

# A Comprehensive Evaluation of Arbitrary Image Style Transfer Methods

Zijun Zhou, Fan Tang, *Member, IEEE*, Yuxin Zhang, Oliver Deussen, Juan Cao, Weiming Dong, *Member, IEEE*, Xiangtao Li, *Senior Member, IEEE*, Tong-Yee Lee, *Senior Member, IEEE*

**Abstract**—Despite the remarkable process in the field of arbitrary image style transfer (AST), inconsistent evaluation continues to plague style transfer research. Existing methods often suffer from limited objective evaluation and inconsistent subjective feedback, hindering reliable comparisons among AST variants. In this study, we propose a multi-granularity assessment system that combines standardized objective and subjective evaluations. We collect a fine-grained dataset considering a range of image contexts such as different scenes, object complexities, and rich parsing information from multiple sources. Objective and subjective studies are conducted using the collected dataset. Specifically, we innovate on traditional subjective studies by developing an online evaluation system utilizing a combination of point-wise, pair-wise, and group-wise questionnaires. Finally, we bridge the gap between objective and subjective evaluations by examining the consistency between the results from the two studies. We experimentally evaluate CNN-based, flow-based, transformer-based, and diffusion-based AST methods by the proposed multi-granularity assessment system, which lays the foundation for a reliable and robust evaluation. Providing standardized measures, objective data, and detailed subjective feedback empowers researchers to make informed comparisons and drive innovation in this rapidly evolving field. Finally, for the collected dataset and our online evaluation system, please see <https://github.com/ZhouZJ-DL/A-Comprehensive-Evaluation-of-Arbitrary-Image-Style-Transfer-Methods>.

**Index Terms**—Image style transfer, assessment system, multiple granularity evaluation

## 1 INTRODUCTION

ARBITRARY image style transfer (AST), or neural style transfer [1], allows users to “paint” in an artistic style of a given painting, using its brushstrokes and textures, onto another image without restrictions. The rapid growth of AST methods mirrors the rise of deep learning and neural networks, with different approaches such as CNN-based models [2]–[8], flow-based models [9], [10], transformer-based models [11]–[13], and recently diffusion-based models [14]–[19]. This dynamic landscape raises crucial questions about the proper assessment of AST methods, the potential limitations of their foundational models, and the value of past approaches.

Currently, AST evaluation primarily relies on self-reported assessments, often utilizing diverse objective metrics and user surveys. However, concerns persist regarding the potential influence of the chosen content and style images on these evaluations. Paper authors seem to have a tendency to show working examples while hiding failures. Figure 1 (a) presents a visual comparison of stylized images generated from two sources: results by the authors (left) and

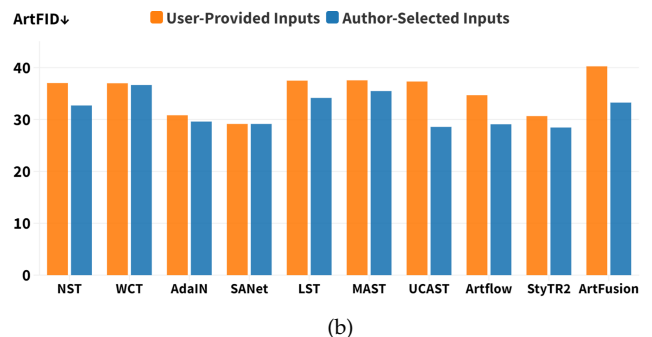
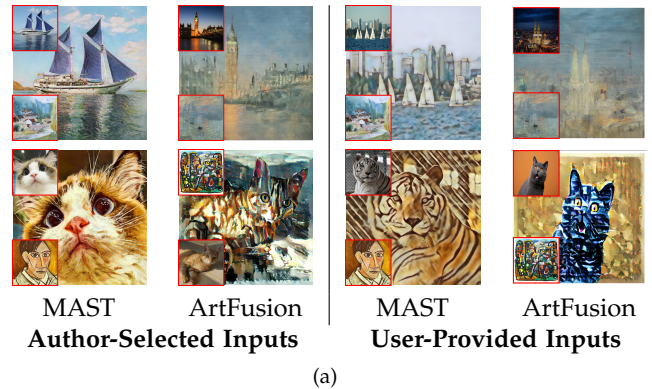


Fig. 1: Comparisons for author-selected inputs or user-provided inputs: (a) visual stylized results by MAST [20] and ArtFusion [14], (b) quantitative statistic by various AST methods.

results by in-the-wild inputs (right), suggesting potential quality differences between the two groups. Notably, author-presented results typically show faithful content preservation

- This work was partly supported by the National Natural Science Foundation of China under Nos. 62102162, U20B2070, the Beijing Science and Technology Plan Project under no. Z231100005923033, the National Science and Technology Council under Grant No. 113-2221-E-006-161-MY3, Taiwan, and the German Research Foundation (DFG) Project under no. 508324734. (Co-corresponding author: Fan Tang and Xiangtao Li.)
- Z. Zhou and X. Li are with Jilin University, Changchun, China.
- F. Tang and J. Cao are with Institute of Computing Technique, Chinese Academy of Sciences, Beijing, China.
- Y. Zhang and W. Dong are with Institute of Automation, Chinese Academy of Sciences, Beijing, China.
- O. Deussen is with University of Konstanz, Germany.
- T.-Y. Lee is with National Cheng Kung University, Tainan, Taiwan.

and successfully capture the intended artistic style. In contrast, the stylized images selected by users suffer from various issues, such as content leaks, lack of a distinct artistic style, or degraded image quality. To quantify these potential differences, we applied several AST [2], [3], [5], [8], [9], [11], [14], [20]–[22] methods to sets of images provided by authors and by users. This is in line with the ArtFID study [23], which consistently revealed superior results when stylizing the original images compared to non-original ones (Figure 1 (b)). This finding underscores the influence of the image domain on style transfer outcomes and highlights the need for a more comprehensive evaluation system for an accurate assessment of AST methods.

Previous efforts for AST assessments [24]–[27] (summarized in Table 1) have made valuable contributions by incorporating objective and subjective tools. For example, Chen et al. developed a new arbitrary style transfer database (AST-IQAD [26]) to facilitate unbiased evaluation of image stylization. This dataset comprises 75 content images, 126 style images, and 1200 stylized images generated using eight representative AST methods, each annotated with human opinion scores. However, two key limitations persist in the aforementioned systems: (a) the intricate interplay between diverse factors influencing AST performance remains insufficiently explored. This hinders a comprehensive understanding of how content features, artistic style characteristics, and selected neural network architectures interact and affect outcomes. (b) Restricted user interaction formats and insufficient data collection, potentially compromising the accuracy and generalizability of the findings.

We present a novel framework for AST assessment, utilizing a fine-grained dataset and a multi-granular evaluation system that integrates objective and subjective perspectives. We first delve deeper into AST performance through a granular objective study. This leverages the parsing information of datasets and the recognition abilities of the RAM model [28] to automatically label content and style images. Such enriched data allows us to explore the influence of various factors on AST performance, including:

- *Scene diversity*: we analyze how diverse scene types, such as landscapes, portraits, and urban environments, impact AST results.
- *Object complexity*: the number and intricacy of objects within images are examined for their effect on performance.
- *Salient regions*: we focus on the area occupied by the most frequently occurring objects in an image, exploring its role in effective style transfer.
- *Content-style consistency*: we investigate the alignment between content and style image characteristics.

Our analysis confirms the pervasive presence of the content-style trade-off phenomenon in AST methods. Specifically, the diffusion-based ArtFusion [14] and the manifold-based MAST [20] excel in metrics that measure content preservation, while the CNN-based SANet [8] (with attentional layer) and AdaIN [5] (with adaptive weight) prioritize style transfer. The CNN-based UCAST [22] utilizes contrastive learning, while the Transformer-based StyTr<sup>2</sup> [11] exhibits a more balanced performance. A finer-grained assessment reveals that (a) images with object counts between 30 and

50 achieve better results compared to other ranges, (b) stylization generally performs better on images with larger salient regions, except for ranges between 60% and 70%, (c) a positive correlation exists between content-style similarity and AST performance.

Building upon this objective analysis, we propose a multi-granular subjective assessment method. It utilizes a diversified spectrum of user surveys, including point-wise (individual image evaluation), pair-wise (comparison between two images), and group-wise (collective evaluation) methods, providing a robust and comprehensive assessment of AST performance. Our experimental methodology for hierarchical evaluation can be found on <http://ivc.ia.ac.cn/>. Our findings suggest that group-wise subjective studies achieve higher consistency compared to individual assessments, indicating a more robust evaluation method for AST performance. Interestingly, the image parsing aspect revealed contrasting results between subjective and objective studies. Images with an object count in the range [10, 20) were rated higher in subjective evaluations, whereas objective metrics favored images with a smaller salient region size. This discrepancy warrants further investigation and highlights the potential limitations of solely relying on objective metrics for comprehensive AST evaluation. Regarding metric correlation, our analysis yielded several key insights: (a) content preservation ability, as measured by various metrics, contributes more significantly to the overall visual quality perceived by human judges, (b) the LPIPS [29] and SRQE [26] metrics demonstrates strong consistency with human perception in terms of content preservation, (c) to some extent, ArtFID [23] and SRQE [26] can measure artistic features in a way that aligns with human judgment.

In summary, our contributions are as follows:

- We propose a multi-granular evaluation system supported by an efficient assessment strategy to obtain robust and valid user feedback on AST performance.
- We collected a fine-grained dataset considering a range of image contexts from multiple sources.
- We selected ten AST methods from four typical architectures for systematic objective and subjective evaluations and conducted comprehensive analysis across different image contexts.

## 2 RELATED WORKS

**AST assessment.** Although the evaluation of AST work faces numerous challenges, there have been attempts made to address this. Initially, researchers only compared the differences in algorithms between different methods [24]. With the development of image quality evaluation metrics, more advanced indicators have been used to assess these AST methods. Chen et al. [26] constructed a dataset (AST-IQAD) comprising subject-rated scores and introduced the SRQE metric to quantitatively predict human perception of stylized images. Furthermore, CLASP [27] employs a style-adaptive pooling strategy for collaborative learning of stylized image quality. Concurrently, user survey methodologies have diversified, with Chen et al. [26], [27] conducting studies based on various quality factors, including content preservation, style resemblance, and overall visual. Several studies [1],



TABLE 1: AST Methods’ Assessment

Works	Subjective Study			Objective Study		Dataset	
	Point-wised	Pair-wised	Group-wised	Image Quality	Artistic Feature	Standard	Semantic Labels
Majumdar et al. [24]	✗	✗	✗	✓	✗	✗	✗
Wang et al. [25]	✗	✗	✗	✓	✗	✗	✗
SRQE [26]	✗	✓	✗	✓	✓	✓	✗
CLSAP [27]	✗	✓	✗	✓	✓	✓	✗
Ours	✓	✓	✓	✓	✓	✓	✓

[30]–[33] have provided reviews and analyses of different AST methods.

At the same time, research in AIGC evaluation offers valuable guidance for designing our framework. Studies such as Li et al. [34] explore Mean Opinion Score (MOS) as a metric for model evaluation. Wang et al. [35] assess six generative models across diverse scene categories, while Zhang et al. [36] propose a standardized evaluation dataset. These approaches aim to establish a standardized evaluation methodology that mitigates bias in AST assessment.

**Image quality assessment.** Initially, people used pixel-based metrics such as MSE distance, PSNR, and SSIM to evaluate image quality, but these methods often did not align well with human perception. Gatys et al. [2] iteratively updated images using content loss and style loss based on Gram matrices by extracting image features from specific layers and calculating their differences. The LPIPS [29] metric is highly aligned with human perception, utilizing deep neural networks to extract features from images, which can be used for evaluation tasks in stylization. With the emergence of image generation models such as GANs and Diffusion models, metrics such as FID [37] and ArtFID [23] have been proposed, aimed at evaluating the diversity and quality of generated images.

**Image style transfer.** The field of image style transfer encompasses domain-specific transfer and arbitrary style transfer (AST). While domain-specific transfer methods [6], [7], [38]–[45] target specific styles within predefined domains, AST allows for transferring any image to any desired style. We chose to focus on AST due to its greater flexibility and potential for wider application. We have developed a parsing dataset and evaluation system designed to accommodate various AST methods. This infrastructure can be readily adapted to different scenarios within the broader domain of image style transfer.

AST began with the work of Gatys et al. [2]. However, the involved optimization-based process is slow and not well-suited for practical applications. As CNN architectures evolved, a large number of feed-forward network image style transfer algorithms emerged [2]–[5], [8], [9], [11], [20]–[22], [38], [46]–[48]. Among them, flow models [9], [10] and manifold algorithms [20], [49], [50] address issues with content leakages. Attention-based networks [8], [13], [51] capture the artistic features of images well. However, these CNN-based methods had issues such as having low generation quality, limited diversity of generated images, as well as deviations from the original content.

With the development of image generation models, high-quality and more realistic image generation models using GANs [52], transformers [53], [54] and diffusion models [55]–[59] emerged. The development of AST algorithms

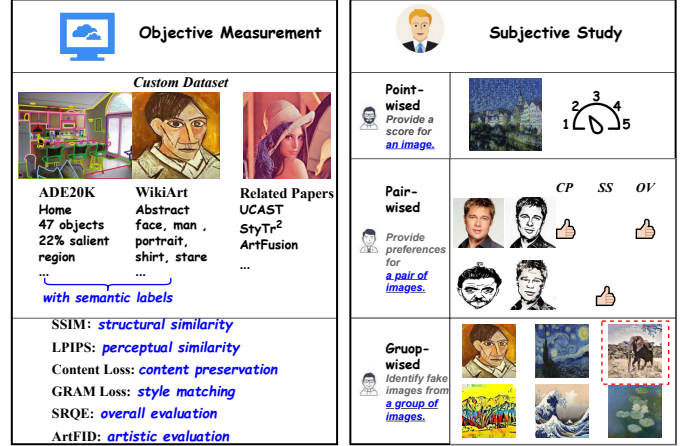


Fig. 2: Proposed experimental system combining subjective and objective evaluation.

has closely paralleled advancements in image generation models. Transformer-based AST [11]–[13], [60] and GAN-based AST [61]–[64] showcase this synergy. Diffusion-based technologies have significantly improved the quality and diversity of image generation, offering even more powerful tools for implementing AST [14], [16]–[19], [39], [44], [65]–[67].

Existing objective assessment methods are susceptible to biases, such as those related to the diverse scene types present in the images under evaluation. This can compromise their validity, hindering accurate comparisons between AST methods. Additionally, existing subjective studies often yield inconsistent and unreliable results due to methodological limitations or individual variations in perception. This lack of robustness makes it difficult to draw definitive conclusions about the performance of these methods. In this paper, we therefore propose a multi-granularity evaluation framework that explores diverse image segmentation information that potentially affects the stylization performance. Additionally, we introduce a group-wise user study to enhance the robustness of the evaluation process.

### 3 EXPERIMENTAL SYSTEM COMBINING SUBJECTIVE AND OBJECTIVE INVESTIGATIONS

This section describes the experimental framework employed to evaluate the performance of various AST methods, the utilized dataset, and the specific configurations for the components of the objective and subjective study. Figure 2 provides a comprehensive overview of the system architecture.

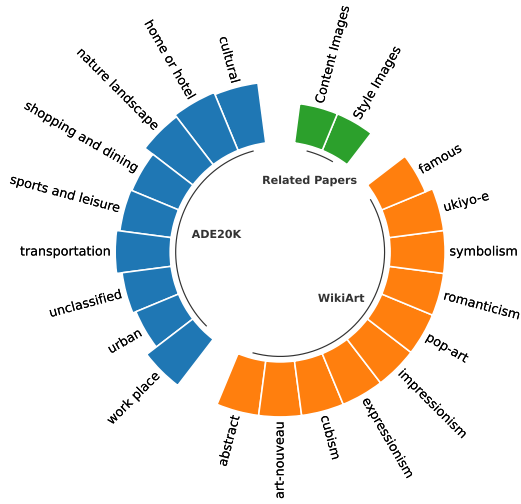


Fig. 3: Our dataset comprises three distinct components: (a) 5,000 diverse images extracted from the ADE20k [68] dataset (b) 5,000 paintings obtained from the official website of WikiArt, encompassing a wide range of artistic styles and genres. (c) 52 content images and 52 style images carefully selected from related research papers, providing specific examples and comparisons.

### 3.1 Dataset Construction

Our research delves into the performances of various AST methods through a comprehensive evaluation system that blends subjective and objective assessments. To ensure a robust and insightful evaluation, we constructed a custom dataset rich in diversity and complexity, catering specifically to the needs of our study. The custom dataset is divided into **Content Dataset** and **Style Dataset**, as shown in Table 2. The detailed setting of the dataset is described below.

#### A. Content Dataset

To construct the content part of our custom dataset, we first randomly sampled 5,000 images from the **ADE20K** dataset by Zhou et al. [68]. This selection encompasses a diverse range of ten coarse-grained image classification labels, including: *cultural, home or hotel, industrial, natural landscape, shopping and dining, sports and leisure, transportation, unclassified, urban, and workplace*. These labels represent various categories such as natural landscapes, indoor scenes, urban environments, and still-life settings, ensuring the richness and diversity of our custom dataset. Beyond providing high-level semantic categories, ADE20K’s rich annotations unlock opportunities for fine-grained analysis: object part labels, instance annotations, and original annotated polygons enable scene parsing, empowering us to explore intricate relationships between object categories, count, areas, and model performance.

To showcase optimal results and facilitate comprehensive performance comparisons with our custom datasets, we further constructed supplemental datasets leveraging 52 original images from related papers.

#### B. Style Dataset

WikiArt (<https://www.wikiart.org/>) serves as a treasure

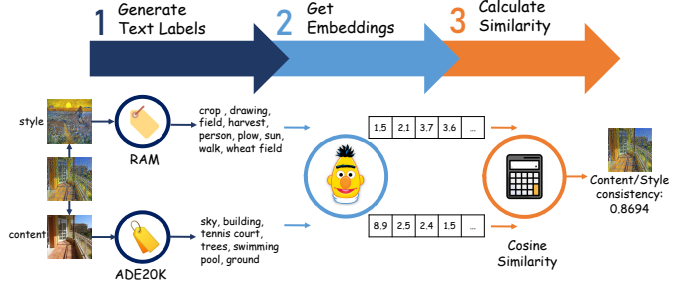


Fig. 4: Dataset preprocessing: generating textual labels for content and style images, facilitating the calculation of content/style similarity for each (content, style) pair.

trove for art enthusiasts and researchers, providing a crowdsourced platform showcasing artists and their works. Each piece comes with rich metadata like genre, style, and artist biography, making it ideal for selection. We carefully selected 5,000 style images from WikiArt, encompassing ten distinct styles: *abstract-expressionism, art nouveau modern, cubism, expressionism, impressionism, pop art, realism, romanticism, symbolism, and ukiyo-e*. This selection ensures a good representativeness for the AST task. To further strengthen the dataset, we also included 100 additional images from famous artworks on WikiArt, frequently featured in representative AST research.

Prior to further analysis, all images underwent preprocessing using the Recognize Anything Model (RAM) [28]. RAM leverages the power of automatic text semantic parsing to generate tags for annotation-free images, making it a valuable tool for large-scale image analysis. We employed RAM to perform content recognition on above 5,000 real artistic images. Each image received a set of relevant tags, such as “artist”, “drawing”, “dress”, “person”, “painter”, “robe”, “stand”, and “woman”. As described in the content part, we also constructed supplementary datasets containing 52 style images based on the original images from related publications.

### 3.2 Dataset Preprocessing

**Basic configuration.** To ensure a fair comparison and consistency in data handling, we relied on the official implementations of all chosen AST methods downloaded from GitHub. While different methods have their own preprocessing pipelines, we standardized the resolution of all images in our custom dataset to 512×512 pixels.

For methods like MAST [20] and ArtFusion [14], capable of both artistic and photorealistic styles, we specifically focused on their artistic image style transfer capabilities in this study. This decision aligns with our aim to analyze the artistic effects of different AST methods, such as brushstrokes, lines, and textures, rather than photorealism.

**Consistency calculating.** We investigate the influence of content-style consistency on image style transfer performance. To measure content-style consistency, the following steps have been taken as shown in Figure 4:

TABLE 2: Custom Dataset for AST’s Assessment

Part	Dataset	Descriptions
Content Dataset	ADE20K [68]	5,000 images from ADE20K dataset. Each image in our dataset is thoroughly annotated with rich scene parsing information. 1) Scene Type: such as nature landscapes, or urban. 2) Object Count: Counting of objects within the image. 3) Main Object Area Ratio: the relative size of the dominant object in the image, 4) Text Label: a descriptive text label using ADE20K [68]’s scene parsing information.
	Papers [2], [3], [5], [8], [9], [11], [14], [20], [21], [46]	52 original content images obtained from the related papers.
Style Dataset	WikiArt <a href="https://www.wikiart.org/">https://www.wikiart.org/</a>	5,000 artistic images from WikiArt, encompassing 10 distinct styles, <i>abstract-expressionism, Art Nouveau, cubism, expressionism, impressionism, pop art, realism, romanticism, symbolism, and ukiyo-e</i> . 100 iconic artistic masterpieces to enrich representativeness. Each image has a descriptive text label generated by RAM [28].
	Papers [2], [3], [5], [8], [9], [11], [14], [20], [21], [46]	52 original style images obtained from the related papers.

- 1) Generate text labels for each content image using the scene parsing information of ADE20K [68] such as {sky, sea water, building, hill, water, bridge, ground, sand beach, palm tree, palm trees}.
- 2) Generate text labels for each style image using the RAM [28] (see Section 3.1) model, such as {artist, drawing, dress, person, painter, robe, stand, woman}.
- 3) Use the BERT [53] model to obtain embeddings for the image’s text labels of style and content.
- 4) Compute the cosine similarity between the style and content embeddings obtained from the BERT model.

### 3.3 AST Methods Participating in the Assessment

To gain comprehensive insights into the performance of arbitrary style transfer and its relationship with different models, we conducted experiments by selecting ten AST methods (listed in Table 3) that represent different architectural approaches:

- **CNN-based** models leverage convolutional neural networks to extract and transfer stylistic elements from a style image onto a content image. NST [2] leverages the Gram matrix for style analysis, achieving artistic style transfer through iterative optimization, while WCT [3] and AdaIN [5] utilize statistical features like mean and variance or feature transforms to capture and transfer style. LST [21] learns the linear transformation matrix instead of relying on second-order statistics. SANet [8] uses a style-attentional layer to integrate the local style patterns according to the semantic spatial distribution of the content image. MAST [20] aligns multi-manifold distributions for semantically consistent style transfer. UCAST [22] improves arbitrary style transfer with a novel style representation by contrastive learning-based optimization.
- **Flow-based** models consist of a sequence of invertible functions that map an input distribution to an output distribution. ArtFlow [9] uses reversible neural flows, producing high-quality stylized images with various styles and avoiding content leaks.
- **Transformer-based** models allow neural networks to focus on specific parts of the input data. StyTr<sup>2</sup> [11]

TABLE 3: Categories of AST Methods and Examples

Category	Method	Publication
CNN-based	NST [2]	CVPR 2016
	WCT [3]	NIPS 2017
	AdaIN [5]	ICCV 2017
	LST [21]	CVPR 2019
	SANet [8]	CVPR 2019
	MAST [20]	ICCV 2021
	UCAST [22]	ACM TOG 2023
Flow-based	ArtFlow [9]	CVPR 2021
Transformer-based	StyTr <sup>2</sup> [11]	CVPR 2022
Diffusion-based	ArtFusion [14]	Preprint 2023

uses a transformer-based framework to capture input image features and achieve unbiased content representation.

- **Diffusion-based** models can perform style transfer on images using conditional diffusion models. ArtFusion [14] introduces a novel probabilistic dual conditional latent diffusion model, which can disentangle the style and content information of an image and use them to guide the reverse diffusion process.

### 3.4 Objective Assessment

We conducted an experiment testing the ten AST methods mentioned in Section 3.3 across five metrics:

- *Image quality metrics*: The Structural Similarity Index (SSIM [69]) is a widely used metric for measuring the similarity between two images. The SSIM index is calculated on various windows of an image: 
$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$
- *Neural network metrics*: measuring similarity between images based on feature extraction and comparison through a neural network. In our work, content Loss and GRAM loss were implemented following the methodology outlined in Gatys et al [2]. The LPIPS metric [29] utilized the default configuration provided within the official LPIPS implementation. It measures the perceptual similarity between two images. A higher LPIPS score indicates that the images are further apart or more different from each other.
- *SRQE*: SRQE [26] is a sparse representation-based image quality evaluation metric utilizing a discrete

dictionary for training, enabling it to provide consistent evaluation scores without retraining. Based on the official implementation, SRQE can directly quantify stylized image quality in terms of content preservation (SRQE\_CP), style resemblance (SRQE\_SR), and overall visual (SRQE\_OV).

- *ArtFID*: ArtFID [23] supports a quantitative measurement for stylization performance, facilitating similar studies to further analyze and improve style transfer methods. When calculating ArtFID, both content and style images are taken into consideration:

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

FID [37] is the Fréchet Inception Distance,  $d$  stands for LPIPS in our experiment (default option), the  $X_c$  are the content images, the  $X_s$  are the style images, and the  $X_g$  are the stylized images.

It's worth noting that beyond conventional metrics like LPIPS and SSIM, which primarily measure content similarity, our evaluation metrics include ArtFID [23]. This metric provides a more comprehensive evaluation that reflects the human judgment of artistic success.

### 3.5 Subjective Assessment

To collect user feedback on the stylized images, we designed a survey that contains three types of questions that capture different levels of granularity: rating questions (point-wise), comparison questions (pair-wise), and identification questions (group-wise). The survey is available online at <http://ivc.ia.ac.cn>. Users can view and evaluate over 7,000 sets of content images, style images, and stylized results generated by various image style transfer methods mentioned in Section 3.3.

#### 3.5.1 Multi-granularity Setting

##### A. Rating Questions (Point-Wise)

Participants saw three images: a real image, an artistic image representing the target style, and a stylized image generated by one of our chosen AST methods. Their task was to evaluate the quality of the AST-generated image by providing a Mean Opinion Score (MOS) on a scale ranging from 0 (poor) to 5 (excellent), with 0.5 increments. The user interface for MOS scoring is displayed in the first row on the right side of Figure 2.

##### B. Comparison Questions (Pair-Wise)

The questions for comparison are illustrated in the second row on the right side of Figure 2. Users are presented with the original image to be stylized, an example of the desired artistic style, and two stylized images rendered using two different AST methods.

For each question, users had to evaluate the two stylized images on three criteria: content preservation, style similarity, and overall visual effect. For each of these criteria, users selected their preference between Method A and Method B. We added two options in comparison questions: users can choose "Both are Good" if they found it difficult to distinguish between the two styles or "Both are Bad"

if they considered the synthesis results of both methods disappointing. This comprehensive set of options allows for an accurate evaluation of all AST methods in the user survey.

To facilitate meaningful comparisons, we leverage an efficient assessment strategy that selects two AST methods with comparable performance for each user question. This approach ensures humans are judging AST methods with similar performance levels. The details of this strategy are provided in Section 3.5.3.

##### C. Identification Questions (Group-Wise)

Recognizing limitations in achieving high user consistency with point-wise and pair-wise evaluations, we design a group-wise user study to assess user preferences and perceptions. In this study, participants encounter sets of ten images containing both real artworks and stylized images generated by a single AST method under investigation. They have to identify the "fake" stylized images within each set, essentially distinguishing human-created art from machine-generated renditions (a kind of Turing art test). This approach leverages the collective intelligence of participants to potentially reveal finer-grained stylistic nuances and enhance consistency compared to point-wise and pair-wise user evaluations. The group-wise user study format is illustrated in Figure 2 (third row, right side). In Section 5.5, we present an experimental evaluation that showcases the advantages of group-wise subjective studies.

Furthermore, to systematically explore the impact of the image ratio on user performance and to establish benchmarks for future research, the website offers three pre-defined ratios of fake images to real artworks: 3:7, 4:6, and 5:5. This controlled variation of image composition allows us to investigate potential shifts in user confidence and accuracy as the proportion of real and stylized images changes.

#### 3.5.2 Investigate Credibility

Maintaining data reliability is crucial for a user survey. Therefore, we implemented user review mechanisms to validate the responses and ensure reliable data for the analysis. These reviews assess the consistency of user judgments within a designated time frame, guaranteeing that their evaluation criteria remain relatively stable. Specifically, we evaluate the validity of both comparison and rating questions for each participant.

**Rating questions.** During the rating task, after completing eight questions, users encounter a "review question". This randomly selected question, already answered earlier, presents the same stylized image for re-evaluation. If the user's rating for the same image differs by more than two levels between the two attempts, we deem the review unsuccessful and invalidate their corresponding survey response. This approach safeguards against inconsistencies in user judgment within the rating.

**Comparison questions.** Following 16 comparison questions, participants also faced a "review question". The question, randomly selected and also previously answered, reappeared with the same images and results from two different AST methods. Users re-evaluated these methods across three criteria: content preservation, style transfer, and overall effect. However, if their judgment differed on two or more of these criteria compared to their initial response, their entire survey response for that question was marked as invalid.



### 3.5.3 Efficient Assessment Strategy

Given the complexity of comparing 45 pairwise relationships of ten AST methods across 6,000 images, traditional user surveys would be impractical due to the extensive number of questions and potential user fatigue. We therefore propose an efficient assessment strategy to effectively address the challenge of allocating limited human resources (Algorithm 1). This strategy aims to concentrate human effort on the most valuable comparisons, maximizing the impact of their feedback.

- To ensure all AST methods receive equal opportunities for comparison, the algorithm leverages an efficient mechanism inspired by Khosla et al. [70]. This mechanism prioritizes methods with lower memory scores, indicating less recent involvement in comparisons. This ensures that all methods fully participate in a comprehensive evaluation while reducing user fatigue.
- To foster meaningful competition, the algorithm implements a point system. Our point system assigns three points to the method of generating the user-preferred stylized image during the evaluation. The non-preferred method receives no points. If users select "Both Are Good," both methods get one point each. The scores of all participating AST methods' are recorded, and the algorithm selects two methods with similar accumulated points for each comparison question. Users then evaluate the results generated by these two methods.

This dynamic approach balances competitive evaluation with ensuring all methods have a fair chance to be seen. Our proposed algorithm balances the need for a comprehensive evaluation with limiting user burden, making it suitable for such a large-scale comparison.

## 4 ANALYSIS OF OBJECTIVE EXPERIMENTS

### 4.1 The Overall Ranking of AST Methods

For our comprehensive evaluation, we first conducted a broad experimental analysis. This allowed us to identify potential trends and outliers before delving into finer-grained comparisons. We utilized six metrics — SSIM [69], Content Loss [2], GRAM Loss [2], LPIPS [29], SRQE [26] and ArtFID [23] — to assess the performance of the AST methods detailed in Section 3.3.

In our study, SSIM [69], content loss [2] and LPIPS [29] assess the degree to which a stylized image retains the visual characteristics of the original content image. They calculate the differences between content and stylized images. GRAM loss [2] measures the similarity between the feature maps of the style image and the stylized image. SRQE [26] quantifies stylized image quality in terms of content preservation (CP), style resemblance (SR), and overall visual (OV) appeal. This section utilizes SRQE\_OV for assessment. ArtFID [23] leverages a set of content, style, and stylized images to provide an overall assessment of the stylized image's quality. The aggregated results are presented in Figure 5. The heatmap visually depicts the performance of each method across all metrics, with darker shades signifying higher scores and the numbers indicating the respective rank within each

---

### ALGORITHM 1: Efficient Assessment Strategy

---

**Data:** Ten AST methods  $AST_{1\sim 10}$  and respective memory scores  $M_{1\sim 10}$ , points  $p_{1\sim 10}$  and length of time not present in comparison  $T_{1\sim 10}$ . Base memory score  $c$ . Weight factor  $\alpha$ .

**Result:** Two methods selected for comparison:  $m_1, m_2$

```

/* Choose the method with the lowest memory score */
1  $j \leftarrow \min_{j \in [1,10]} M_j$ ;
2  $m_1 \leftarrow AST_j$ ;
/* Find 3 methods closest to the points of  $AST_1$  */
3  $distances \leftarrow \emptyset$ ;
4 for  $i \leftarrow 1$  to  $|M|$  do
5    $d \leftarrow |p_i - p_j|$ ;
6    $distances.append(d)$ 
7 end
8  $distances.sort()$ ;
9  $distances \leftarrow distances.First3Items()$ ;
10  $j \leftarrow \text{random}(1, 3)$ ;
11  $m_2 \leftarrow AST_j$ ;
/* Update memory scores */
12 for  $i \leftarrow 1$  to  $|M|$  do
13   if  $i = j$  then
14      $T_i = 0$ ;
15      $M_i \leftarrow c$ ;
16   end
17   else
18      $T_i \leftarrow T_i + 1$ ;
19      $M_i \leftarrow \alpha \log(T_i) + c$ ;
20   end
21 end

```

---

metric. SANet [8] and AdaIN [5] excel in style matching, as shown by their strong performance on ArtFID [23] and GRAM loss [2]. NST [2] excels in content preservation, as indicated by its performance on content loss and GRAM loss. StyTr<sup>2</sup> [11] demonstrates a balanced performance across both content and style metrics. Notably, ArtFusion [14] achieves outstanding performance on SSIM, LPIPS, and content loss. However, its lower performance on ArtFID than other methods warrants further investigation. This discrepancy might be attributed to FID, a core component of ArtFID that measures the distance between image distributions. Unlike other methods directly using the content image, ArtFusion operates by sampling from Gaussian noise conditioned on content and style. This sampling approach could lead to minor distribution discrepancies between generated and original content images, potentially impacting ArtFID scores.

Our experimental results reveal a trade-off between content preservation and the fidelity of the style transfer. We found that the CNN-based SANet with attentional layer [8] and the CNN-based AdaIN [5] stand out for their low ArtFID and GRAM Loss, indicating their ability to capture artistic nuances. However, their performance on content preservation metrics like LPIPS and SSIM is relatively weaker, suggesting a potential compromise in maintaining the original image's fidelity. Conversely, the diffusion-based ArtFusion [14] and the manifold-based MAST [20] excel in SSIM, LPIPS, and Content Loss, demonstrating strong content preservation, but their lower performance for ArtFID, SQRE\_OV and GRAM losses shows challenges in faithfully reproducing the target style. A similar, albeit more nuanced,

	SSIM	LPIPS	Content Loss	GRAM	SRQE	ArtFID	Overall	AST Ranking
StyTr <sup>2</sup>	1	3	5	4	1	1	1	1
UCAST	3	4	7	9	5	2	2	2
Artflow	4	6	6	7	4	4	3	3
SANet	7	9	9	3	3	3	4	4
NST	8	7	1	1	2	6	5	5
ArtFusion	5	2	3	8	8	7	6	6
AdaIN	9	8	8	2	6	5	6	7
MAST	2	1	2	10	10	9	8	8
LST	6	5	4	5	7	8	9	9
WCT	10	10	10	6	9	10	10	10

Fig. 5: Ranking performance of AST methods on SSIM, LPIPS, Content Loss, GRAM Loss, SRQE\_OV and ArtFID with overall ranking.

trade-off is observed for UCAST [22].

The transformer-based StyTr<sup>2</sup> [60] appears to strike a notable balance between these competing aspects. Compared to the attention-based SANet, it exhibits improved content preservation while maintaining satisfactory style transfer performance, potentially mitigating the “content leak” issue observed in SANet.

#### 4.2 The Impact of Fine-Grained Factors on AST Performance

In this section, we delve deeper into the intricate interplay between specific image characteristics and their influence on AST effectiveness. We focus on four factors:

- *Scene diversity*: each image is annotated with its primary scene type, such as nature landscape, indoor, or urban.
- *Object complexity*: used the number of annotated instances per image to represent object complexity.
- *Salient regions*: the area ratio of the two main objects in the image. The polygon annotations by ADE20K enable an accurate area calculation for each object instance.
- *Content-style consistency*: the semantic alignment between content and style images in AST. Content image semantics are defined by ADE20K, and style image content is recognized by RAM [28].

To comprehensively evaluate the performance of the AST methods using the metrics introduced in Section 3.4, we analyze their performance across diverse scene types and object complexity levels within the images. To assess

the level of salient region preservation and content-style consistency within the content images, we utilize ArtFID [23]. As detailed in Section 3.4, this robust metric combines the Fréchet Inception Distance (FID) [37] for evaluating image distribution alignment with LPIPS [29] for measuring content preservation. A lower ArtFID score indicates a superior AST performance, signifying a successful integration of style and content while maintaining high perceptual fidelity.

**A. Scene diversity** We analyze SSIM, LPIPS content loss, GRAM loss, SRQE\_OV and ArtFID values across various scenes in our custom dataset. This analysis reveals how different content types influence the performance of AST methods.

We first investigated the performance of different AST methods across various scene categories, as visualized in Figure 6. (In the radar charts, all metrics have been transformed to follow a “larger is better” principle for visual clarity. This includes inverting the values of LPIPS, content loss, GRAM loss, and ArtFID, which originally indicated better performance with lower values.) StyTr<sup>2</sup> demonstrated exceptional performance on all scene types based on the SSIM and SRQE\_OV metrics. For the LPIPS metric, MAST achieved the best results on *natural landscapes*, *shopping and dining*, and *cultural scene* types, while ArtFusion excelled in *sports and leisure*, *urban*, and *transportation* scenes. Optimization-based NST dominated most scene types when considering content loss and style loss metrics. SANet achieved the best performance on most scene types using the ArtFID metric, while StyTr<sup>2</sup> excelled in *natural landscapes* and *sports and leisure*.

In the subsequent analysis, we explored the performance across distinct scene categories, as depicted in Figure 7. Scenes with *natural landscapes*, typically featuring fewer objects and more prominent salient regions, demonstrated enhanced performance in terms of ArtFID, SSIM, content loss, and GRAM loss metrics. These results are in concordance with intuitive human judgments. Conversely, these *natural landscape* scenes displayed the weakest performance in the LPIPS metric. This observation highlights the potential for bias when solely relying on individual metrics for AST evaluation.

In summary, the performance of AST methods depends on the scenes and the metric used. Some AST methods are better at matching the style of the target image, such as SANet and AdaIN, while others do a better job of keeping the content of the original image intact, such as ArtFusion. Relying on single metrics can be misleading, so it is important to consider using multiple metrics for evaluation. Scene types with fewer objects and larger prominent areas, like wide-open *nature landscapes*, generally achieved higher performance across most evaluation metrics. In contrast, indoor scenes with cluttered objects, such as *home or hotel*, *shopping and dining* tended to perform less favorably. This disparity might be attributed to the inherent challenges of stylizing complex scenes with numerous intricate details, potentially overwhelming the AST methods.

**B. Object complexity** To explore the influence of object complexity on AST effectiveness, we leveraged the number of objects in each image as a representative measurement of complexity. We categorized these images into five distinct groups based on object count:

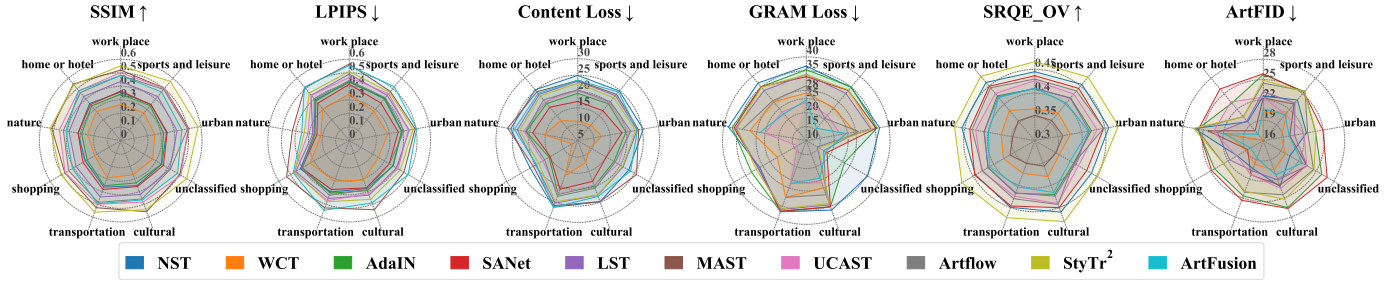


Fig. 6: AST methods performance on SSIM, LPIPS, Content Loss, Gram Loss SRQE\_OV and ArtFID across different scenes. In the radar charts, all metrics have been transformed to a “larger is better” scale for visual consistency.

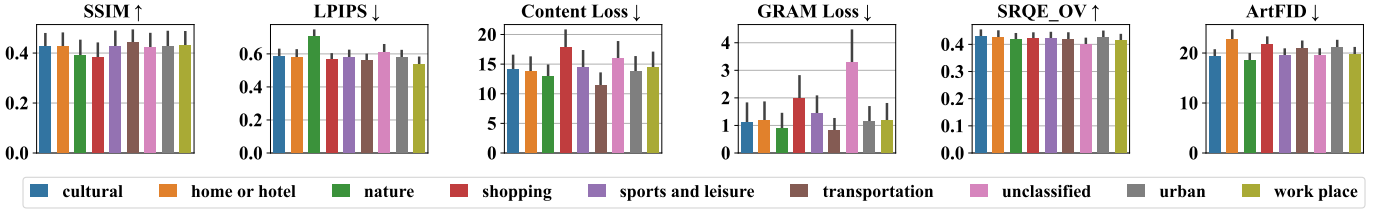


Fig. 7: SSIM, LPIPS, Content Loss, Gram Loss, SRQE\_OV and ArtFID across different scenes.

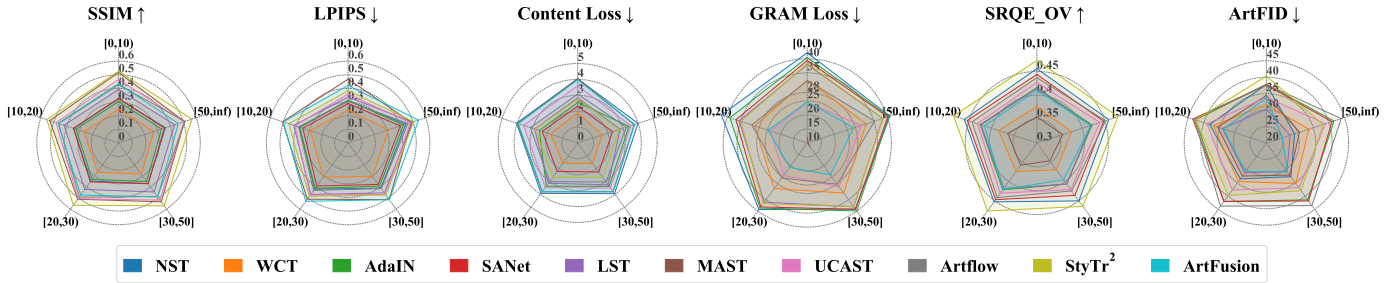


Fig. 8: AST methods performance on SSIM, LPIPS, Content Loss, Gram Loss SRQE\_OV and ArtFID across different levels of object complexity. In the radar charts, all metrics have been transformed to a “larger is better” scale for visual consistency.

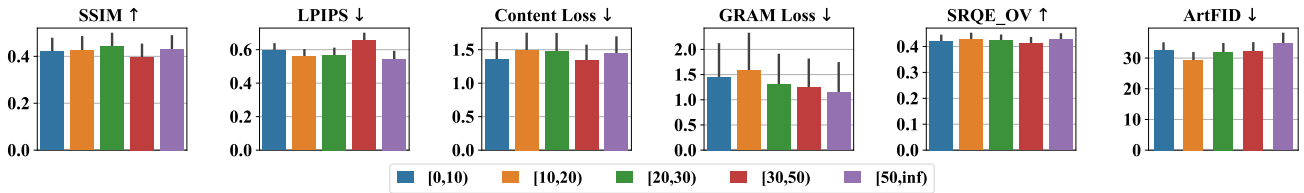


Fig. 9: SSIM, LPIPS, Content Loss, Gram Loss, SRQE\_OV and ArtFID across different levels of object complexity.

- Group [0, 10): Images containing less than 9 objects.
- Group [10, 20): Images containing 10 to 19 objects.
- Group [20, 30): Images containing 20 to 29 objects.
- Group [30, 50): Images containing 30 to 49 objects.
- Group [50, inf): Images containing more than 50 objects.

This grouping strategy aligns with the object count distribution within the ADE20K dataset, facilitating a balanced representation of different object density levels.

The performance of the AST methods for the different object counts is shown in Figure 8. The performance is consistent with the overall performance shown in Figure 5. There is a content-style trade-off in AST methods. Compared

to the evaluation of different scene type groups, the stylization methods show high consistency in performance for each evaluation metric and each object count group.

Figure 9 analyzes the correlation between the average performance of the AST methods and the number of objects in an image. Contrary to intuition, the results indicate that a minimal number of objects does not necessarily lead to superior performance. Images with 10 to 20 objects achieved the best average performance on ArtFID, SRQE\_OV and LPIPS, metrics associated with style matching and perceptual fidelity. Interestingly, images with 50 or more objects achieved the best average performance on GRAM loss, a metric measuring style similarity. This divergence highlights the potential for bias inherent in different metrics,

as further explored in Section 5.4, which examines how each metric relates to human perception.

In conclusion, images with high object complexity exhibit superior performance in terms of GRAM loss and the LPIPS metric. Images with medium object complexity achieve the best results on SSIM (Group [30,50]), content loss (Group [30,50]), SRQE\_OV(Group [10,20]) and ArtFID (Group [10,20]). Further investigation is warranted to elucidate the correlation between object count and these metrics, assessing the potential implications for specific applications.

**C. Salient regions of the Image**

To examine how variations in salient regions influence style transfer performance, we segmented our dataset into five groups based on the relative area occupied by salient regions within each image. These groups are:

- Group [0,0.4): salient regions cover less than 40% of the image area.
- Group [0.4,0.5): salient regions cover between 40% and 50% of the image area.
- Group [0.5,0.6): salient regions cover between 50% and 60% of the image area.
- Group [0.6,0.7): salient regions cover between 60% and 70% of the image area.
- Group [0.7,1]: salient regions cover 70% or more of the image area.

This categorization ensures a balanced distribution across groups, reflecting the overall distribution of salient region sizes in our dataset. This approach allows us to isolate and analyze the effect of different levels of salient region prominence on the performance of diverse style transfer methods. Figure 10 shows how the salient region size of an image impacts the ArtFID score.

From AST methods: SANet and StyTr<sup>2</sup> perform well. One key factor contributing to this observed trend is the ability of attention-based models to focus and refine stylistic features when dealing with smaller main areas. In scenarios where the main subject or focal point occupies a modest portion of the image, attention-based methods demonstrate a heightened precision in capturing and replicating the stylistic elements onto the content image.

Regarding the main area, we observed that methods tend to exhibit better performance when the main area falls in the group [0.6, 0.7) instead of a larger main area. This implies that lower image complexity leads to better performance. This finding challenges conventional expectations and prompts a deeper examination into the dynamics of style transfer in relation to image composition.

**D. Content-Style Consistency**

This analysis is underpinned of three distinct groups delineated by cosine similarity’s range: [0,0.9), [0.9,0.95) and [0.95,1]. Our analysis reveals a positive correlation between content-style consistency and the performance of AST methods. As Figure 11 illustrates, images with more consistent content and style elements (e.g., both depicting portraits) tend to achieve higher performance on ArtFID metrics (except from NST and ArtFusion). This suggests that aligning the content and style domains can be a valuable strategy for optimizing AST quality. For example, utilizing a portrait artistic style for a content image containing a person can leverage the inherent similarities between the domains,

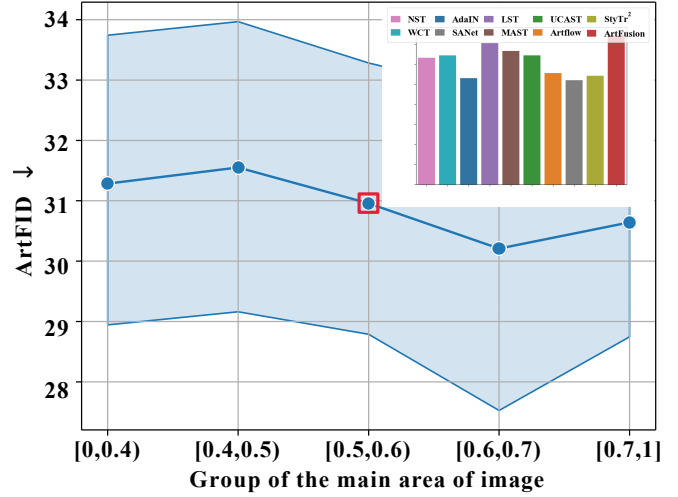


Fig. 10: Impact of **salient region** size on AST performance. This bar plot visualizes the ArtFID scores achieved by different AST methods when the salient region occupies 50% to 60% of the image area.

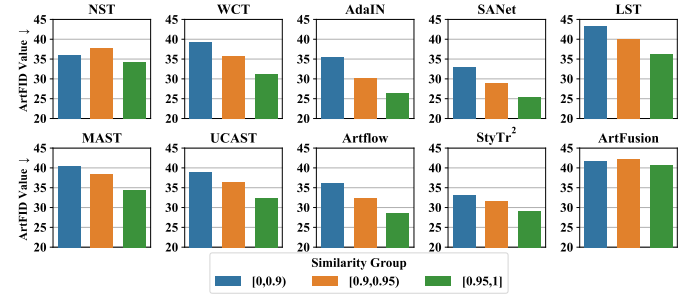


Fig. 11: The relationship between ArtFID scores and **content-style similarity**: lower ArtFID scores are observed in AST results as the content and style images exhibit greater semantic similarity.

potentially leading to a more faithful and aesthetically pleasing style transfer.

Our findings reveal a crucial factor influencing the success of AST: the content-style consistency between the content and style images. As illustrated in Figure 12, the [0.95, 1] consistency group consistently exhibits the best performance across most AST methods. This suggests that when the content and style images share a high degree of semantic similarity, the AST process can achieve more accurate and visually pleasing results, as shown in Figure 13.

**4.3 Analyzing a Specific Style**

We evaluated the performance of the AST method on a diverse set of ten styles: *abstract expressionism, art nouveau, cubism, expressionism, impressionism, pop art, realism, romanticism, symbolism, and ukiyo-e*. We measured the stylized results using all five metrics described in Section 4.1 and obtained an overall ranking evaluation.

The overall ranking distribution of the AST method across these styles aligns closely with the overall performance ranking presented in Section 4.1. As illustrative examples,



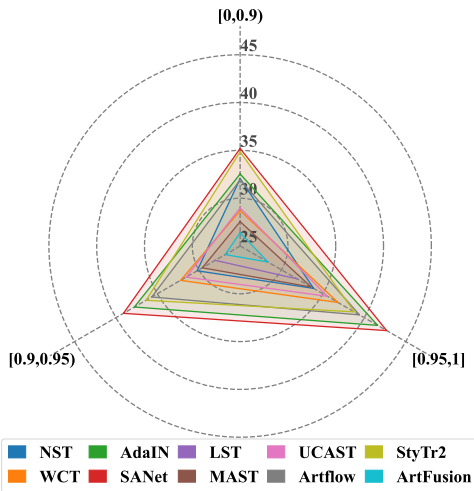


Fig. 12: AST performance across different **content-style similarity** levels.

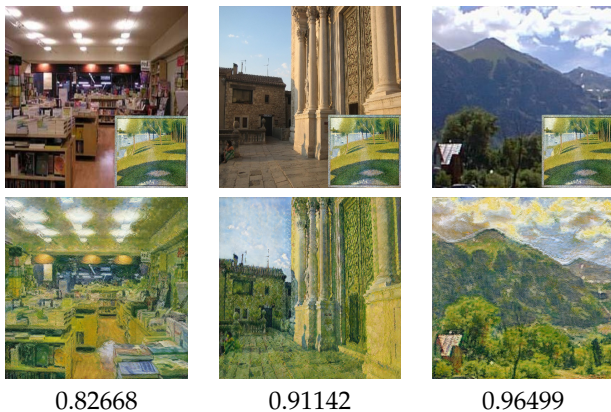


Fig. 13: This figure presents stylized images generated by UCAST [22] belonging to different content/style similarity groups. The annotations at the bottom represent the calculated similarity between the content and style images. Notably, the higher semantic similarity between content and style images leads to better-stylized performance.

we present renderings in the *abstract expressionism* and *impressionism* styles (Figure 14).

## 5 ANALYSIS OF SUBJECTIVE EXPERIMENTS

To gain a comprehensive understanding of user perception, we conducted a multi-granularity subjective study encompassing point-wise, pair-wise, and group-wise user evaluations. A total of 55 participants were recruited for the subjective study, with 43 having a background in computer graphics and 25 possessing knowledge of AST research, indicating a degree of artistic appreciation ability. Approximately 3500 valid responses were collected, including 700 point-wise, 2100 pair-wise, and 700 group-wise questions. The overall results of this study are presented in Figure 15, with a detailed analysis provided in the subsequent sections.

### 5.1 Analysis of Point-Wise Questions

In our evaluation of different AST methods, we meticulously considered user ratings as a crucial metric for assessing the subjective satisfaction and preference of individuals engaging with stylized images.

The user ratings, indicative of individual perceptions and preferences, exhibited a substantial degree of variability as shown in Figure 16. The wide distribution suggests diverse opinions and responses to the stylization outputs generated by different AST methods. This variability could be attributed to varying artistic tastes, subjective interpretations of visual aesthetics, and the diverse nature of content and style images used in the evaluation process.

In the diverse landscape of user ratings, two AST methods, LST and UCAST, demonstrated remarkable performance. Their median values were significantly higher than those of other methods, and their distributions were concentrated in the high mean opinion score (MOS) zone. ArtFusion’s scores are also aggregated at a high zone. However, its median value is not as good as LST and UCAST’s. In terms of MOS scores, StyTr<sup>2</sup> and Artflow has a high median value. However, their MOS score distribution is not as concentrated, indicating their robustness is not as good.

The success of UCAST, StyTr<sup>2</sup>, and ArtFusion in garnering superior average scores may be attributed to several factors. These methods may exhibit enhanced capabilities in preserving the artistic essence of the style image, achieving better visual fidelity, and adapting more seamlessly to diverse content images. Additionally, user-friendly interfaces, faster processing times, or other user-centric features could contribute to the positive reception of these methods.

### 5.2 Analysis of Pair-Wise Questions

To evaluate the performance of each AST method, we analyzed user responses in three key aspects as the setting in [26]: content preservation (CP), style similarity (SS), and overall effect (OV). We tracked the *win rate*, the percentage of times one method outperformed the other in direct comparisons. Subsequently, the collected *win rate* were standardized to a zero mean, facilitating visual representation in Figure 17.

The results reveal three outstanding performers: CNN-based UCAST [22] (contrastive learning), transformer-based StyTr<sup>2</sup> [11], and diffusion-based ArtFusion [14]. ArtFusion excels in content preservation, StyTr<sup>2</sup> shines in style similarity, and UCAST achieves a balanced performance across both. NST surpasses the average in both content preservation and style similarity, suggesting a more stable optimization process compared to AdaIN, SANet, and LST. ArtFusion shines in content preservation, leading to its strong performance in overall visual effect, second only to UCAST. This highlights the significant contribution of content preservation to the overall visual experience. These findings align with the objective study presented in Section 4.1.

Users often struggle to choose between two stylistically transferred images when evaluating their relative quality. However, they find it easier to assess whether both images achieve a desired level of quality, regardless of which method was used. In response to this observation, we introduce the



Fig. 14: Stylized images generated with *abstract expressionism* and *impressionism* styles. The red numbers below each image indicate the overall ranking performance of the corresponding AST method.

TABLE 4: Pair-Wised User Study Results. CP, SS, and OV are the *Win Rate* for content preservation, style similarity, and overall visual effect, respectively. BGCP, BGSS, and BGOV are the *Both Good Rate* for content preservation, style similarity, and overall visual effect.

Categories	AST	Content Preserve		Style Similarity		Overall Visual Effect	
		CP	BGCP	SS	BGSS	OV	BGOV
CNN-based	NST [2]	44.15%	64.89%	42.29%	54.65%	44.15%	47.67%
	WCT [3]	4.31%	35.48%	18.10%	13.64%	8.05%	10.91%
	AdaIN [5]	28.03%	56.04%	35.84%	40.79%	30.64%	41.24%
	LST [21]	34.59%	88.98%	35.95%	70.64%	36.49%	63.55%
	SANet [8]	25.58%	57.95%	38.79%	53.95%	30.46%	41.98%
	MAST [20]	44.35%	58.93%	34.46%	41.82%	40.40%	25.81%
	UCAST [22]	<b>58.97%</b>	<b>89.53%</b>	<b>58.01%</b>	<b>82.81%</b>	<b>61.22%</b>	<b>68.75%</b>
Flow-based	Artflow [9]	27.66%	79.22%	35.37%	57.75%	30.85%	49.25%
Transformer-based	StyTr <sup>2</sup> [11]	52.38%	73.27%	53.70%	53.06%	53.44%	47.37%
Diffusion-based	ArtFusion [14]	<b>69.33%</b>	<b>90.65%</b>	46.32%	61.70%	58.28%	42.59%

*both good rate* metric, which quantifies the proportion of users who choose “both are good” when comparing stylistically transferred images perceived as similar in quality (both “good” or “bad”).

We make statistics on *win rate* and *both good rate* from the three aspects of CP (content preservation), SS (style similarity), and OV (overall visual effect), respectively. As shown in Figure 18, UCAST exhibits a significant advantage in achieving “both good” performance across all three aspects, evidenced by the remarkably thick flows for “Both Good CP”, “Both Good SS”, and “Both Good OV”. ArtFusion demonstrates strong performance in content preservation (very thick “Both Good CP” flow compared to its thin “Both Bad CP” flow), but its performance in stylistic similarity and overall visual quality is ordinary.

LST and Artflow exhibit a high *both good rate* but low *win rate*. This suggests they’re consistently decent across all aspects but lack the edge to consistently outperform other methods in direct comparisons. UCAST, StyTr<sup>2</sup>, and ArtFusion, on the other hand, demonstrate the ability to clearly outshine their competitors. The entire detailed result is shown in Table 4.

### 5.3 Analysis of Group-Wise Questions

To evaluate the realism of AST results, we designed two user study metrics within the “Identification Question (group-wise)” group. The proposed metrics assess the

TABLE 5: Group-Wised Subjective Study Results.

Categories	AST	PO% ↓	PC% ↓
CNN-based	NST [2]	72.01%	76.15%
	WCT [3]	86.39%	82.29%
	AdaIN [5]	82.32%	85.91%
	LST [21]	71.72%	79.18%
	SANet [8]	79.03%	82.18%
	MAST [20]	77.26%	77.26%
	UCAST [22]	<b>54.86%</b>	72.81%
Flow-based	Artflow [9]	70.98%	77.19%
Transformer-based	StyTr <sup>2</sup> [11]	61.69%	71.97%
Diffusion-Based	ArtFusion [14]	<b>51.91%</b>	<b>67.64%</b>

effectiveness of AST in generating visually deceptive stylistic transformations that confound human perception:

- *pick correct rate*: the ratio of stylized images that are correctly detected as artificial (correctly chose “yes” when presented with a stylized image).
- *pick out rate*: the ratio of fake images that users can identify correctly out of all the fake images.

Low values for both *pick correct rate* and *pick out rate* indicate strong AST performance, suggesting that the stylized images closely resemble real photographs and are thus difficult for humans to identify as artificial. Detailed results can be found in Table 5.

To investigate the interplay between user performance and the proportion of fake images in identification questions, we varied the number of stylized images presented



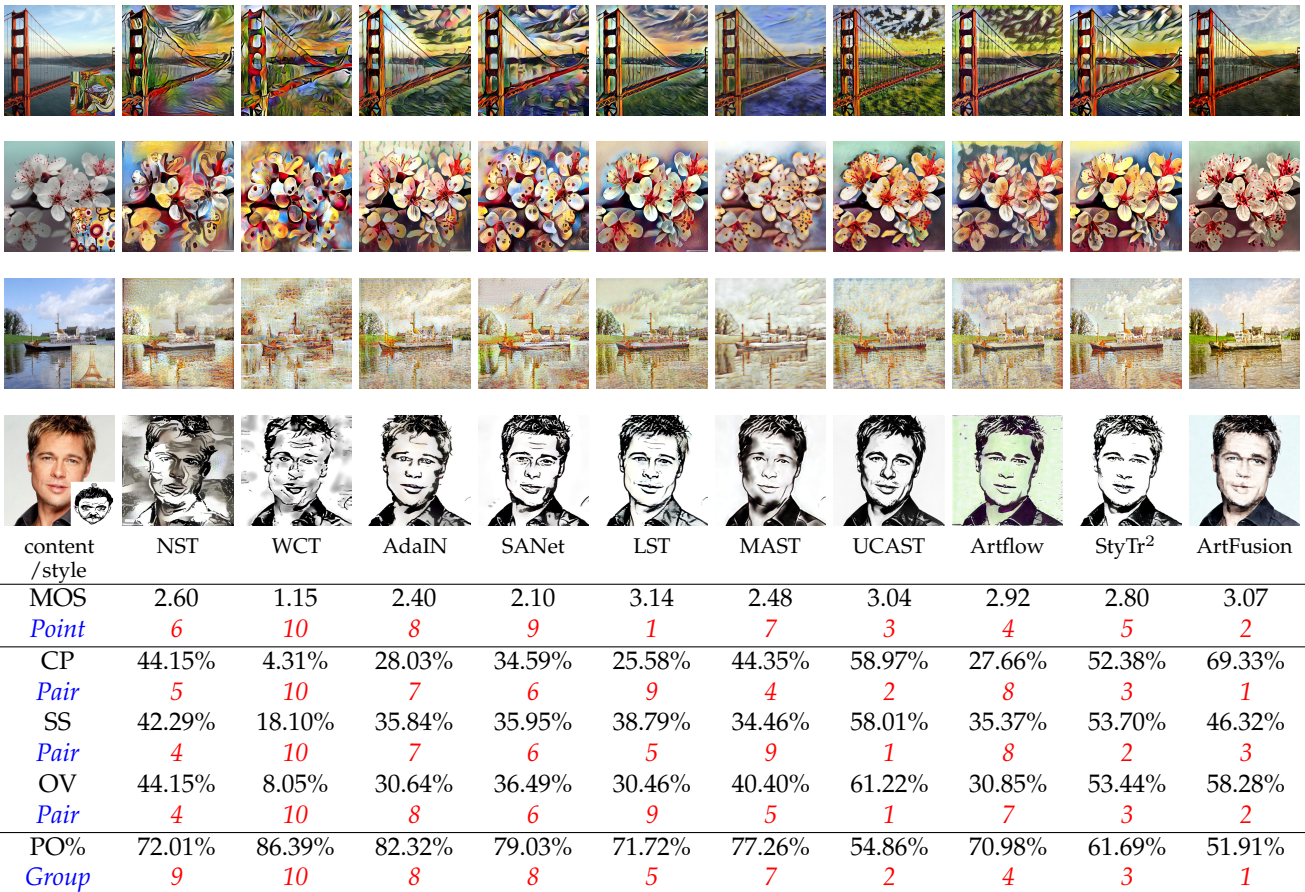


Fig. 15: The provided images showcase the outputs of different AST methods. CP, SS, and OV denote the *win rate* (Section 5.2) for content preservation, style similarity, and overall visual effect, respectively. PO refers to the *pick out rate* (Section 5.3). Notably, CNN-based AdaIN and SANet (with attention layer) exhibit high fidelity in replicating the target style. While Diffusion-based ArtFusion excels at content preservation, its artistic expression may be less pronounced. CNN-based UCAST (with contrast learning) and Transformer-based StyTr<sup>2</sup> achieve a compromise, maintaining content quality while incorporating stylistic elements.

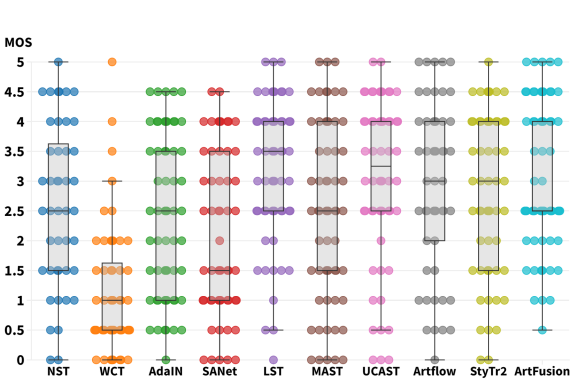


Fig. 16: Point-wise user study results: the scatter plot displays the distribution of AST MOS. Scores range from 0 to 5 in increments of 0.5. The line across the box highlights the median score, with the bottom and top edges indicating the values where 25% of ratings fall below and above, respectively.

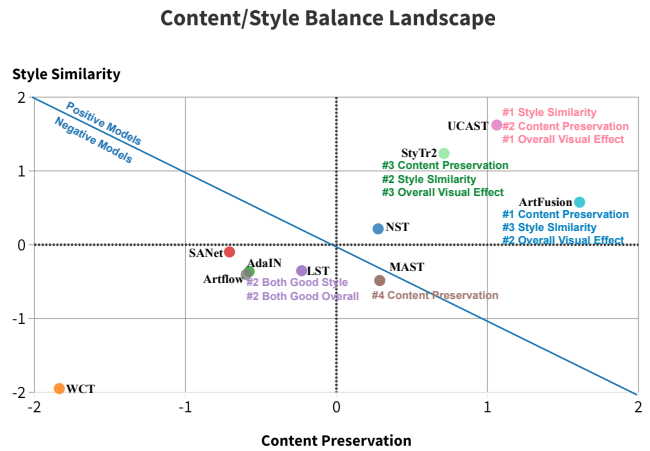


Fig. 17: The performance of AST methods was evaluated on pair-wise questions in terms of content preservation (CP) and style similarity (SS). The upper-right quadrant of the plot indicates methods that performed well on both metrics.

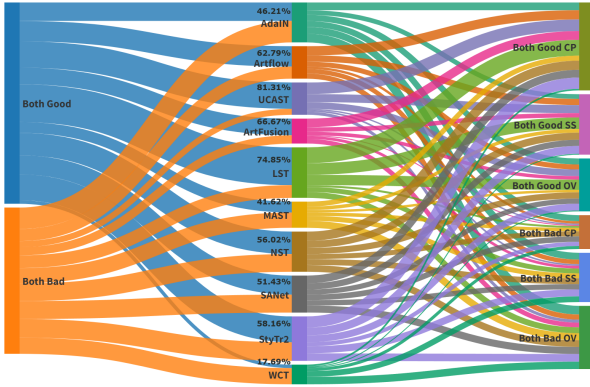


Fig. 18: A comprehensive sankey plot visualizing the *both good rate* of AST: the width of each flow represents the relative frequency of images falling into each performance category.

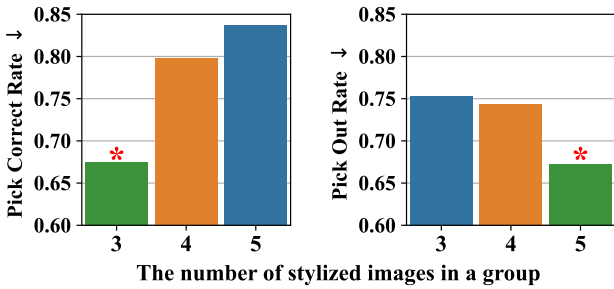


Fig. 19: The *pick out rate* ↓ and *pick correct rate* ↓ for different numbers of stylized images (3, 4, and 5) in each group. A total of ten images are included in each group.

(3, 4, and 5) per question, creating three distinct groups with different fake image ratios. Figure 19 depicts the resulting pick correct rate and pick out rate for each group. Our analysis reveals a trade-off relationship between these metrics. While the *pick correct rate* increased with more stylized images (facilitating the identification of individual fakes), the *pick out rate* conversely decreased (making it harder to identify all fakes within the group). This suggests that for the group-wise method to effectively assess the realism of AST, employing a smaller number of images per group offers a more balanced and informative approach.

### 5.4 Investigating Fine-Grained Influences on AST Performance

Building upon the group-wise analysis, we further investigate the specific image characteristics that may influence AST performance. To achieve this, we categorize the images used in the group-wise study into distinct semantic groups based on pre-defined criteria detailed in Section 4.2. This allows us to visualize and statistically analyze the impact of these fine-grained factors on the user’s ability to identify fake stylized images, as measured by the *pick out rate*.

Figure 20 reveals a fascinating interplay between image characteristics and the *pick out rate*. Distinct trends emerge for groups based on object count, salient region, and content-style consistency.

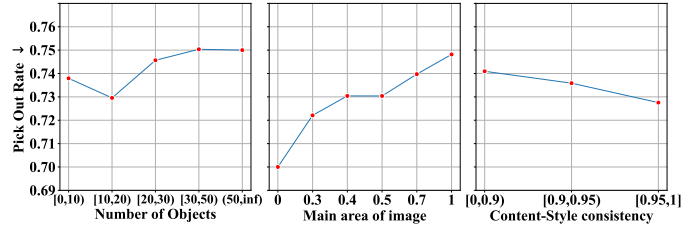


Fig. 20: The *pick out rate* ↓ varies when images from different semantic groups, as described in Section 4.2.

**Object Complexity.** Notably, the *pick out rate* peaks for images with 10 to 20 objects, declining as the object count increases. This aligns with the objective findings presented in Section 4.2, where ArtFID metric scores correlate with human judgment. This suggests that ArtFID, which measures the distance between image distributions, captures aspects of perceptual fidelity that resonate with human evaluation, particularly for images with moderate object complexity.

**Salient Region.** The main area of the image reveals a departure from the objective metrics observed in Section 4.2. Contrary to expectations, the *pick out rate* decreases (meaning users perform better at identifying fakes) as the main area of the image shrinks, and this correlation is evident.

One potential explanation for this phenomenon is that larger main areas pose greater challenges for AST methods in preserving the details of prominent objects, as shown in Figure 21. This could lead to more noticeable artifacts or inconsistencies in the stylized image, particularly around the main object. These discrepancies might become easier for users to identify, contributing to the higher *pick out rate* for images with larger main areas.

**Content-style Consistency.** Our findings resonate with both the objective study presented in Section 4.2 and human intuition: as the semantic similarity between content and style images increases, the *pick out rate* decreases, indicating improved AST performance. This relationship is visually evident in the right part of Figure 20, where the lowest *pick out rate* aligns with higher content-style consistency scores.

### 5.5 Analyze the Consistency of Group-Wise User Study

To compare the consistency of the group-wise user study and pair-wise user study, we measure their reliability, which is a measurement of how consistently a method measures something.

Our research focused on assessing the Test-Retest reliability of user study results conducted through pair-wise, point-wise, and group-wise evaluation methods. Test-retest reliability measures the consistency of results by repeating the same test on the same sample at a different point in time. We apply this method by extracting specific questions answered by the same user at least twice time and comparing the results of the two answers.

Following our review mechanism in Section 3.5.2, we chose user responses that had the same questions and came from the same participants, enabling a rigorous analysis of the consistency in their evaluations across various time points. To measure the degree of agreement among the replies to repeated questions, we applied the Cohen Kappa






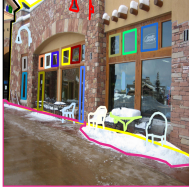


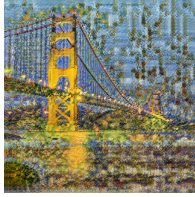
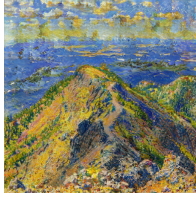

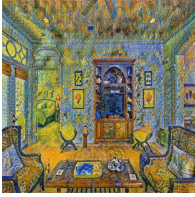
Target Style	Parsing					
	Stylized Images					
Salient Region Percentage		56.31%	56.09%	40.87%	33.24%	24.17%

Fig. 21: A group of stylized images generated by UCAST [22]. When the salient region of an image is larger, it becomes more straightforward for users to spot artifacts, leading to easier identification of the image as stylized or fake. Conversely, when the salient region is smaller, the artifacts become relatively harder to discern.

TABLE 6: Cohen’s kappa results. We use it to evaluate the user study results at three levels of granularity: Point-Wise, Pair-Wise, and Group-Wise.

Method	Cohen’s Kappa
<b>Group-Wise</b>	<b>0.5778</b>
Pair-Wise content preservation	0.4732
Pair-Wise style similarity	0.5726
Pair-Wise overall visual effect	0.4933
Point-Wise	0.5172

coefficient to estimate the Test-retest reliability for each level of question.

The results are presented in Table 6. The group-wise evaluations demonstrated a remarkable Cohen Kappa value of 0.5778. This exceptional result indicates a substantial agreement among participants, showcasing the method’s great performance in capturing consistent user opinions over time. The higher Cohen Kappa values, especially in group-wise and style-focused pair-wise assessments, signify a notable degree of reliability in user evaluations.

The variations in Cohen Kappa values across different evaluation methods suggest that the reliability of assessments may be influenced by the specific aspects being evaluated (content, style, overall impression). Group-wise assessments demonstrate a robust ability to consistently capture user opinions.

## 5.6 Factors Influencing User’s Preference

Having gathered both subjective and objective evaluation results, and considering that these metrics can be broadly categorized into measures of content preservation and style

matching abilities, we aimed to explore how these aspects contribute to user preference for stylized images.

We assume the following metrics as the user’s preference for stylized images: point-wise metric *MOS*, pair-wise metrics *win rate*, *both good rate*, and group-wise metric *pick out rate*. From each of the selected subjective metric, we are able to gain a ranking performance of AST methods. Figure 22 shows the Kendall rank correlation coefficients (KRCC, denoted by  $\tau$ ) and Spearman’s rank correlation coefficients (SRCC, denoted by  $\rho$ ) between the AST rankings derived from selected subjective metrics.

We observe that *MOS* and *pick out rate* have the strongest positive correlation with *both good rate* of content preservation (BGCP) ( $\tau = 0.87, \rho = 0.95$ ). Similarly, *win rate* of overall visual effect (OV) has the highest correlation with *win rate* of content preservation (CP) ( $\tau = 0.87, \rho = 0.96$ ). These findings highlight the strong affinity between user preference for stylized images and content preservation metrics. While style matching metrics demonstrate a considerable correlation with user preference, this association is notably weaker than the correlation observed with content preservation metrics.

Figure 23 examines the Kendall’s  $\tau$  and Spearman’s  $\rho$  coefficients between the AST rankings derived from the objective metrics presented in Section 4.1 and the subjective user preferences. We observe that *MOS* exhibits a stronger positive correlation with content preservation metrics compared to style matching metrics. With SRQE\_CP, the correlation coefficients are ( $\tau = 0.56, \rho = 0.76$ ), followed by ( $\tau = 0.51, \rho = 0.62$ ) with LPIPS and ( $\tau = 0.38, \rho = 0.49$ ) with content loss. Conversely, the correlation with SRQE\_SR loss is negative ( $\tau = -0.47, \rho = -0.54$ ), and that with GRAM loss is ( $\tau = -0.2, \rho = -0.33$ ). Similar trends are observed for OV and *pick out rate*. Our analysis reveals a

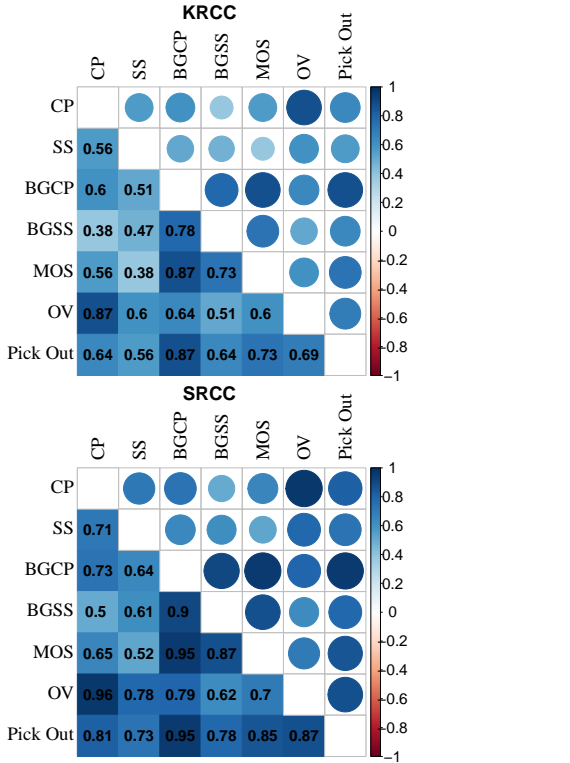


Fig. 22: Kendall rank correlation coefficient (KRCC) and Spearman’s rank correlation coefficient (SRCC) were employed to assess the consistency between subjective metrics. The figure illustrates that content preservation exerts a stronger influence on user preference compared to style similarity.

noteworthy emphasis on content preservation in subject user preferences that the perceived success of AST methods is more closely linked to the retention of content features than the adherence to the target style.

Additionally, the MOS exhibits an extremely high correlation with the *both good rate* of content preservation (BGCP) ( $\tau = 0.87, \rho = 0.95$ ) and the *both good rate* of style similarity (BGSS) ( $\tau = 0.73, \rho = 0.87$ ), demonstrating the meaningfulness of *both good rate* as a subjective metric.

### 5.7 Consistency between Subjective and Objective Study

To assess the validity of both subjective and objective studies, we first obtained the AST rankings for all metrics included in the study, encompassing both objective and subjective metrics. Next, we calculated the Kendall’s  $\tau$  and Spearman’s  $\rho$  coefficients between these rankings, as shown in Figure 23.

Regarding point-wise granularity, the Mean Opinion Score (MOS) has the highest ( $\tau = 0.56, \rho = 0.76$ ) with SRQE\_CP, indicating that users rate the stylized image higher if its content is similar to the original image. Therefore, point-wise questions may be biased towards AST methods that generate better image quality.

As for pair-wise granularity, the *win rate* for both content preservation (CP) and overall visual (OV) effect demonstrate

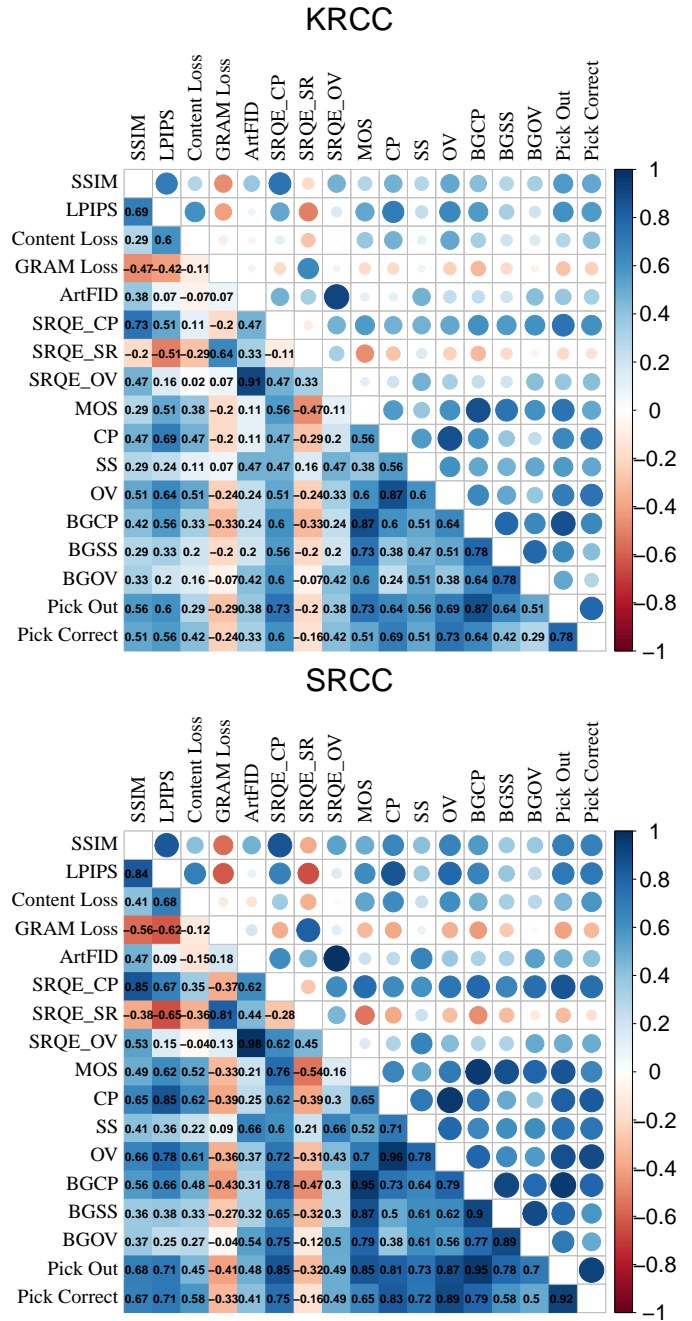


Fig. 23: Kendall rank correlation coefficient (KRCC) and Spearman’s rank correlation coefficient (SRCC) were used to assess consistency between objective and subjective metrics. LPIPS correlates with *win rate*, SRQE\_CP with *both good rate*, *pick out rate*, and *pick correct rate*.

the strongest correlation with LPIPS, with coefficients of ( $\tau = 0.69, \rho = 0.85$ ) and ( $\tau = 0.64, \rho = 0.78$ ), respectively. Additionally, the *both good rate* for CP, SS, and OV (BGCP, BGSS, BGOV) show the highest correlation with SRQE\_CP, with coefficients of ( $\tau = 0.6, \rho = 0.78$ ), ( $\tau = 0.56, \rho = 0.65$ ), and ( $\tau = 0.6, \rho = 0.75$ ) respectively. This suggests that LPIPS effectively captures the comparative preference reflected in win rates, while SRQE\_CP aligns with the perception of overall quality indicated by *both good rate*. These findings

reveals that LPIPS and SRQE\_CP can partially represent human judgment in AST evaluation. The *win rate* of style similarity (SS) exhibits a weak correlation with almost all objective metrics, except for the correlation coefficients of ( $\tau = 0.47, \rho = 0.66$ ) with ArtFID and SRQE\_OV, which suggests that ArtFID and SRQE do provide a quantitative measurement of the artistic style of stylized images to some extent. However, the evaluation of artistic features cannot fully correspond with human perception.

At the group-wise granularity level, the *pick out rate* and the *pick correct rate* exhibit the strongest correlations with SRQE\_CP ( $\tau = 0.73, \rho = 0.85$ ) and ( $\tau = 0.76, \rho = 0.75$ ), respectively. This result indicates a close correspondence between SRQE\_CP and subjective metrics, further supporting the notion that content preservation significantly influences user preference. Both the *pick out rate* and the *pick correct rate* demonstrate correlations with LPIPS, content loss, SSIM, SRQE, and ArtFID, as well as with MOS, *win rate*, and *both good rate*. These correlations validate the reliability and relevance of these two subjective metrics.

## 6 CONCLUSION

In this study, we present a comprehensive framework for evaluating Arbitrary Style Transfer (AST) performance by combining a rigorous objective study with a multifaceted subjective study. Objectively, we evaluated AST performance across multiple metrics by analyzing its variations across diverse image groups based on scene type, object complexity, salient regions, and content-style consistency. Subjectively, we employed a group-wise user study, complemented by point-wise and pair-wise studies, resulting in a multi-granular evaluation. We introduced three metrics - *Both Good Rate, Pick Out Rate, and Pick Correct Rate* - to obtain robust and nuanced user feedback. Furthermore, we explored the consistency between objective and subjective results, investigating the factors influencing human perception of AST quality and the reliability of various metrics in reflecting human judgment.

Although our research offers a valuable contribution to the field of AST assessment, we acknowledge potential avenues for future refinement and improvement.

- 1) Given the observed misalignment between objective and subjective metrics in AST, designing an objective metric that perfectly aligns with human perception remains a challenge.
- 2) We employ Spearman's rank correlation coefficient (SRCC) and Kendall's rank correlation coefficient (KRCC) to represent the consistency between objective and subjective metrics. These coefficients measure the consistency between rankings, which may not necessarily reflect the precise correlation between the two sets of metrics.
- 3) Future research may focus on optimizing AST methods, particularly in mitigating the identified limitations and weaknesses, as well as aligning more closely with user preferences.

## REFERENCES

- [1] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, "Neural style transfer: A review," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 11, pp. 3365–3385, 2019.
- [2] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2414–2423.
- [3] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," *Advances in neural information processing systems*, vol. 30, 2017.
- [4] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "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 (CVPR)*, July 2017.
- [5] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501–1510.
- [6] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images," *arXiv preprint arXiv:1603.03417*, 2016.
- [7] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," *arXiv preprint arXiv:1610.07629*, 2016.
- [8] D. Y. Park and K. H. Lee, "Arbitrary style transfer with style-attentional networks," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5880–5888.
- [9] J. An, S. Huang, Y. Song, D. Dou, W. Liu, and J. Luo, "Artflow: Unbiased image style transfer via reversible neural flows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 862–871.
- [10] K. Zhu, Z. Tian, R. Luo, and X. Mao, "Styleflow: Disentangle latent representations via normalizing flow for unsupervised text style transfer," *arXiv preprint arXiv:2212.09670*, 2022.
- [11] Y. Deng, F. Tang, W. Dong, C. Ma, X. Pan, L. Wang, and C. Xu, "Stytr2: Image style transfer with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 326–11 336.
- [12] J. Yoo, Y. Uh, S. Chun, B. Kang, and J.-W. Ha, "Photorealistic style transfer via wavelet transforms," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9036–9045.
- [13] N. Dai, J. Liang, X. Qiu, and X. Huang, "Style transformer: Unpaired text style transfer without disentangled latent representation," *arXiv preprint arXiv:1905.05621*, 2019.
- [14] D.-Y. Chen, "Artfusion: Arbitrary style transfer using dual conditional latent diffusion models," *arXiv preprint arXiv:2306.09330*, 2023.
- [15] J. Jeong, M. Kwon, and Y. Uh, "Training-free style transfer emerges from h-space in diffusion models," *arXiv preprint arXiv:2303.15403*, 2023.
- [16] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," *arXiv preprint arXiv:2208.01626*, 2022.
- [17] Z. Wang, L. Zhao, and W. Xing, "Styldiffusion: Controllable disentangled style transfer via diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7677–7689.
- [18] S. Yang, H. Hwang, and J. C. Ye, "Text-guided diffusion image style transfer with contrastive loss fine-tuning," 2022.
- [19] Y. Zhang, N. Huang, F. Tang, H. Huang, C. Ma, W. Dong, and C. Xu, "Inversion-based style transfer with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 146–10 156.
- [20] J. Huo, S. Jin, W. Li, J. Wu, Y.-K. Lai, Y. Shi, and Y. Gao, "Manifold alignment for semantically aligned style transfer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 861–14 869.
- [21] X. Li, S. Liu, J. Kautz, and M.-H. Yang, "Learning linear transformations for fast image and video style transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [22] Y. Zhang, F. Tang, W. Dong, H. Huang, C. Ma, T.-Y. Lee, and C. Xu, "A unified arbitrary style transfer framework via adaptive contrastive learning," *ACM Transactions on Graphics*, 2023.
- [23] M. Wright and B. Ommer, "Artfid: Quantitative evaluation of neural style transfer," in *Pattern Recognition: 44th DAGM German Conference, DAGM GCPR 2022, Konstanz, Germany, September 27–30, 2022, Proceedings*. Springer, 2022, pp. 560–576.
- [24] S. Majumdar, A. Bhoi, and G. Jagadeesan, "A comprehensive comparison between neural style transfer and universal style transfer," *arXiv preprint arXiv:1806.00868*, 2018.



- [25] Z. Wang, L. Zhao, H. Chen, Z. Zuo, A. Li, W. Xing, and D. Lu, "Evaluate and improve the quality of neural style transfer," vol. 207, p. 103203.
- [26] H. Chen, F. Shao, X. Chai, Y. Gu, Q. Jiang, X. Meng, and Y.-S. Ho, "Quality evaluation of arbitrary style transfer: subjective study and objective metric," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.
- [27] H. Chen, F. Shao, X. Chai, Q. Jiang, X. Meng, and Y.-S. Ho, "Collaborative learning and style-adaptive pooling network for perceptual evaluation of arbitrary style transfer," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [28] Y. Zhang, X. Huang, J. Ma, Z. Li, Z. Luo, Y. Xie, Y. Qin, T. Luo, Y. Li, S. Liu *et al.*, "Recognize anything: A strong image tagging model," *arXiv preprint arXiv:2306.03514*, 2023.
- [29] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.
- [30] Q. Cai, M. Ma, C. Wang, and H. Li, "Image neural style transfer: A review," *Computers and Electrical Engineering*, vol. 108, p. 108723, 2023.
- [31] A. Singh, V. Jaiswal, G. Joshi, A. Sanjeev, S. Gite, and K. Kotecha, "Neural style transfer: A critical review," *IEEE Access*, vol. 9, pp. 131 583–131 613, 2021.
- [32] Q. Shang, L. Hu, Q. Li, W. Long, and L. Jiang, "A survey of research on image style transfer based on deep learning," in *2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture (AIAM)*. IEEE, 2021, pp. 386–391.
- [33] L. Liu, Z. Xi, R. Ji, and W. Ma, "Advanced deep learning techniques for image style transfer: a survey," *Signal Processing: Image Communication*, vol. 78, pp. 465–470, 2019.
- [34] C. Li, Z. Zhang, H. Wu, W. Sun, X. Min, X. Liu, G. Zhai, and W. Lin, "Agiqa-3k: An open database for ai-generated image quality assessment," *arXiv preprint arXiv:2306.04717*, 2023.
- [35] J. Wang, H. Duan, J. Liu, S. Chen, X. Min, and G. Zhai, "Aigciqa2023: A large-scale image quality assessment database for ai generated images: from the perspectives of quality, authenticity and correspondence," in *CAAI International Conference on Artificial Intelligence*. Springer, 2023, pp. 46–57.
- [36] Y. Zhang, Y. Zhang, Y. Yao, J. Jia, J. Liu, X. Liu, and S. Liu, "Unlearncanvas: A stylized image dataset to benchmark machine unlearning for diffusion models," *arXiv preprint arXiv:2402.11846*, 2024.
- [37] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.
- [38] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 694–711.
- [39] J. Jeong, M. Kwon, and Y. Uh, "Training-free style transfer emerges from h-space in diffusion models," *arXiv preprint arXiv:2303.15403*, 2023.
- [40] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
- [41] H. Zhang and K. Dana, "Multi-style generative network for real-time transfer," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.
- [42] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Diversified texture synthesis with feed-forward networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3920–3928.
- [43] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "Stylebank: An explicit representation for neural image style transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1897–1906.
- [44] X.-C. Liu, Y.-C. Wu, and P. Hall, "Painterly style transfer with learned brush strokes," *IEEE Transactions on Visualization and Computer Graphics*, 2023.
- [45] B. Sheng, P. Li, C. Gao, and K.-L. Ma, "Deep neural representation guided face sketch synthesis," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 12, pp. 3216–3230, 2018.
- [46] Y. Zhang, F. Tang, W. Dong, H. Huang, C. Ma, T.-Y. Lee, and C. Xu, "Domain enhanced arbitrary image style transfer via contrastive learning," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1–8.
- [47] G. Ghiasi, H. Lee, M. Kudlur, V. Dumoulin, and J. Shlens, "Exploring the structure of a real-time, arbitrary neural artistic stylization network," *arXiv preprint arXiv:1705.06830*, 2017.
- [48] M.-M. Cheng, X.-C. Liu, J. Wang, S.-P. Lu, Y.-K. Lai, and P. L. Rosin, "Structure-preserving neural style transfer," *IEEE Transactions on Image Processing*, vol. 29, pp. 909–920, 2019.
- [49] X. Luo, Z. Han, L. Yang, and L. Zhang, "Consistent style transfer," *arXiv preprint arXiv:2201.02233*, 2022.
- [50] X. Luo, Z. Han, and L. Yang, "Progressive attentional manifold alignment for arbitrary style transfer," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 3206–3222.
- [51] Y. Wang, "An arbitrary style transfer network based on dual attention module," in *2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, vol. 4. IEEE, 2021, pp. 1221–1226.
- [52] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.
- [53] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [55] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [56] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8162–8171.
- [57] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.
- [58] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.
- [59] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5775–5787, 2022.
- [60] Y. Deng, F. Tang, W. Dong, H. Huang, C. Ma, and C. Xu, "Arbitrary video style transfer via multi-channel correlation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1210–1217.
- [61] H. Chen, L. Zhao, Z. Wang, H. Zhang, Z. Zuo, A. Li, W. Xing, and D. Lu, "Dualast: Dual style-learning networks for artistic style transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 872–881.
- [62] Z. Zheng and J. Liu, "P<sup>2</sup>-gan: efficient style transfer using single style image," *arXiv preprint arXiv:2001.07466*, 2020.
- [63] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1–10.
- [64] Y. Shu, R. Yi, M. Xia, Z. Ye, W. Zhao, Y. Chen, Y.-K. Lai, and Y.-J. Liu, "Gan-based multi-style photo cartoonization," *IEEE Transactions on Visualization and computer graphics*, vol. 28, no. 10, pp. 3376–3390, 2021.
- [65] W. Xu, C. Long, R. Wang, and G. Wang, "Drb-gan: A dynamic resblock generative adversarial network for artistic style transfer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6383–6392.
- [66] W. Song, X. Jin, S. Li, C. Chen, A. Hao, and X. Hou, "Finestyle: Semantic-aware fine-grained motion style transfer with dual interactive-flow fusion," *IEEE Transactions on Visualization and Computer Graphics*, 2023.
- [67] Y. Shi, P. Liu, S. Chen, M. Sun, and N. Cao, "Supporting expressive and faithful pictorial visualization design with visual style transfer," *IEEE Transactions on Visualization & Computer Graphics*, vol. 29, no. 01, pp. 236–246, 2023.
- [68] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641.
- [69] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity,"



*IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

- [70] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, “Understanding and predicting image memorability at a large scale,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2390–2398.



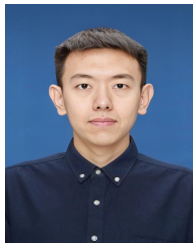
**Juan Cao** received the PhD degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2008. She is a professor with the Institute of Computing Technology, Chinese Academy of Sciences. Her research interests include multimedia content analysis and fake multimedia detection. She has more than 90 publications in international journals and conferences, including IEEE TKDE, TIP, KDD, CVPR, etc.



**Zijun Zhou** received the B.Sc degree in computer science from Jilin University in 2022. He is now pursuing the M.E. degree at the School of Artificial Intelligence, Jilin University. His research interests include computer graphics, computer vision, and machine learning.



**Weiming Dong** (Member, IEEE) is a Professor at the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences. He received his BSc and MSc degrees in 2001 and 2004, both from Tsinghua University, China. He received his PhD in Computer Science from the University of Lorraine, France, in 2007. His research interests include image synthesis, image recognition, and computational creativity.



**Fan Tang** (Member, IEEE) received the B.Sc. degree in computer science from North China Electric Power University, Beijing, China, in 2013, and the Ph.D. degree from Institute of Automation, Chinese Academy of Sciences, Beijing, in 2019. He is an Assistant Professor with the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include computer graphics, computer vision, and machine learning.



**Xiangtao Li** (Senior Member, IEEE) received the B.Eng., M.Phil., and Ph.D. degrees in computer science from the Northeast Normal University, Changchun, China, in 2009, 2012, and 2015, respectively. He is currently a Professor with the School of Artificial Intelligence, Jilin University, Changchun. He has published over 100 technical papers in prominent journals such as *Nature Communications*, *Advanced Science*, *IEEE TCYB*, *AAAI*, *IEEE TCBB*, *IEEE TNB*, *IEEE TEM*, etc.



**Yuxin Zhang** received B.Sc. degree in Automation from Tsinghua University, Beijing, China, in 2020. She is now a Ph.D. candidate of Institute of Automation, Chinese Academy of Sciences, and the School of Artificial Intelligence at the University of Chinese Academy of Sciences. Her research interests include computer vision, computer graphics, and machine learning.



**Oliver Deussen** graduated at Karlsruhe Institute of Technology and is now a professor of visual computing at the University of Konstanz (Germany). He is one of the speakers of the Excellence Cluster “Centre for the Advanced Study of Collective Behavior” and vice speaker of the SFB Transregio “Quantitative Methods for Visual Computing”. He is President of the Eurographics Association and served as Co-editor-in-chief of the Computer Graphics Forum from 2012 to 2015.



**Tong-Yee Lee** (Senior Member, IEEE) received the PhD degree in computer engineering from Washington State University, Pullman, in May 1995. He is currently a chair professor with the Department of Computer Science and Information Engineering, National Cheng-Kung University, Tainan, Taiwan. He leads the Computer Graphics Group, Visual System Laboratory, National Cheng-Kung University (<http://graphics.csie.ncku.edu.tw>). His current research interests include computer graphics, nonphotorealistic rendering, medical visualization, virtual reality, and media resizing. He is a senior member of the IEEE Computer Society and a member of the ACM. He also serves on the editorial boards of the *IEEE Transactions on Visualization and Computer Graphics*.