD*21] chose to quantify temporal smoothness by defining $Diff_{F(t)} = ||F_t - F_{t-1}||$ and calculating the mean and variance of $Diff_{F(t)}$, where F_t and F_{t-1} are two adjacent frames of a T-frame rendered video. Similarly, Lu and Wang [LW22] compute the average pixel distances using D^* defined as:

$$D^* = \frac{1}{T-1} \sum_{t=0}^{T-2} ||F_{t+1} - F_t||_2 \tag{19}$$

Except for the different notation used in [DTD*21] and [LW22], Deng et al. [DTD*21] also reports the computed variance in addition to the mean error.

Similarly to image style transfer methods focused on preserving depth or structural information [LPSB17, CLW*19, IM22] utilizing depth data to gauge perceptual similarity performance, Xia et al. [XXL*21] also use depth maps for their evaluation. To quantify the temporal stability performance of their Photorealistic Video style transfer method, they employ a Temporal Change Consistency metric (TCC), defined as:

$$\mathrm{TCC}(D, G) = \frac{\sum_{i=1}^{n-1} SSIM(abs(d^i - d^{i+1}), abs(g^i - g^{i+1}))}{n-1} \tag{20}$$

where $D = (d^1, d^2, \ldots, d^n)$, and $G = (g^1, g^2, \ldots, g^n)$ are the estimated depth maps of the stylized frames, and the corresponding ground truth depth maps, respectively. Here, TCC uses SSIM [WBSS04] to compute the differences between consecutive depth maps.

3.4.4. *Efficiency*

Another essential aspect of the quantification of the performance of a stylization algorithm is efficiency. This basically accounts for speed, memory and control. NST algorithms compare the time their model needs to output a stylized image through a forward pass given an input content image (e.g. [RWB17, JAFF16, DTD*22, LLH*21, AHS*21]). Occasionally, the memory a trained model requires is compared to state-of-the-art models, especially when approaches are focused on improving the stylization network and its composition (e.g. [LW16b, SKLO18, ZD18, SYZ18]). Another consideration has to do with user control (e.g. allowing users to stylize specific areas of an image differently). Unless there is an obvious difference in the number of styles a model can reproduce (e.g. PSPM vs. ASPM), the number of styles a method is capable of emulating is also compared [GLYY20].

3.5. Ablation studies

Inspired by the field of neuroscience, where ablation studies have been used to unveil and analyze more precisely the structure and
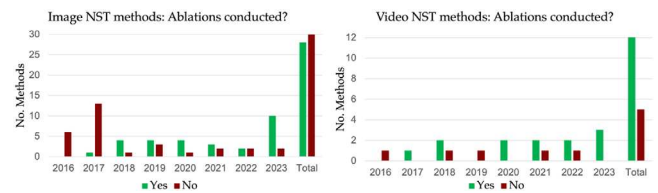

Figure 5: *The number of Image and Video NST studies including Ablation Studies. The graph depicts the reviewed papers of the NST literature. NST, neural style transfer.*

organization of the human brain, artificial intelligence, machine learning and subsequently computer vision research has resorted to the concept of removing (or substituting) specific components from artificial neural networks to effectively analyze their behaviour [MLdPM19]. NST systems have also used this idea, which is capable of deriving useful insights and understanding, resulting in an in-depth evaluation of the performance of the proposed models. Ablations cannot be explicitly regarded as qualitative or quantitative evaluation techniques, as they can be used to produce both types of data – for the examination of particular components or aspects of an approach, qualitative results (e.g. stylized images/videos) can be provided, or the computational metrics can be re-run for a quantitative assessment.

Figure 5 provides an overview of the involvement of ablation studies in Image and Video NST evaluation approaches. More Video NST studies include ablations than not, based on the papers reviewed in this work (Table 7).

A common practice is the study of the effect of the different losses the NST methods employ during training [SKLO18, KSLO19, GCLY18, LLH*21, PL19, LHYZ22]. Yet, as the design and implementation of NST models can be widely diverse, the ablation studies employed assess different aspects in different methods. For example, Deng et al. [DTD*21] performed ablations to analyze the effect of the content-aware positional encoding (CAPE) method proposed, whereas Liu et al. [LLH*21] provided qualitative results to verify the effectiveness of the shallow feature used in their suggested method (*AdaAttN*). Ablation studies can be crucial for enhancing interpretability and understanding.

3.6. Experiments

In this section, to show an example of a current evaluation procedure and reveal some of the potential issues that arise, we experimentally compare nine state-of-the-art methods: *AdaIN* [HB17], *AdaAttN* [LLH*21], *ArtFlow* [AHS*21], *CS-BNet* [LW22], *IEContraAST* [CWZ*21], *MCCNet* [DTD*21], *RAST* [MZLB23], *SANet* [PL19] and *StyTr²* [DTD*22]. For AdaIN and AdaAttN, we used the PyTorch implementations available at https://github.com/naoto0804/pytorch-AdaIN and https://github.com/Huage001/AdaAttN , respectively, whereas, for the rest of the approaches, their official code implementations were used. We utilize the dataset proposed by Mould and Rosin [MR17] comprised of 20 content images, and 10 style images that are frequently employed by NST methods and which encompass a diverse range of stylistic attributes (e.g. genre, colour, texture). Example results are

Figure 6: *Results of state-of-the-art NST approaches. More results are provided in the supplementary material. NST, neural style transfer.*

Table 9: *Quantitative comparisons between state-of-the-art NST approaches. $\mathcal{L}_c$ and $\mathcal{L}_s$ account for Content and Style Loss (Error) respectively.*

Method	$\mathcal{L}_c \downarrow$	SSIM $\uparrow$	LPIPS $\downarrow$	$\mathcal{L}_s \downarrow$	SIFID $\downarrow$
AdaIN	1.6170	0.2588	0.5174	**0.1023**	**0.6223**
AdaAttN	1.2542	0.4705	0.4766	<u>0.1049</u>	1.3530
ArtFlow	<u>0.9575</u>	0.4547	0.4824	0.1079	0.9657
CSBNet	1.2494	0.3601	0.4780	0.1141	5.0953
IEContraAST	1.3530	0.4065	<u>0.4439</u>	0.1064	<u>0.8600</u>
MCCNet	1.1193	0.4547	0.4714	0.1091	2.3755
RAST	**0.9272**	**0.5383**	**0.3101**	0.1113	1.7580
SANet	1.4260	0.3096	0.5373	0.1073	1.1592
StyTr^2	1.0415	<u>0.4882</u>	0.4693	0.1073	1.1089

The best results are in bold and the second-best results are marked with an underline.

4. Analysis and Recommendations

This section provides analysis and recommendations for the different aspects of evaluation: benchmark datasets, qualitative evaluation, human evaluation studies and quantitative evaluation. The aim of the suggested recommendations is to establish an evaluation benchmark and standard practices to be used universally across all approaches. This would benefit the NST community, and the results reported by each NST method would be more meaningful and allow for direct comparisons. We argue that the quantitative metrics can provide useful insights but a deeper understanding is required to support their utilization.

4.1. Benchmark datasets

As discussed in Section 3.1, for Video NST there is some agreement as to what data is used in the evaluation. Similarly, in the domain of photorealistic Image NST, there is some consensus on the dataset employed. However, there is no consensus within the domain of artistic Image NST.

Only a few approaches have focused on addressing the lack of a benchmark dataset in NST research. The method of Mould and Rosin [MR17] suggested a set of content images that fulfil a set of criteria and can potentially be utilized as a benchmark for evaluation. However, they don't consider style images and the potential characteristics that a benchmark style image set should include. The recent approach of Ruta et al. [RGC*23] proposed a large-scale dataset for style transfer, containing both content and style images. The suggested *BBST-4M* dataset categorizes images into style and content based on how stylistic they are. While this dataset holds significant potential, it falls short in comparison to the work of Mould and Rosin [MR17] in terms of considering the distinct characteristics and attributes that should be encompassed by both the content image set and the style image set. The difficulty in composing a suitable evaluation dataset can be attributed to the existence of various types of evaluation and to the diverse intentions and goals of the proposed stylization solutions.

The following subsections discuss the research questions that should be considered when compiling a sufficient evaluation dataset, which are:

shown in Figure 6. Although the techniques differ, the stylizations produced are quite similar. Based on the visual side-by-side comparisons alone, it is difficult to quantify the performance of the methods and to estimate which approach preserves the content information better, or which approach generates results that capture the style image more accurately.

Table 9 gives quantitative comparisons for the state-of-the-art methods. Three different computational metrics are used to capture the content preservation performance while two different computational metrics are employed to quantify style resemblance. *RAST* generates consistently more content-aware results that preserve the global structure of the input content images. However, for the rest of the methods, there is a lot of variability in their performance across the different metrics. For example, *ArtFlow* performs better than *StyTr^2* in Content Error ($\mathcal{L}_c$), but it falls behind in SSIM. Similarly, when quantifying the style performance of the algorithms, *IEContraAST* achieves better results than *AdaAttN* when considering SIFID, but *AdaAttN* outperforms *IEContraAST* when measuring Style Error ($\mathcal{L}_s$). Additionally, while *AdaAttN* performs competently to the rest of the approaches on the Style Error metric (second best), its efficacy drops significantly on the SIFID metric (sixth best).

Although quantitative computational metrics can offer valuable insights into the efficiency of the algorithms, their application varies across NST methods, leading to inconsistencies in evaluation. The next section provides an in-depth analysis of the considerations arising from the utilization of the various evaluation techniques and proposes ways to alleviate them.

- What are the properties or attributes the content image set and style image set should include? In accordance with this, does the dataset distribution cover a sufficient amount of the identified image characteristics?

- Can an NST approach utilize the dataset effectively to demonstrate a new solution to an existing or new problem? Does the evaluation benchmark consider a sufficient amount of possible use cases, covering a wide range of potential problems and applicability of NST?
- Is the evaluation benchmark suitable for both qualitative and quantitative evaluation, including human evaluation studies? Or, should the dataset be divided into three partitions, with each image set comprising the relevant images based on the specific type of evaluation? What is the optimal size of the dataset?

4.1.1. *Content and style benchmark datasets*

Mould and Rosin [MR17] discussed a range of policies for selecting the proposed image set. Although they only provide a small set of content images, the principles with which the set of images is composed are worth revisiting. Their suggested list of image properties includes colourfulness, complexity, contrast, sharpness, lineness, mean, standard deviation and noise. For these, a numerical value can be derived using computational methods, thus enabling an evaluation of how broad the distribution of a set of images is and the range of properties it covers. A list of image properties, explicitly suited for designing the benchmark set of content images is also provided – such consideration of representative coverage is a general challenge for visual datasets in computer vision. This includes variation in scale, fine detail, variation in texture, regular structure, vivid and varied colours, muted colours, thin features, human faces and more. Independently from the discussion on the size of the evaluation benchmark and the number of content images and style images that should be utilized in each stage of the evaluation process, the set of 20 images (*NPGeneral*) presented by Mould and Rosin [MR17] could serve as a promising initial foundation and offer valuable insights for the design of a comprehensive evaluation benchmark dataset.

Although there have not been any efforts to establish a benchmark video dataset, the literature suggests a convergence on utilizing videos from the MPI Sintel dataset for evaluation (Section 3.1.2). This dataset is useful as it also contains ground truth optical flows for measuring the warping error and gauging the temporal coherence performance of the methods in comparison. Nevertheless, its synthetic nature does not allow for a comprehensive and detailed examination of the results, as the performance might degrade when tested on real-world videos [MC22]. We argue that a similar effort to the approach of Mould and Rosin [MR17] should be made in order to develop a video dataset that encompasses a wide variety of characteristics. Videos could be retrieved from multiple sources, combining both synthetic and real-world scenes. Establishing such a benchmark dataset for Video NST approaches could result in more robust comparisons and insightful evaluations.

Large-scale art datasets (e.g. Wikiart) can greatly facilitate the creation of a well-suited set of style images. An evident evaluation dataset design decision would be to compose the style images set by incorporating a diverse range of artworks encompassing various artistic genres, for example as defined in Wikiart. It would therefore be necessary to provide an analysis and discussion regarding the definition of an artistic genre and its interrelation

with style. As defined in *Tate* [Tat23], genres are types of painting codified in the 17th century as history, portrait, genre painting (scenes of everyday life), landscape and still life. Developed in European culture, the genre system is not particularly relevant to contemporary art but is a system that can be used to divide artworks according to depicted themes and objects. The Wikiart dataset includes 68 different genres of art (not all of them consider visual 2D artworks), for example 'abstract', 'graffiti'. This visual art encyclopaedia (https://www.wikiart.org/en/paintings-by-style) distinguishes between genres and styles, defining style as the distinctive visual elements of the artwork, its techniques and methods, usually corresponding with an art movement (e.g. 'Cubism'). Arguably, a comprehensive benchmark style image set should include artworks spanning a broad spectrum of artistic genres and styles. However, a range of image properties (e.g. colourfulness, complexity, contrast) should be also covered.

4.1.2. *Different use cases and different types of evaluation*

Since the seminal work of Gatys et al. [GEB16], style transfer approaches have emerged that consider different aspects and introduce different perspectives. Proposed methods suggest an aesthetic standpoint, decomposing style into colour and texture [HJL*20], attempt to disentangle content and style [KSLO19], introduce geometric warps [LYH21], or utilize depth information to retain global structure and depth effect [LCLR17, CLW*19, IM22]. When considering the benchmark set of content images, it is essential to incorporate a wide variety of examples in order to accommodate all the different use cases. An informed design of the content image space should consider a comprehensive list of image properties and attributes to ensure a representative distribution is covered.

Another crucial consideration regarding the design of an evaluation benchmark is the existence of different types of evaluation. The number of subjective visual side-by-side comparisons that can be presented is limited and there are constraints to the amount of questions recruited human participants are able to respond to in the setting of a user study. For quantitative metrics, an important consideration that should be taken into account is that the larger the dataset, the more reliable the quantitative results should be.

4.1.3. *Suggested policies*

Based on these considerations, policies regarding the design of benchmark sets of images for use in the different evaluation practices can be formulated. For the visual comparisons accompanied by the subjective judgements of the authors in an NST paper, it would be beneficial for the NST community if the authors included a universally agreed-upon small set of content and style images in the results. The authors could complement the suggested benchmark images with their own images, but including some common images in all studies would allow for fairer and more insightful comparisons. For the Human Evaluation Studies, another set of content and style images can be established. This can be larger than the set of images presented in this paper, as a larger sample can provide more statistical significance to the results of a user study. Further discussion regarding the data and best practices are discussed in the next sections.

Except for the image properties that the content and style image set should satisfy, the size of the benchmark test dataset for the quantitative evaluation is another significant aspect. The larger the benchmark evaluation dataset for the computation of the quantitative metrics, the more representative and accurate the resulting comparisons would be. Theoretically, NST research could follow the standard practices in computer vision research using deep learning approaches. A split (e.g. 80%:20%) on the training dataset could accommodate a test set. However, the most commonly used datasets (Wikiart and MS COCO) might not sufficiently satisfy the aforementioned requirements for a good evaluation benchmark – further analysis is required to ensure their coverage of image characteristics is representative.

Considering the size of the evaluation benchmark in relation to the average training set is also important. If an 80%:20% training-test dataset split is hard to accommodate, a recommended guideline could be to ensure that the evaluation benchmark is at least 10% of the size of the average training set. This ensures that the evaluation dataset is representative and provides sufficient coverage of different content and style variations. An appropriate dataset, for example, would be composed of a combination of content and styles that when multiplied yields a number of stylized results that is proportional to the magnitude of the average training set utilized. If we consider MS COCO as the most common dataset, the total number of stylizations should approximately be 8000 (10% of MS COCO). Another important design decision is the number of samples in the content set and the styles set. If *NPGeneral* [MR17] is used as the benchmark content image set, the size of the benchmark styles set should be 400. As it is more crucial to evaluate how well a model generalizes in reproducing the artistic effect of a variety of artworks, the style images set could be designed first to be the smallest possible that can encompass all the different styles. Then, the content image set can be fixed accordingly. Sanakoyeu et al. used 18 styles and produced 300 stylizations per style to measure the deception rate. An image set of approximately 50 styles, combined with a content image set of size 100, would result in 5000 stylizations that could suffice for the requirements of quantitative evaluation. Such a benchmark dataset would enhance the reliability of the results reported in Table 9 which were based on only 200 stylizations for each method.

Another suggestion regarding the benchmark styles set could be to divide the set into different subsets depending on form/style [Tat24, Fer95]. As the content information (semantics) of a referenced style image (painting) is not considered for the style transfer process (since content is provided by another input image), the set of style images could be categorized according to the qualities of colour, brushwork, form and texture. For example, a subset could contain abstract paintings, while another subset could contain realistic paintings. Another subset could contain colourful paintings while another subset could contain more dull/pale artworks or artworks where texture is more dominant. Providing evaluation results for the proposed NST algorithms on each subset would enable researchers and practitioners to selectively utilize the method that best suits their intended styles and use cases. This also aligns with the guidelines by Burnell et al. [BSB*23] suggesting that reporting aggregate metrics limits understanding.

Currently, there does not exist a benchmark evaluation dataset specifically tailored for NST methods, and no substantial efforts have been made to compile a large-scale dataset that can fit the evaluation requirements of NST research. Hence, it is essential to emphasize the significance of providing all the relevant and necessary information regarding the data that is utilized during the evaluation process, so that any presented results are easily repeatable and reproducible. Initiatives (e.g. *Plan S*: https://www.coalition-s.org/) are already in place which seek to establish an open-research practice among the scientific community. High-impact journals have committed to making published research open-access [ACM23]. We encourage the NST community to also adopt these disciplines, releasing both the data and code of any novel approach. Thus, it is of great significance for the field of NST that the evaluation data is not only published but also made openly accessible. By doing so, the validity and reliability of the research findings could be significantly enhanced. This practice promotes transparency and would allow for independent verification and further exploration of the results.

4.2. Qualitative evaluation

Despite relying on subjective judgements, qualitative evaluation can provide useful insights regarding the effectiveness of a proposed method. By presenting visual results side-by-side with state-of-the-art methods (e.g. Figure 6), the differences and novelty in a proposed technique's results become evident. Additionally, the inclusion of additional zoomed-in cut-outs can effectively emphasize significant distinctions and provide enhanced visibility of the improvements made.

As described in the previous section, it could be feasible and potentially valuable if a universally agreed-upon image dataset is used for the qualitative evaluations. This could consist of a reasonably small amount of content and style images, resulting in a small number of stylizations that are derived using the same combinations of content and style in all the NST papers. Definitely, the size of this set of images would probably not be adequate to capture and highlight the effectiveness of the proposed method, therefore, the authors could complement these stylizations with more images that they chose that could be useful for the readers. However, as it would be beneficial for the field if each method included the same comparisons, a website maintained by the authors of each paper or the corresponding supplementary material section could be used to accommodate a substantial amount of image comparisons.

In the supplementary material, we include an illustration of the most commonly used content and style images. Considering that the existing literature predominantly consists of qualitative evaluations relying on this particular image set, we argue that the small benchmark dataset, encompassing a limited number of content and style images, can be constructed using images from this set. This will enable more consistent literature and allow NST researchers to draw visual comparisons with chronologically older methods.

4.3. Human evaluation studies

For the Human Evaluation Studies, the dataset utilized to produce the results shown to the participants is not the only point of

disagreement among the NST methods. Different formulations of the questions, different presentations of the comparisons, and varied numbers of participants are also among the aspects contributing to inconsistencies in conducting user studies in the field. For a comprehensive analysis and overview of the most effective practices in user studies for computer graphics research, readers are encouraged to refer to the tutorial by Malpica et al. [MSK*23].

4.3.1. Data in NST user studies

As argued in Section 4.1, it is possible to establish a benchmark evaluation dataset for user evaluation studies. Depending on the total number of questions presented, a set of content and style images can be designed to capture a wide variety of image characteristics. As already noted, the greater the number of questions posed, the more significant the results can be. For example, a user study of 50 questions can employ results synthesized from a benchmark dataset of 10 content and 5 style images. This is still a small dataset (50 images), and might not sufficiently capture the novelty and efficiency of a proposed method. To enhance the dataset, the authors could introduce a set of images they deem to be suitable, or if the dataset is adequately large, a subset could be replaced to encompass images that are more representative of the problem that is addressed. In any case, it is recommended that the authors make the user study datasets available, as this is not only good practice for the reproducibility of the results, but could also encourage subsequent studies to utilize the same dataset or build on it.

4.3.2. Participant considerations

Unlike other machine learning or computer vision tasks (e.g. object detection) that might not require crowdsourcing for evaluating the effectiveness of the results, evaluating results in NST research can be a subjective activity. Collecting and reporting supplementary data from participants regarding their experiences could significantly enrich comprehension of the results and shed light on participants' behaviour. Such additional information – demographics and task-related subjective data – could provide valuable insights that enhance the overall understanding and interpretation of the findings [CNGR*22]. Importantly, the relation of the participants to the task and to the broader NST research or any art-adjacent field should be reported, as it can potentially influence their choices. The number of participants should be kept as high as possible, but without compromising the reliability and quality of the responses.

4.3.3. Compared methods

The abundance of NST methods (Tables 1–7) renders it impractical to directly compare any single approach to the entirety of techniques that are typically considered to be state-of-the-art. Yet, it is important to compare with methods that attempt to solve a similar problem or view NST with a similar lens. For example, depth-aware methods are required to compare against state-of-the-art methods that utilize depth and aim for global structure and detail preservation. By including a comprehensive set of methods in the comparison, including newly published approaches and approaches that address the same or a similar issue, researchers can establish a more robust

evaluation framework and provide stronger evidence to support their effectiveness in synthesizing stylized visual imagery.

4.3.4. *Presentation and question formulation*

One of the most essential and critical aspects of a user evaluation study is the presentation of the questions, their exact formulation and the format of the responses that are collected. As discussed in Section 3.3, there is a variety of methods employed to present the questions and collect responses from the participants. Asking the participants to select their favourite stylization amongst all the stylizations displayed together is the most common approach. However, A/B testing and one-by-one presentation asking the participants to rate on a Likert scale are also employed. It is also worth noting that the two-alternative-forced-choice task (2AFC) has been shown to be more precise when measuring aesthetic preference [So23]. Nevertheless, it is difficult to reach a consensus as to which approach is most appropriate for each aspect of stylization (content preservation, style resemblance, overall aesthetics). Independent of the technique that is followed, we suggest the use of accompanying questions that allow the participants to justify their responses. Collecting textual feedback for each question (or a majority) regarding the subjective judgement of the participants (combined with participant-related data, as mentioned in Section 4.3.2), would allow for an improved understanding of human cognitive behaviour and the interpretation of the results. This would require further work to analyze this type of data, however, modern large language models [LZD*23] could offer valuable assistance. In addition, asking the participants to justify their selection of one stylization over another would allow the users to point out details or highlight particular regions of the images (possibly using graphical tools) where the differences are more obvious.

It is also important to acknowledge that the validity of any results is significantly impacted by the reliability of the recruited participants (Section 4.3.2). The participants should be sampled representatively; ideally, the participants would have an interest in the NST software, and ultimately they could potentially use it [BHH*23]. Nevertheless, this is hard to control, especially when utilizing crowdsourcing platforms, such as Amazon Mechanical Turk (AMT). It is possible, however, to filter out poor-quality data and avoid the potential carelessness of remote crowdworkers. An example is to remove unreliable data labelled too soon [XWF*18]. Both Bylinskii et al. [BHH*23] and Cowley et al. [CNGR*22] propose the utilization of checks randomly inserted through the study. Embedding checks/questions where there is an objectively correct answer (e.g. produce obviously bad stylizations, maybe through the content leak phenomenon, as described in Section 3.4.1) can allow detection of when participants are not paying careful attention, and help eliminate substandard responses.

Another question presentation issue is the display and positioning of the images on the screen. Certainly, all the images from all the methods in comparison should maintain the same resolution and size. If the number of methods that are compared allows, the images should be positioned next to each other, otherwise, regardless of the number of methods, the order of the images should be randomized. It has been shown that including the content and style images as part of the questions has an impact on the results

Table 10: *Suggested approach for detecting statistical significance in user studies' results.*

Setup	Collecting data from multiple participants results in an $N \times N$ contingency table where each cell represents the number of times a particular method was chosen to produce the favourite image over the other $N - 1$ methods.
Hypothesis	The null hypothesis (H_0) would be that there is no significant difference between the methods in terms of being chosen to produce the most preferred synthesized stylizations. The alternative hypothesis (H_1) would be that there is a significant difference between the performance of the approaches in comparison. A thorough explanation of hypothesis testing is given in [Duc22].
Statistical tests	*t-test:* A commonly used statistical test that compares the means of the two samples. One approach is to perform a series of pairwise *t*-tests to determine if there are significant differences between each pair of methods. Essentially, this would mean comparing the performance of the suggested NST technique with the other state-of-the-art methods.
	Rank aggregation: The overall ranking of each method can be calculated based on the number of times it was chosen as the favourite across all questions. Statistical methods such as the Friedman test [Fri37] or the Wilcoxon signed-rank test [Wil92] can be employed to determine if there are significant differences in the rankings amongst methods.
	Effect size: Effect size measures such as Cohen's d [Coh13] or Cliff's delta [Cli93] can be used to quantify the magnitude of the differences between methods, providing extra information about the practical significance of the findings.
Documenting the results	*Statistical test results:* Reporting the p-values obtained from the *t*-tests for each pair of methods could enhance the reliability of the results. Any significant differences found can also be highlighted.
	Effect sizes: Effect size measures can be included to provide a more comprehensive understanding of the differences between methods.
	Limitations: An essential part of the presentation of the results is the discussion of any limitations of the study, such as sample size, potential biases, or specific characteristics of the images or methods used.
	Visual aids: Using graphs/tables to present the results will allow readers to understand the findings more easily.

[IM22], thus, depending on the question the participants are asked, an appropriate choice should be made regarding the inclusion of the images used to infer the stylizations. For example, if the question considers content preservation or style resemblance, the content and style images should be revealed; if the question is in the form of 'choose the favourite stylization' or 'select the image with the overall highest quality', the authors should consider carefully if it is appropriate to reveal the content and style images.

4.3.5. *Reporting results*

Another consideration that lacks attention in the current literature is the concept of statistical significance in the presentation and reporting of the results. The results of user studies are mostly reported as aggregate preference rates for each method or the total number of votes accumulated for each method during the comparison, or the proportion of times the method in question was selected in comparison to each of the other methods. Yet, as is essential in literature in the area of psychology, user studies should utilize statistical tests that can better interpret the collected data. This can account for the inherent uncertainty in the results, considering that various factors such as participants' reliability, the number of participants, the number of questions asked per participant, and the experimental setting can influence the reproducibility of the findings.

The concept of statistical significance is used in psychological research to help determine whether the differences or relationships we observe in data are statistically meaningful or if they could have occurred by chance [Lyk68]. There is a range of statistical tests that can be carried out, most of them being suitable for paired data. A suggested approach for detecting statistical significance in the results of a user study and determining the best approach amongst the compared methods is given in Table 10.

The concept of statistical power can also be useful. Statistical power is the probability that the test correctly rejects the null hypothesis [Ell10]; reporting this by complementing the statistical tests can allow for a better understanding of the significance of the results. It is also a practical way to estimate the minimum sample size required for an experiment.

4.4. Quantitative evaluation metrics

There is no consensus on the most suitable quantitative metrics to use in NST studies. Nonetheless, as discussed in Section 3, we can divide them into four categories depending on the aspect of the NST process they assess: perceptual metrics, stylization performance, video metrics and efficiency. Speed and memory comparisons make up the efficiency category. However, there is a range of different metrics that compose each of the other categories. Different metrics examine specific aspects, and in some cases, even the same aspects, but employ distinct methodologies to assess them.

To suggest only a selection of those metrics as the benchmark evaluation methodology is a challenging task. However, it is a necessary task as it would be extremely inefficient to quantify the performance of an NST system by employing all the metrics depicted in Figure 3. Depending on the aim of a method, a few of the metrics can be discarded. For example, for a study proposing an algorithm for improving the stylization effect or attempting to produce an aesthetic effect [HJL*20], computing depth map or edge map differences may not be appropriate. Other metrics such as content and style error may also be discarded if seen from a more critical lens. As these are defined and utilized in the training loop, they are not providing anything useful – their use points to a circular definition since the evaluation of the content preservation or style performance qualities of the stylized results relies on the same means that were employed to create them. In addition, in contrast to LPIPS, the VGG

network used to calculate Content and Style error is not trained on distorted images, which potentially results in features that are not as good at capturing perceptual behaviour. Similarly, the SSIM metric, not relying on deep features, is shown to be inconsistent with human judgements for colour images [NAM20].

Future work will be required to necessitate a more in-depth analysis of the employed computational metrics, scrutinizing their utility and performance, with the ultimate aim of establishing a universally adopted benchmark array of metrics, consisting of a singular metric per aspect of the evaluated NST performance. Such an effort would contribute to standardizing the evaluation process and promoting consistent and meaningful comparisons among different NST methods.

As this could be a challenging task, statistical analysis, as described in the previous section on User Studies, can aid in the validity and reliability of the reported results. For any computational metric that is employed, instead of the computation of an average of the performance of each method on a test dataset, again, pairwise t-tests can be utilized to dictate the statistical significance and as a consequence, the reliability of the results. This would be useful, particularly if a benchmark dataset is not universally adopted – the statistical tests can provide a useful approximation of the repeatability of the results for different data.

Recently, several methods have considered evaluation in NST, proposing new quantitative metrics. The approach of Chen et al. [2024] is capable of assessing how well a stylized image resembles an authentic artwork. The system suggested by Wright and Ommer [WO22] combines both content preservation and style matching evaluation in a single metric, namely *ArtFID*. Also, Chen et al. [CSC*23] developed a network (*CLSAP-Net*) that uses collaborative learning which is composed of a content preservation estimation network (CPE-Net), a style resemblance estimation network (SRE-Net) and an overall vision target network (OVT-Net) attempting to provide a robust metric that effectively assesses all aspects of stylization. These metrics have exhibited a significant correlation with human judgement, suggesting that the reliance on user studies for subjective judgements can be alleviated. These efforts hold great promise in the field of NST, offering a pathway to establish a standardized evaluation procedure that reduces the reliance on costly and potentially flawed user surveys.

Although the definition of NST does not explicitly consider the aesthetics of the results, arguably, this is a concept encapsulated in the artistic nature of the stylizations. The goal of NST is to create artistic imagery, and often the recruited participants in the setting of a user study are asked to choose their favourite artistic synthesis. Despite the content and style considerations, the selection of a 'favourite' image is also based on what the viewer considers to be aesthetically pleasing. The aesthetics of the NST results is something that requires further investigation in order to adequately capture the full spectrum of the human subjective experience and ultimately replace the widely used user studies. The recently developed computational aesthetic evaluation field [MMP12, MK18, AOH*21, TM18, YTG*23] encompasses conventional (based on hand-crafted features) and learning-based (based on Deep Learning) approaches that attempt to derive aesthetic judgements on images. As the aims of the computational aesthetic evaluation field strongly correlate

with the aims of NST, resorting to such approaches could be catalytic in converging to a more robust quantitative evaluation that considers the aesthetics of the results.

4.5. Summary of evaluation issues and recommendations

An exhaustive analysis of the state-of-the-art methods has led to a collection of recommendations that could alleviate the present evaluation issues and serve as a basis for the development of a standardized evaluation protocol. Table 11 provides a condensed summary of the prevailing evaluation challenges and our recommendations to address them.

5. Conclusion

The success and widespread adoption of deep learning and computer vision techniques has undoubtedly influenced and shaped the trajectory of NST research. Along with the unparalleled success of diffusion models and text-to-image generative approaches, the field of image and video NST has continued to progress and produce remarkable ideas and implementations. Yet, the considerable progress and maturation of the field over the years has not led to the establishment of a robust and reliable evaluation protocol.

This review paper has examined the different evaluation techniques in image and video NST literature. Qualitative evaluations offer useful insights but suffer from subjectivity and reproducibility issues. Human evaluation studies, while subjective, gather quantitative data through user surveys with inconsistent methodologies across NST methods. Quantitative evaluation includes multiple computational metrics that can reliably assess the various facets of the performance of NST algorithms. Despite being reproducible and repeatable, the quantitative computational metrics are not utilized identically by each NST method. Different data and different metrics are selectively employed by each approach, suggesting a requirement for a universally agreed benchmark protocol.

Another point of contention relates to the use of data in the evaluation process. No benchmark evaluation dataset exists and the nature of the diverse evaluation approaches further exacerbates the complexity of this issue. In this paper, we have underscored the challenges arising from the utilization of disparate and often undisclosed or undefined data sources in the NST evaluation methods and emphasized the significance of establishing benchmark test datasets. Such datasets would not only facilitate fair comparisons among different methods but also enhance the transparency and reproducibility of the reported results.

The rapid advancements in NST techniques and the proliferation of diverse approaches have created a diverse landscape of evaluation methods and metrics. However, there is a lack of consensus on the most appropriate and effective evaluation criteria, leading to inconsistencies and limitations in the evaluation process. This inherent complexity highlights the need for a comprehensive and standardized evaluation protocol that can effectively assess and compare the performance of different NST methods. Such a protocol would provide a solid foundation for advancing the field, ensuring the credibility of research findings and facilitating meaningful comparisons between different approaches. The analysis in our paper provides a

Table 11: *Summary of issues in evaluation and suggested recommendations.*

Evaluation aspect	Issues [I] and recommendations [R]
Datasets	[I]: No universally agreed benchmark evaluation dataset exists. Different data is used for each type of evaluation. [R]: Development of benchmark evaluation datasets for the different aspects of evaluation: for qualitative evaluation, a small set of content images/video frames and style images that is universally shared and presented for each NST study; for user studies, a set of content and style images/video frames that capture a wide variety of characteristics and suffice for the development of adequate questions; for quantitative evaluation, a larger-scale test dataset would provide more accurate and reproducible results. Additionally, creating subsets for the style and content datasets, depending on the paintings' style or the content's characteristics, and reporting results for multiple subset datasets would further aid understanding and help distinguish which method performs better for particular image genres.
Qualitative evaluation	[I]: Inconsistencies in the data presented and commented on. Lack of common practices in deriving subjective judgements. [R]: A small dataset to be commonly shared among the methods. The subjective judgements to be clearly focused on addressing essential aspects, such as content preservation, stylization performance, observed artifacts and overall aesthetic quality.
Human evaluation studies	[I]: Dissimilarities in the design and formulation of user studies, and consequently in the presentation of the results. Lack of consensus regarding the data utilized, the number and background of participants recruited, the quantity and presentation of the questions and the reporting of the results. [R]: Development of a standardized practice for conducting user surveys. The different aspects of human evaluation studies should be considered as they can significantly influence the results and their interpretation. Essential aspects include the data that is utilized, the recruitment of participants, the number of methods in a comparison, the presentation of the questions and their formulation, the reporting of the results and potential statistical analysis.
Quantitative evaluation metrics	[I]: Different methods employ different metrics. Absence of a standardized set of quantitative evaluation metrics. [R]: Achieve a consensus on a handful of metrics that can adequately and effectively evaluate the NST results. Resort to the field of Computational Aesthetics to allow for metrics that can quantify not only content preservation and stylization performance but also the aesthetic quality of the results.

foundation upon which a standardized evaluation process could be designed and developed. A universal evaluation framework could ensure the reliability, repeatability and reproducibility required for further robust developments in the NST landscape.

Data Availability

Data sharing is not applicable to this paper as no new data were created or analysed in this study.

Conflict of Interest

The authors declare no conflict of interest.

References

[ACM23] ACM: ACM Transactions on Graphics Author Guidelines: ACM Digital Library, 2023. https://dl.acm.org/journal/tog/author-guidelines.

[AHS*21] AN J., HUANG S., SONG Y., DOU D., LIU W., LUO J.: Artflow: Unbiased image style transfer via reversible neural flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), IEEE, pp. 862–871.

[AOH*21] ACHLIOPTAS P., OVSJANIKOV M., HAYDAROV K., EL-HOSEINY M., GUIBAS L. J.: Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), IEEE, pp. 11569–11579.

[AXHL20] AN J., XIONG H., HUAN J., LUO J.: Ultrafast photorealistic style transfer via neural architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence 34*, 07 (April 2020), 10443–10450. doi: https://doi.org/10.1609/aaai.v34i07.6614.

[BHH*23] BYLINSKII Z., HERMAN L., HERTZMANN A., HUTKA S., ZHANG Y.: Towards better user studies in computer graphics and vision. *Foundations and Trends® in Computer Graphics and Vision 15*, 3 (2023), 201–252.

[BSB*23] BURNELL R., SCHELLAERT W., BURDEN J., ULLMAN T. D., MARTINEZ-PLUMED F., TENENBAUM J. B., RUTAR D., CHEKE L. G., SOHL-DICKSTEIN J., MITCHELL M., et al.: Rethink reporting of evaluation results in ai. *Science 380*, 6641 (2023), 136–138.

[Buc21] BUCHNER J.: Imagehash, 2021. https://pypi.org/project/ImageHash/.

[BWSB12] BUTLER D. J., WULFF J., STANLEY G. B., BLACK M. J.: A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision (ECCV)* (October 2012), A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, C. Schmid, (Eds.) Part IV, LNCS 7577, Springer-Verlag, pp. 611–625.

[CALL23] CHEN J., AN J., LYU H., KANAN C., LUO J.: Learning to evaluate the artness of AI-generated images. IEEE Transactions on Multimedia (2024), 1–10.

[CFYD16] CHEN W., FU Z., YANG D., DENG J.: Single-image depth perception in the wild. Advances in Neural Information Processing Systems 29 (2016).

[Cli93] CLIFF N.: Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin 114*, 3 (1993), 494.

[CLW*19] CHENG M.-M., LIU X.-C., WANG J., LU S.-P., LAI Y.-K., ROSIN P. L.: Structure-preserving neural style transfer. *IEEE Transactions on Image Processing 29* (2019), 909–920.

[CNGR*22] COWLEY H. P., NATTER M., GRAY-RONCAL K., RHODES R. E., JOHNSON E. C., DRENKOW N., SHEAD T. M., CHANCE F. S., WESTER B., GRAY-RONCAL W.: A framework for rigorous evaluation of human performance in human and machine learning comparison studies. *Scientific Reports 12*, 1 (2022), 5444.

[Coh13] COHEN J.: *Statistical Power Analysis for the Behavioral Sciences*. Routledge, New York (1988).

[CS16] CHEN T. Q., SCHMIDT M.: Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337* (2016).

[CSC*23] CHEN H., SHAO F., CHAI X., JIANG Q., MENG X., HO Y.-S.: Collaborative learning and style-adaptive pooling network for perceptual evaluation of arbitrary style transfer. *IEEE Transactions on Neural Networks and Learning Systems.* (2023).

[CWZ*21] CHEN H., WANG Z., ZHANG H., ZUO Z., LI A., XING W., LU D.: Artistic style transfer with internal-external learning and contrastive learning. *Advances in Neural Information Processing Systems 34* (2021), 26561–26573.

[CYL*17] CHEN D., YUAN L., LIAO J., YU N., HUA G.: Stylebank: An explicit representation for neural image style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), IEEE, pp. 1897–1906.

[DFI*15] DOSOVITSKIY A., FISCHER P., ILG E., HAUSSER P., HAZIRBAS C., GOLKOV V., VAN DER SMAGT P., CREMERS D., BROX T.: Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), IEEE, pp. 2758–2766.

[DSK17] DUMOULIN V., SHLENS J., KUDLUR M.: A learned representation for artistic style. In *International Conference on Learning Representations* (2017). https://openreview.net/forum?id=BJO-BuT1g.

[DTD*21] DENG Y., TANG F., DONG W., HUANG H., MA C., XU C.: Arbitrary video style transfer via multi-channel correlation. In *Proceedings of the AAAI Conference on Artificial Intelligence 35*, 2 (May 2021), pp. 1210–1217. doi: https://doi.org/10.1609/aaai.v35i2.16208.

[DTD*22] DENG Y., TANG F., DONG W., MA C., PAN X., WANG L., XU C.: Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), IEEE, pp. 11326–11336.

[Duc22] DUCA A. L.: Hypothesis testing explained, May 2022. https://www.kdnuggets.com/2021/09/hypothesis-testing-explained.html.

[Ell10] ELLIS P. D.: *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press, Cambridge, UK, 2010.

[Fer95] FERNIE E. C.: Art history and its methods: A critical anthology. *Phaidon* (1995).

[Fri37] FRIEDMAN M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association 32*, 200 (1937), 675–701.

[GCLY18] GU S., CHEN C., LIAO J., YUAN L.: Arbitrary style transfer with deep feature reshuffle. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2018), pp. 8222–8231.

[GEB15] GATYS L. A., ECKER A. S., BETHGE M.: A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* (2015).

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

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

[GFZ23] GU B., FAN H., ZHANG L.: Two birds, one stone: A unified framework for joint learning of image and video style transfers. In *Proceedings of the IEEE/CVF Conference on Computer Vision* (2024), pp. 23545–23554.

[GGZY18] GAO C., GU D., ZHANG F., YU Y.: ReCoNet: Real-time coherent video style transfer network. In *Asian Conference on Computer Vision* (2018), pp. 637–653.

[GLK*17] GHIASI G., LEE H., KUDLUR M., DUMOULIN V., SHLENS J.: Exploring the structure of a real-time, arbitrary neural artistic stylization network. *arXiv preprint arXiv:1705.06830* (2017).

[GLYY20] GAO W., LI Y., YIN Y., YANG M.-H.: Fast video multi-style transfer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2020), IEEE, pp. 3222–3230.

[HAW*23] HUANG S., AN J., WEI D., LUO J., PFISTER H.: Quantart: Quantizing image style transfer towards high visual fidelity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), IEEE, pp. 5947–5956.

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

[HJL*20] HU Z., JIA J., LIU B., BU Y., FU J.: Aesthetic-aware image style transfer. In *Proceedings of the 28th ACM International Conference on Multimedia* (2020), ACM, pp. 3320–3329.

[HJL*21] HUO J., JIN S., LI W., WU J., LAI Y.-K., SHI Y., GAO Y.: Manifold alignment for semantically aligned style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), IEEE, pp. 14861–14869.

[HJL*23] HONG K., JEON S., LEE J., AHN N., KIM K., LEE P., KIM D., UH Y., BYUN H.: AesPA-Net: Aesthetic pattern-aware style transfer networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), IEEE, pp. 22758–22767.

[HRU*17] HEUSEL M., RAMSAUER H., UNTERTHINER T., NESSLER B., HOCHREITER S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems30* (2017).

[HWL*17] HUANG H., WANG H., LUO W., MA L., JIANG W., ZHU X., LI Z., LIU W.: Real-time neural style transfer for videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), IEEE, pp. 783–791.

[IKT*17] IGNATOV A., KOBYSHEV N., TIMOFTE R., VANHOEY K., VAN GOOL L.: DSLR-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), IEEE, pp. 3277–3285.

[IM22] IOANNOU E., MADDOCK S.: Depth-aware neural style transfer using instance normalization. In *Computer Graphics & Visual Computing (CGVC)* (2022), Eurographics Digital Library.

[IM23] IOANNOU E., MADDOCK S.: Depth-aware neural style transfer for videos. *Computers 12*, 4 (2023), 69.

[JAFF16] JOHNSON J., ALAHI A., FEI-FEI L.: Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision* (2016), Springer, pp. 694–711.

[JvST*19] JAMRIŠKA O., ŠÁRKA SOCHOROVÁ, TEXLER O., LUKÁČ M., FIŠER J., LU J., SHECHTMAN E., SÝKORA D.: Stylizing video by example. *ACM Transactions on Graphics 38*, 4 (2019).

[JWY*13] JIANG H., WANG J., YUAN Z., WU Y., ZHENG N., LI S.: Salient object detection: A discriminative regional feature integration approach. In *2013 IEEE Conference on Computer Vision and Pattern Recognition* (2013), IEEE, pp. 2083–2090. doi: https://doi.org/10.1109/CVPR.2013.271.

[JYF*19] JING Y., YANG Y., FENG Z., YE J., YU Y., SONG M.: Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics 26* (2020), 3365–3385.

[KKM19] KITOV V., KOZLOVTSEV K., MISHUSTINA M.: Depth-aware arbitrary style transfer using instance normalization. *arXiv preprint arXiv:1906.01123* (2019).

[KLA*20] KARRAS T., LAINE S., AITTALA M., HELLSTEN J., LEHTINEN J., AILA T.: Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), IEEE, pp. 8110–8119.

[KLSH09] KOHAVI R., LONGBOTHAM R., SOMMERFIELD D., HENNE R. M.: Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery 18* (2009), 140–181.

[KLZ*23] KE Z., LIU Y., ZHU L., ZHAO N., LAU R. W.: Neural preset for color style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), IEEE, pp. 14173–14182.

[KSLO19] KOTOVENKO D., SANAKOYEU A., LANG S., OMMER B.: Content and style disentanglement for artistic style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), IEEE, pp. 4422–4431.

[KTH*14] KARAYEV S., TRENTACOSTE M., HAN H., AGARWALA A., DARRELL T., HERTZMANN A., WINNEMOELLER H.: Recognizing image style. In *Proceedings of the British Machine Vision Conference* (2014), BMVA Press. doi: https://doi.org/10.5244/C.28.122.

[KVL19] KURZMAN L., VAZQUEZ D., LARADJI I.: Class-based styling: Real-time localized style transfer with semantic segmentation, 2019. arXiv:1908.11525.

[LCH*17] LIU Y., CHENG M.-M., HU X., WANG K., BAI X.: Richer convolutional features for edge detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), IEEE, pp. 3000–3009.

[LCLR17] LIU X.-C., CHENG M.-M., LAI Y.-K., ROSIN P. L.: Depth-aware neural style transfer. In *Proceedings of the Symposium on Non-Photorealistic Animation and Rendering* (2017), pp. 1–10.

[LFY*17a] LI Y., FANG C., YANG J., WANG Z., LU X., YANG M.-H.: Diversified texture synthesis with feed-forward networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), IEEE, pp. 3920–3928.

[LFY*17b] LI Y., FANG C., YANG J., WANG Z., LU X., YANG M.-H.: Universal style transfer via feature transforms. *Advances in Neural Information Processing Systems30* (2017).

[LHW*18] LAI W.-S., HUANG J.-B., WANG O., SHECHTMAN E., YUMER E., YANG M.-H.: Learning blind video temporal consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), ECCV, pp. 170–185.

[LHYZ22] LUO X., HAN Z., YANG L., ZHANG L.: Consistent style transfer. *arXiv preprint arXiv:2201.02233* (2022).

[LLH*21] Liu S., Lin T., He D., Li F., Wang M., Li X., Sun Z., Li Q., Ding E.: AdaAttN: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), IEEE, pp. 6649–6658.

[LLKY19] Li X., Liu S., Kautz J., Yang M.-H.: Learning linear transformations for fast image and video style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), IEEE, pp. 3809–3817.

[LLL*18] Li Y., Liu M.-Y., Li X., Yang M.-H., Kautz J.: A closed-form solution to photorealistic image stylization. In *Proceedings of the European Conference on Computer vision (ECCV)* (2018), ECCV, pp. 453–468.

[LLW*23] Li D., Luo H., Wang P., Wang Z., Liu S., Wang F.: Frequency domain disentanglement for arbitrary neural style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence* (June 2023), AAAI, vol. 37, pp. 1287–1295. doi: https://doi.org/10.1609/aaai.v37i1.25212.

[LMB*14] Lin T.-Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P., Zitnick C. L.: Microsoft coco: Common objects in context. In *European Conference on Computer Vision* (2014), Springer, pp. 740–755.

[LPSB17] Luan F., Paris S., Shechtman E., Bala K.: Deep photo style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), IEEE, pp. 4990–4998.

[LW16a] Li C., Wand M.: Combining Markov random fields and convolutional neural networks for image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016), IEEE.

[LW16b] Li C., Wand M.: Precomputed real-time texture synthesis with Markovian generative adversarial networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14* (2016), Springer, pp. 702–716.

[LW22] Lu H., Wang Z.: Universal video style transfer via crystallization, separation, and blending. In *Proceedings of the International Joint Conferences on Artificial Intelligience(IJCAI)* (2022), vol. 36, pp. 4957–4965.

[LWLH17] Li Y., Wang N., Liu J., Hou X.: Demystifying neural style transfer. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (2017), IJCAI'17, AAAI Press, pp. 2230–2236.

[LXNC17] Li S., Xu X., Nie L., Chua T.-S.: Laplacian-steered neural style transfer. In *Proceedings of the 25th ACM International Conference on Multimedia* (2017), ACM, pp. 1716–1724. doi: https://doi.org/10.1145/3123266.3123425.

[LYH21] Liu X.-C., Yang Y.-L., Hall P.: Learning to warp for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), IEEE, pp. 3702–3711.

[Lyk68] Lykken D. T.: Statistical significance in psychological research. *Psychological Bulletin 70*, 3, Pt. 1 (1968), 151–159.

[LZ21] Liu S., Zhu T.: Structure-guided arbitrary style transfer for artistic image and video. *IEEE Transactions on Multimedia 24*, (2022), 1299–1312.

[LZD*23] Liu X., Zheng Y., Du Z., Ding M., Qian Y., Yang Z., Tang J.: GPT understands, too. *AI Open* (2023).

[MC22] Man K., Chahl J.: A review of synthetic image data and its use in computer vision. *Journal of Imaging 8*, 11 (2022), 310.

[MK18] Mohammad S., Kiritchenko S.: WikiArt emotions: An annotated dataset of emotions evoked by art. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (2018).

[MLdPM19] Meyes R., Lu M., de Puiseau C. W., Meisen T.: Ablation studies in artificial neural networks. *arXiv preprint arXiv:1901.08644* (2019).

[MMP12] Murray N., Marchesotti L., Perronnin F.: AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012), IEEE, pp. 2408–2415.

[MR17] Mould D., Rosin P. L.: Developing and applying a benchmark for evaluating image stylization. *Computers & Graphics 67* (2017), 58–76.

[MSK*23] Malpica S., Sun Q., Kellnhofer P., Beacco A., Senel G., McDonnell R., Flores Vargas M.: Effective user studies in computer graphics. In *Eurographics 2023 - Tutorials* (2023), A. Serrano, P. Slusallek, (Eds.), The Eurographics Association. doi: https://doi.org/10.2312/egt.20231030.

[MSZM17] Mechrez R., Shechtman E., Zelnik-Manor L.: Photorealistic style transfer with screened poisson equation. In *British Machine Vision Conference* (2017).

[MZLB23] Ma Y., Zhao C., Li X., Basu A.: RAST: Restorable arbitrary style transfer via multi-restoration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2023), IEEE, pp. 331–340.

[NAM20] Nilsson J., Akenine-Möller T.: Understanding SSIM. *arXiv preprint arXiv:2006.13846* (2020).

[Ost] Ostagram: https://www.ostagram.me/.

[PL19] Park D. Y., Lee K. H.: Arbitrary style transfer with style-attentional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), IEEE, pp. 5880–5888.

[PS19] PENHOUËT S., SANZENBACHER P.: Automated deep photo style transfer. *arXiv preprint arXiv:1901.03915* (2019). http://arxiv.org/abs/1901.03915.

[PTPC*17] PONT-TUSET J., PERAZZI F., CAELLES S., ARBELÁEZ P., SORKINE-HORNUNG A., VAN GOOL L.: The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675* (2017).

[RDB16] RUDER M., DOSOVITSKIY A., BROX T.: Artistic style transfer for videos. In *Pattern Recognition* (2016), Rosenhahn B., Andres B., (Eds.), Springer International Publishing, Cham, pp. 26–36.

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

[RGC*23] RUTA D., GILBERT A., COLLOMOSSE J., SHECHTMAN E., KOLKIN N.: Neat: Neural artistic tracing for beautiful style transfer. *arXiv preprint arXiv:2304.05139* (2023).

[RKH*21] RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., KRUEGER G., SUTSKEVER I.: Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (2021), PMLR, pp. 8748–8763.

[RLH*20] RANFTL R., LASINGER K., HAFNER D., SCHINDLER K., KOLTUN V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2020), *44* (3), IEEE.

[RVRK16] RICHTER S. R., VINEET V., ROTH S., KOLTUN V.: Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14* (2016), Springer, pp. 102–118.

[RWB17] RISSER E., WILMOT P., BARNES C.: Stable and controllable neural texture synthesis and style transfer using histogram losses. *arXiv preprint arXiv:1701.08893* (1 2017). https://arxiv.org/abs/1701.08893v2.

[SAOM20] SVOBODA J., ANOOSHEH A., OSENDORFER C., MASCI J.: Two-stage peer-regularized feature recombination for arbitrary image style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), IEEE, pp. 13816–13825.

[SDM19] SHAHAM T. R., DEKEL T., MICHAELI T.: Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), IEEE, pp. 4570–4580.

[SJJ*21] SINGH A., JAISWAL V., JOSHI G., SANJEEVE A., GITE S., KOTECHA K.: Neural style transfer: A critical review. *IEEE Access 9* (2021), 131583–131613.

[SKLO18] SANAKOYEU A., KOTOVENKO D., LANG S., OMMER B.: A style-aware content loss for real-time hd style transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), ECCV, pp. 698–714.

[So23] SO C.: Measuring aesthetic preferences of neural style transfer: More precision with the two-alternative-forced-choice task. *International Journal of Human–Computer Interaction 39*, 4 (2023), 755–775.

[SYZ18] SHEN F., YAN S., ZENG G.: Neural style transfer via meta networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), IEEE, pp. 8061–8069.

[Tat23] Tate: Art terms, 2023. https://www.tate.org.uk/art-terms/.

[Tat24] Tate: 2024. https://www.tate.org.uk/art-terms/f/formalism.

[TLL*23] TANG H., LIU S., LIN T., HUANG S., LI F., HE D., WANG X.: Master: Meta style transformer for controllable zero-shot and few-shot artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), IEEE, pp. 18329–18338.

[TM18] TALEBI H., MILANFAR P.: NIMA: Neural image assessment. *IEEE Transactions on Image Processing 27*, 8 (2018), 3998–4011.

[ULVL16] ULYANOV D., LEBEDEV V., VEDALDI A., LEMPITSKY V. S.: Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML* (2016), vol. *1*, p. 4.

[UVL17a] ULYANOV D., VEDALDI A., LEMPITSKY V.: Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), IEEE, pp. 6924–6932.

[UVL17b] ULYANOV D., VEDALDI A., LEMPITSKY V. S.: Instance normalization: The missing ingredient for fast tylization, 2017. arXiv:1607.08022.

[Vid19] Videvo: 2019. https://www.videvo.net/.

[WBSS04] WANG Z., BOVIK A. C., SHEIKH H. R., SIMONCELLI E. P.: Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing 13*, 4 (2004), 600–612.

[WFJ*17] WILBER M. J., FANG C., JIN H., HERTZMANN A., COLLOMOSSE J., BELONGIE S.: BAM! the behance artistic media dataset for recognition beyond photography. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), IEEE, pp. 1202–1211.

[Wil92] WILCOXON F.: Individual comparisons by ranking methods. *Breakthroughs in Statistics: Methodology and Distribution*. Springer, New York, 1992, pp. 196–202.

[WO22] WRIGHT M., OMMER B.: Artfid: Quantitative evaluation of neural style transfer. In *Pattern Recognition: 44th DAGM German Conference, DAGM GCPR 2022, Konstanz, Germany, September 27–30, 2022, Proceedings* (2022), Springer, pp. 560–576.

[WYXL20] WANG W., YANG S., XU J., LIU J.: Consistent video style transfer via relaxation and regularization. *IEEE Transactions on Image Processing 29* (2020), 9125–9139.

[WZC*21] WANG Z., ZHAO L., CHEN H., ZUO Z., LI A., XING W., LU D.: Evaluate and improve the quality of neural style transfer. *Computer Vision and Image Understanding 207* (2021), 103203.

[WZDB22] WU Z., ZHU Z., DU J., BAI X.: CCPL: Contrastive coherence preserving loss for versatile style transfer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI* (2022), Springer, pp. 189–206.

[WZZ*22] WANG Z., ZHANG Z., ZHAO L., ZUO Z., LI A., XING W., LU D.: AesUST: towards aesthetic-enhanced universal style transfer. In *Proceedings of the 30th ACM International Conference on Multimedia* (2022), ACM, pp. 1095–1106.

[XLN23] XU W., LONG C., NIE Y.: Learning dynamic style kernels for artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), IEEE, pp. 10083–10092.

[XT15] XIE S., TU Z.: Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), IEEE, pp. 1395–1403.

[XWF*18] XU Z., WILBER M., FANG C., HERTZMANN A., JIN H.: Learning from multi-domain artistic images for arbitrary style transfer. *arXiv preprint arXiv:1805.09987* (2018).

[XXL*21] XIA X., XUE T., LAI W.-s., SUN Z., CHANG A., KULIS B., CHEN J.: Real-time localized photorealistic video style transfer.

In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2021), IEEE, pp. 1089–1098.

[YTG*23] YI R., TIAN H., GU Z., LAI Y.-K., ROSIN P. L.: Towards artistic image aesthetics assessment: A large-scale dataset and a new method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), IEEE, pp. 22388–22397.

[YUC*19] YOO J., UH Y., CHUN S., KANG B., HA J.-W.: Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), IEEE, pp. 9036–9045.

[ZD18] ZHANG H., DANA K.: Multi-style generative network for real-time transfer. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops* (2018).

[ZHW*23] ZHU M., HE X., WANG N., WANG X., GAO X.: All-to-key attention for arbitrary style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), IEEE, pp. 23109–23119.

[ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), IEEE, pp. 586–595.

[ZTD*23] ZHANG Y., TANG F., DONG W., HUANG H., MA C., LEE T.-Y., XU C.: A unified arbitrary style transfer framework via adaptive contrastive learning. *ACM Transactions on Graphics 42*, (5) (2023).

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information