Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 6 Iss. 4 October 2026, pp: 2002-2013
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Systematic Review of Arbitrary Neural Style Transfer: Eight Deep Learning Approaches Compared

Dwika Lovitasari Yonia[1*], Naufal Zahir Rizqullah[2], Sigit Mulyanto[3], Bambang Sutejo[4]

[1]Department of Information Systems, Faculty of Computer Science, Universitas Darwan Ali, Sampit, Indonesia
[2]Department of Software Engineering, SMKN 4 Tanjungpinang, Tanjungpinang, Indonesia
[3,4]Departement of Management, Faculty of Business, Universitas Darwan Ali, Sampit, Indonesia

E-Mail: [1]dwika.lovitasari.yonia@gmail.com, [2]naufalzahirr@yahoo.com, [3]sigitmul@gmail.com, [4]tejosampit@gmail.com

Received May 31th 2026; Revised Jun 22th 2026; Accepted Aug 30th 2026; Available Online Oct 30th 2026
Corresponding Author: Dwika Lovitasari Yonia
Copyright © 2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

Neural style transfer has become an important computer vision technique for combining the semantic content of one image with the visual style of another. Arbitrary style transfer extends this capability by allowing a trained model to process previously unseen styles without retraining for each reference style. However, comparative evidence concerning architectural mechanisms, evaluation metrics, computational requirements, and reproducibility remains fragmented. This study presents a directed, systematic literature review of eight representative, arbitrary neural style transfer methods: AdaIN, SANet, MCCNet, ArtFlow, AdaAttN, IEContraAST, CSBNet, and RAST. The review protocol follows PRISMA 2020 principles and includes research questions, database searching, eligibility criteria, quality assessment, data extraction, and comparative synthesis. The findings show that normalization-based methods offer efficient inference, attention-based methods improve local style adaptation, flow-based and restoration-based methods address content leakage, contrastive learning strengthens style representation, and video-oriented methods emphasize temporal consistency. Recent transformer- and diffusion-based studies further improve global feature modeling and controllability but generally incur higher computational costs. This review contributes an architectural taxonomy, an evaluation metric map, a reproducibility analysis, and recommendations for standardized image and video benchmarks that integrate objective metrics, computational measurements, and human preference evaluation.

Keywords: Arbitrary Style Transfer, Computer Vision, Deep Learning, Neural Style Transfer, Systematic Literature Review

1. INTRODUCTION

The rapid advancement of artificial intelligence and computer vision has strengthened deep learning's role in creative image processing, particularly in image synthesis, visual manipulation, and artistic image stylization. Neural style transfer has become an important technique because it can combine the structural content of one image with the artistic characteristics of another image, such as texture, color distribution, brush strokes, and visual patterns, while maintaining the main semantic structure of the original image [1]. Recent studies also show that neural style transfer has evolved beyond conventional CNN-based methods toward Transformer-based, contrastive learning, implicit representation, and diffusion-based approaches, which improve controllability, content preservation, style diversity, and visual quality of generated images [2]. Therefore, neural style transfer provides significant opportunities for graphic design, digital art, animation, social media content production, fashion visualization, and other creative industries because it supports the automated generation of visually appealing images while still allowing users to control artistic expression and content structure [3], [4].

In neural style transfer, the content image supplies the object arrangement, semantic structure, and spatial composition that should remain recognizable, whereas the style image provides the artistic appearance to be transferred. Early optimization-based methods generated visually attractive results by iteratively optimizing an output image to balance content and style representations [5]. Although effective, iterative optimization is computationally expensive and unsuitable for many real-time applications [1], [6]. Feed-forward transformation networks subsequently improved efficiency through perceptual training objectives, but

many early models were restricted to one style or a limited set of predefined styles [7]. Arbitrary style transfer addresses this limitation by applying previously unseen styles during inference without model retraining.

AdaIN became an influential normalization-based baseline because it aligns the channel-wise mean and variance of content features with those of style features [8]. Despite its efficiency, global statistic alignment may not adequately capture complex local style patterns or spatial relationships. Attention-based methods such as SANet and AdaAttN subsequently improved local content-style correspondence [9], [10]. ArtFlow and RAST addressed content leakage through reversible flow and restoration mechanisms [11], [12], while IEContraAST introduced internal-external learning and contrastive objectives to produce more harmonious style representations [13]. MCCNet and CSBNet extended arbitrary style transfer to video by emphasizing temporal consistency and frame-to-frame stability [14], [15].

Recent developments also indicate a transition toward Vision Transformer and diffusion-based architectures. Transformer-based methods model long-range dependencies between content and style features, which can improve global structure and semantic alignment [2], [14], [16]. Diffusion-based approaches formulate stylization as a controlled iterative generation process and may produce visually rich and flexible outputs [17], [18]. Nevertheless, these advances often introduce greater training complexity, memory consumption, and inference latency. Consequently, architectural progress should be evaluated not only through visual quality but also through computational cost, reproducibility, and deployment feasibility.

Although numerous arbitrary neural style transfer methods have been introduced, cross-study comparison remains difficult because studies frequently use different datasets, resolutions, baselines, hardware configurations, and evaluation protocols [6], [19]. Metrics such as SSIM, LPIPS, FID, ArtFID, content loss, style loss, temporal loss, and user preference capture different dimensions of stylization quality and cannot be interpreted as interchangeable measures [20]. In addition, public source code does not automatically guarantee reproducibility because repositories may lack checkpoints, environment files, preprocessing instructions, random seeds, or evaluation scripts [21], [22].

The novelty of this study lies in a focused systematic comparison of eight representative arbitrary neural style transfer methods that reflect normalization-based, attention-based, flow-based, contrastive learning-based, restoration-based, and video-oriented architectural directions. The review compares their mechanisms, training requirements, evaluation practices, strengths, limitations, computational implications, and implementation availability. It also places transformer-based and diffusion-based approaches within the emerging research landscape and proposes a standardized benchmark framework for fair comparison across image and video style transfer tasks.

2. METHOD

This study uses a directed Systematic Literature Review approach to compare representative arbitrary neural style transfer methods. The review procedure includes research design, research questions, database searching, eligibility assessment, PRISMA reporting, quality assessment, data extraction, and comparative synthesis, as shown in Figure 1. The reporting protocol follows the principles of PRISMA 2020, particularly in ensuring transparency throughout the article identification, screening, eligibility assessment, and inclusion processes.

2.1. Research Design

This study adopts a directed systematic literature review design because the principal units of analysis were determined in advance: AdaIN, SANet, MCCNet, ArtFlow, AdaAttN, IEContraAST, CSBNet, and RAST. These methods were selected to represent distinct architectural directions rather than to claim exhaustive coverage of every arbitrary style transfer model. Each core study was required to present an identifiable methodological contribution, sufficient architectural or experimental information, and an official scientific publication.

The study is descriptive and comparative rather than meta-analytic. Statistical effect sizes were not calculated because the reviewed studies use heterogeneous datasets, image resolutions, hardware configurations, evaluation metrics, and experimental protocols. Instead, evidence was synthesized through architectural classification, metric mapping, quality assessment, reproducibility analysis, and narrative comparison. The reporting procedure follows PRISMA 2020 principles to improve transparency in article identification, screening, eligibility assessment, and inclusion [23].

2.2. Research Questions

Five research questions that guide the systematic literature review and ensure that the search, selection, extraction, and synthesis processes remain aligned with the study objectives. RQ1 focuses on the development of arbitrary neural style transfer methods across the reviewed publication period, while RQ2 examines the architectural categories and transfer mechanisms represented by the eight selected methods. RQ3 addresses the datasets, evaluation metrics, and experimental settings used in previous studies, whereas RQ4 evaluates the strengths and limitations of each method in terms of content preservation, style fidelity, temporal consistency,

computational efficiency, deployment feasibility, and reproducibility. Finally, RQ5 identifies the remaining research gaps and future opportunities, particularly those related to transformer-based and diffusion-based approaches, standardized benchmarks, and unified evaluation protocols. Therefore, the research questions presented in Table 1 provide a structured framework for organizing the comparative analysis and interpreting the findings of this review.

Table 1. Research questions

Code	Description
RQ1	How have arbitrary neural style transfer methods developed across the reviewed publication period?
RQ2	What architectural categories and transfer mechanisms are represented by the eight selected methods?
RQ3	What datasets, evaluation metrics, and experimental settings are used to assess arbitrary neural style transfer?
RQ4	What are the strengths and limitations of each method in terms of content preservation, style fidelity, temporal consistency, computational efficiency, deployment feasibility, and reproducibility?
RQ5	What research gaps and future opportunities remain, particularly regarding transformer-based and diffusion-based methods, standardized benchmarks, and unified evaluation protocols?

2.3. Literature Search and Study Selection Strategy

The literature search covered studies published between 2017 and 2025, beginning with the publication period of AdaIN and extending to the most recent studies available during the review process. Searches were conducted through IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, CVF Open Access, AAAI, IJCAI, NeurIPS, arXiv, and Google Scholar. GitHub was used only to verify implementation availability, pretrained checkpoints, and supporting materials, and was not treated as a primary scientific publication source.

The primary Boolean search string was: ("arbitrary neural style transfer" OR "arbitrary style transfer") AND (AdaIN OR SANet OR AdaAttN OR ArtFlow OR MCCNet OR IEContraAST OR CSBNet OR RAST). A supplementary search was performed using: ("arbitrary style transfer") AND (transformer OR "Vision Transformer" OR diffusion). Search terms were adjusted to the syntax of each database while preserving the same conceptual coverage. The scientific review covered publications from 2017 to 2025, while repository availability and implementation status were rechecked during the final revision.

2.4. PICOS Framework and Eligibility Criteria

The PICOS framework used to define the scope and eligibility criteria of this systematic literature review shown in Table 2. The population component covers studies on arbitrary neural style transfer for images or videos in computer vision, while the intervention component includes the eight core deep learning methods together with relevant transformer-based and diffusion-based approaches. The comparison component focuses on architectural design, transfer mechanisms, datasets, evaluation metrics, computational characteristics, strengths, limitations, and reproducibility. The expected outcomes include an architectural taxonomy, comparative synthesis, evaluation metric map, identification of research gaps, and benchmark recommendations. Finally, the study design component limits the evidence base to peer-reviewed journal articles, conference papers, recognized preprints, comparative studies, and official repositories used only as supporting implementation evidence. The PICOS framework was applied to define the population, intervention, comparison, outcome, and study design used during literature identification and selection.

Table 2. PICOS framework for article selection

Component	Criteria	Description
P	Population/ Problem	Studies concerning arbitrary neural style transfer for images or videos in computer vision.
I	Intervention	The eight core methods and supporting transformer-based, attention-based, or diffusion-based approaches.
C	Comparison	Architecture, transfer mechanism, dataset, metrics, strengths, limitations, computational characteristics, deployment feasibility, and reproducibility.
O	Outcome	Architectural taxonomy, comparative synthesis, metric map, research gaps, and benchmark recommendations.
S	Study Design	Full-text journal articles, conference papers, recognized preprints, and comparative studies; repositories are used only as supporting implementation evidence.

Studies were included when they: (1) addressed arbitrary or universal neural style transfer; (2) provided sufficient information about architecture, training, datasets, or evaluation; (3) represented one of the eight core methods or supported the analysis of recent architectural developments; (4) were available as full-

text journal articles, conference papers, or recognized preprints; and (5) were written in English. Studies were excluded when they: (1) focused exclusively on single-style transfer without direct relevance to arbitrary transfer; (2) duplicated another publication; (3) were blogs, tutorials, or non-technical online articles; (4) lacked sufficient methodological or experimental detail; or (5) consisted of unsupported repositories without an identifiable scientific publication. Full-text eligibility decisions followed predefined criteria consistent with protocol-oriented systematic review practice [24].

2.5. Study Selection and PRISMA Reporting

The directed identification process combined method-name searching, database searching, citation tracking, and recent-topic searching. The finalized and auditable evidence corpus contained 29 unique scientific publications. Eight publications were designated as core method studies, while 21 supporting studies provided evidence concerning reviews, evaluation metrics, transformer and diffusion developments, systematic review procedures, and reproducibility shown in Figure 1.

After manual record management, 29 unique scientific records remained and no duplicate record was retained. All 29 publications were screened and assessed in full text against the eligibility criteria. No additional report was excluded during the final corpus audit because only eligible records were retained in the consolidated evidence set. The qualitative synthesis therefore included 29 scientific publications, comprising eight core method studies and 21 supporting studies. Eight implementation repositories were examined separately and were not counted as scientific publications. Figure 1 summarizes this PRISMA-based evidence selection flow.

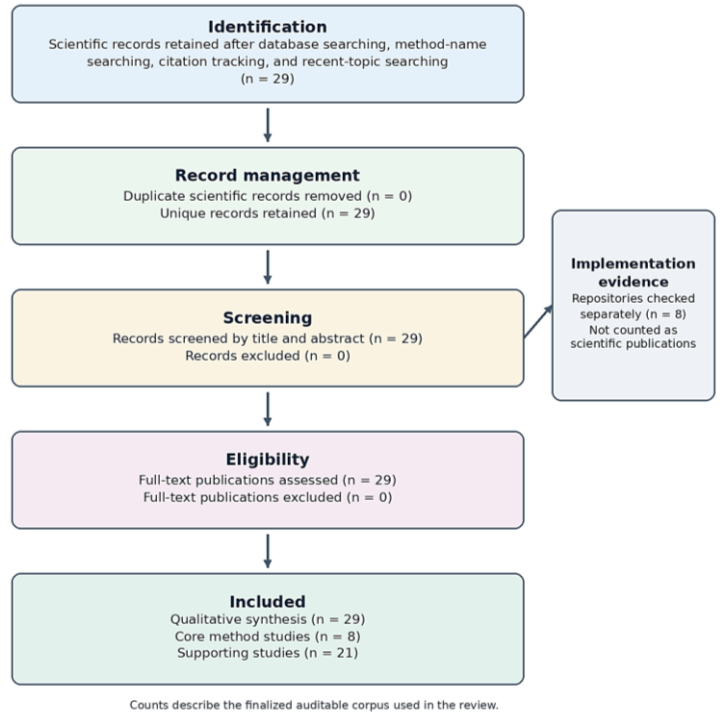


Figure 1. PRISMA-Based Flow Diagram of The Finalized 29-Publication Evidence Corpus

2.6. Quality Assessment and Data Extraction

Quality assessment was used to ensure that the reviewed studies provided sufficient evidence for comparative synthesis. Six criteria evaluated research objectives, architecture, datasets, evaluation metrics, limitations, and supporting materials shown in Table 3. Each item was scored as 0 when not reported, 0.5 when partially reported, and 1 when clearly reported.

Table 3. Quality Assessment Checklist

Code	Quality criterion	Score
QA1	The objective and methodological contribution are explicitly explained.	0, 0.5, or 1
QA2	The architecture and transfer mechanism are adequately described.	0, 0.5, or 1
QA3	The dataset and experimental scenario are reported.	0, 0.5, or 1
QA4	Objective or subjective evaluation metrics are explained.	0, 0.5, or 1
QA5	Methodological limitations can be identified from the paper.	0, 0.5, or 1

Code	Quality criterion	Score
QA6	The scientific paper is traceable and reproducibility materials, such as code, checkpoints, or implementation documentation, are available.	0 to 1

The total quality score was obtained by summing QA1 to QA6, producing a maximum score of 6. A minimum score of 4.0 was used as the inclusion threshold for the core synthesis. As shown in Table 4, all eight core studies met this threshold. QA5 was scored as 0.5 when limitations could be inferred from experiments or discussion but were not presented in a dedicated limitation statement. QA6 was scored as 0.5 for AdaAttN because the verified PyTorch repository is a supporting reimplementation rather than an author-official release, and for SANet because the official project page explicitly states that the source code will not be released, although an unofficial implementation is available.

Table 4. Quality Assessment Results for The Eight Core Studies.

Method	QA1	QA2	QA3	QA4	QA5	QA6	Total	Decision
AdaIN	1	1	1	1	0.5	1	5.5	Included
SANet	1	1	1	1	0.5	0.5	5.0	Included
MCCNet	1	1	1	1	0.5	1	5.5	Included
ArtFlow	1	1	1	1	0.5	1	5.5	Included
AdaAttN	1	1	1	1	0.5	0.5	5.0	Included
IEContraAST	1	1	1	1	0.5	1	5.5	Included
CSBNet	1	1	1	1	0.5	1	5.5	Included
RAST	1	1	1	1	0.5	1	5.5	Included

3. RESULTS AND DISCUSSION

This section presents the characteristics of the reviewed literature, summarizes the eight core methods, compares their architectural contributions, examines emerging transformer and diffusion directions, maps evaluation practices, assesses reproducibility and computational requirements, and identifies research gaps.

3.1. Characteristics of the Reviewed Literature

The eight core methods represent six major architectural directions. AdaIN represents normalization-based stylization; SANet and AdaAttN represent attention-based transfer; ArtFlow represents flow-based modeling; IEContraAST represents contrastive learning; MCCNet and CSBNet represent video-oriented style transfer; and RAST represents restoration-based stylization. This classification illustrates how arbitrary style transfer has evolved from global statistical alignment toward more adaptive mechanisms for local correspondence, content preservation, style representation, and temporal consistency.

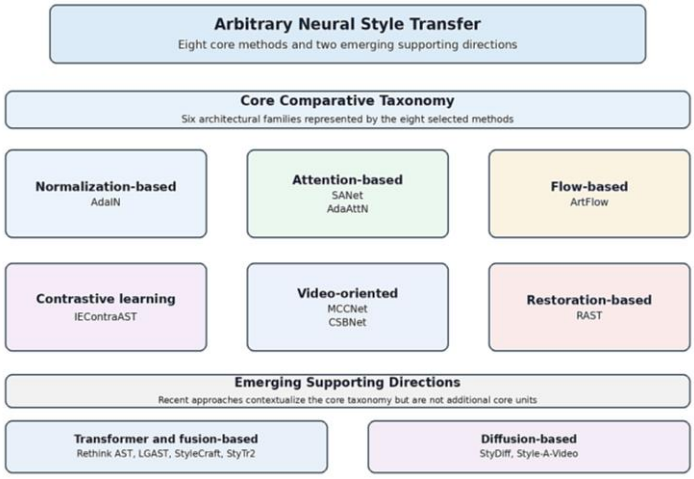


Figure 2. Taxonomy of Arbitrary Neural Style Transfer Methods Reviewed in This Study

The core methods were published between 2017 and 2023, while supporting literature extends the analysis to recent evaluation frameworks, transformer-based architectures, diffusion-based models, and reproducibility practices. The core set is therefore used for direct architectural synthesis, whereas newer approaches are discussed as emerging directions that broaden the research landscape without changing the eight primary units of analysis. Figure 3 separates the six categories represented by the eight core methods from transformer-based and diffusion-based approaches, which are treated as contextual emerging directions rather than additional core units.

The Figure 2 also shows that each method is reviewed based on its architectural focus and contribution to solving common problems in neural style transfer, such as content leakage, weak style representation, temporal inconsistency, and limited evaluation consistency [25]. RAST is included because it explicitly treats style transfer as a restorable process to reduce content leakage and preserve content information [26]. Video-oriented methods are also important because video style transfer must handle temporal consistency, flickering, and content stability across frames [6]. In addition to the eight core papers, supporting literature on neural style transfer foundations, transformer-based stylization, contrastive learning, evaluation methods, and SLR guidelines is used to strengthen the context without shifting the main focus from the selected methods.

3.2. Summary of the Eight Core Arbitrary Neural Style Transfer Methods

Table 5 summarizes the publication year, venue, architectural category, and principal contribution of each core method. The year column refers to the original peer-reviewed publication year, not the repository creation date or the latest GitHub update. Therefore, AdaIN remains 2017 and IEContraAST remains 2021 even when their repositories are updated later. RAST is listed as WACV 2023 because the conference paper is the core study; its extended ACM journal version appeared in 2024.

Table 5. Summary of The Eight Core Arbitrary Neural Style Transfer Methods.

Method	Year	Venue	Category	Main focus
AdaIN	2017	ICCV	Normalization-based	Aligns the channel-wise mean and variance of content and style features for efficient arbitrary transfer [7].
SANet	2019	CVPR	Attention-based	Uses style attention and multi-level feature embeddings for local semantic and spatial style integration [8].
MCCNet	2021	AAAI	Video correlation	Uses multi-channel correlation to fuse content and style features while supporting temporal coherence [14].
ArtFlow	2021	CVPR	Flow-based	Uses reversible neural flows and unbiased feature transfer to reduce content leakage [28].
AdaAttN	2021	ICCV	Attention-normalization	Computes per-point style statistics through adaptive attention normalization [9].
IEContraAST	2021	NeurIPS	Contrastive learning	Combines internal style statistics, external information, and contrastive objectives [12].
CSBNet	2022	IJCAI	Video modular network	Uses crystallization, separation, and blending modules for stable video stylization [13].
RAST	2023	WACV	Restoration-based	Uses multi-restoration to control content preservation and style embedding [27].

AdaIN provides a simple and efficient baseline, while SANet and AdaAttN demonstrate how attention can improve local adaptation. ArtFlow and RAST place stronger emphasis on content preservation by addressing content leakage through reversible or restorative processes. IEContraAST strengthens style representation through contrastive learning. MCCNet and CSBNet broaden arbitrary style transfer toward video by prioritizing temporal stability, which introduces evaluation requirements that differ from those of static image transfer.

3.3. Comparative Synthesis of the Reviewed Methods

The comparative synthesis evaluates the methods according to their main strengths, limitations, and research implications. Direct numerical comparison is avoided when the original studies use different datasets, hardware, resolutions, or evaluation settings. Table 6 compares the principal strengths, limitations, and research implications of the core methods. Direct numerical ranking is avoided because the original studies use different datasets, hardware, image resolutions, and evaluation settings. The comparison, therefore, emphasizes architectural mechanisms and reported methodological behavior rather than treating heterogeneous numerical results as directly equivalent.

Table 6. Comparative Synthesis of Strengths, Limitations, and Research Implications

Method	Strengths	Limitations	Research implications
AdaIN	Simple, efficient, and suitable as a real-time baseline.	Global statistics may not represent complex local style patterns or spatial correspondence.	Useful as a baseline for testing improvements in local adaptation and style fidelity.
SANet	Improves local style integration through attention and semantic spatial matching.	Attention adds computation and may still struggle with difficult content-style balance.	Relevant for evaluating attention-based local correspondence.
MCCNet	Supports image and video transfer and explicitly considers temporal continuity.	Comparison with image-only methods is not fully equivalent; temporal evaluation is required.	Important for video benchmark tracks and temporal consistency analysis.

Method	Strengths	Limitations	Research implications
ArtFlow	Reduces content leakage through reversible neural flows.	Architecturally more complex than statistic-based normalization methods.	Supports research on repeated stylization, invertibility, and content preservation.
AdaAttN	Combines attention with local statistical normalization and improves local detail.	Multi-level attention increases memory and computation relative to AdaIN.	Useful for examining the contribution of local attentive statistics.
IEContra AST	Improves style harmony using internal-external learning and contrastive objectives.	Training objectives are more complex and require broader style information.	Relevant for studying representation learning and contrastive supervision.
CSBNet	Improves video stability through crystallization, separation, and blending.	Video-specific modules increase deployment complexity and limit equivalence with static image evaluation.	Should be compared within a dedicated video style transfer track.
RAST	Addresses content leakage through restoration and explicit content-style control.	Restoration operations may increase architectural and computational complexity.	Opens further research on restoration accuracy as a control mechanism.

3.4. Architectural Patterns and Development Directions

The literature shows a progression from efficient feature-statistic alignment toward architectures that model local relationships, invertible transformations, learned style distributions, restoration processes, and temporal dependencies. Normalization-based transfer is computationally attractive because it directly modifies feature statistics [8]. However, global statistics may underrepresent complex local patterns, motivating attention-based mechanisms such as SANet and AdaAttN [9], [10].

Table 7. Emerging Transformer-Based and Diffusion-Based Development Directions.

Direction	Representative methods	Main mechanism	Principal challenge
Transformer sequence modeling	StyTr2 [27]; Rethink AST [2]	Long-range content-style interaction, contrastive learning, and domain-specific transformer encoding.	Higher memory use and attention cost.
Local-global and unified fusion	LGAST [3]; StyleCraft [12]	Local-global style learning and multi-level content-style fusion.	More complex training and component interactions.
Diffusion-based image transfer	StyDiff [17]	Iterative denoising or conditional generation guided by content and style information.	Multiple inference steps and high computational cost.
Diffusion-based video transfer	Style-A-Video [18]	Diffusion generation with temporal guidance or consistency constraints.	Temporal stability and real-time deployment.

ArtFlow and RAST reflect stronger concern for content leakage, which occurs when style embedding unintentionally alters or weakens the original content structure. ArtFlow uses reversible neural flows to support unbiased projection and reconstruction [28], whereas RAST uses multi-restoration to regulate content preservation and style embedding [11], [29]. For video, MCCNet and CSBNet introduce correlation or modular mechanisms that preserve temporal continuity and reduce frame-to-frame instability [14], [15]. Image and video methods should therefore not be placed in one undifferentiated ranking because their evaluation objectives differ.

Transformer-oriented research extends arbitrary style transfer by enabling long-range interactions and more structured content-style fusion. Rethink AST combines transformer modeling with contrastive learning [2], LGAST integrates local and global style information [3], StyleCraft uses unified multi-level content-style fusion [12], and StyTr2 employs separate content and style transformer encoders followed by a transformer decoder [27]. These approaches strengthen global modeling and semantic alignment, but they also increase architectural complexity and attention-related computational cost.

Diffusion-based approaches formulate stylization as conditional generation or iterative denoising. StyDiff refines image style transfer through a diffusion framework [17], while Style-A-Video extends diffusion-based stylization to temporally guided video generation [18]. These methods offer stronger controllability and visual diversity but typically require multiple inference steps and greater computational resources.

Diffusion-based approaches provide another emerging direction by formulating stylization as an iterative denoising or conditional generation process. Such models may improve controllability, visual diversity, and semantic consistency [17], [18]. Their principal limitation is computational cost because multiple denoising steps can increase inference latency. Diffusion-based video stylization additionally requires temporal conditioning to avoid flickering and inconsistent appearance across frames.

3.5. Evaluation Metrics Used

Neural style transfer cannot be evaluated reliably using a single metric because stylization quality contains multiple dimensions. Content preservation, style fidelity, perceptual quality, temporal stability, computational efficiency, and human aesthetic preference represent different objectives [6], [19]. A method can preserve content effectively while producing weak style resemblance, or achieve strong style transfer while introducing structural distortion.

Objective metrics should therefore be interpreted as complementary measurements. Content loss and SSIM primarily describe structural preservation. Style loss, FID, and ArtFID focus on style or distributional correspondence. LPIPS evaluates perceptual distance, while user studies assess aesthetic preference and visual harmony. Video methods additionally require temporal loss, warping error, or inter-frame consistency. Computational evaluation should report latency, frames per second, model size, peak memory, and hardware specifications.

Table 8. Evaluation Metric Map for Arbitrary Neural Style Transfer

Evaluation dimension	Metrics	Purpose	Main limitation
Content preservation	Content loss, SSIM, LPIPS	Measures structural or perceptual similarity to the content image.	May penalize acceptable changes introduced by stylization.
Style fidelity	Style loss, Gram loss, FID, ArtFID	Measures correspondence between the generated output and target style characteristics.	Depends on the feature extractor, reference set, and metric assumptions.
Perceptual quality	LPIPS and human preference evaluation	Assesses perceptual similarity, harmony, and perceived quality.	Human assessment is subjective and requires careful study design.
Temporal consistency	Temporal loss, warping error, inter-frame consistency	Evaluates stability between consecutive video frames.	Requires video data, temporal correspondence, and consistent optical flow assumptions.
Computational efficiency	Latency, FPS, FLOPs, model size, peak memory	Evaluates processing speed and resource requirements.	Strongly affected by hardware, software, batch size, and input resolution.
Reproducibility	Code, checkpoint, environment, seed, evaluation script	Assesses whether the reported experiment can be rerun consistently.	Public code alone does not guarantee complete reproducibility.

Future experiments should report a metric portfolio rather than a single score. At minimum, an image benchmark should combine content preservation, style fidelity, perceptual quality, inference efficiency, and blind human preference. A video benchmark should add temporal consistency measurements and frame-processing performance. Metric definitions, feature extractors, image resolutions, and implementation settings must be documented to support fair comparison [6], [8].

3.6. Reproducibility and Repository Availability

Implementation availability supports reproducibility, but repository access alone is insufficient. A reproducible study should provide training and inference scripts, pretrained checkpoints, environment or dependency files, dataset preparation instructions, preprocessing configurations, evaluation scripts, and random seeds [21], [22]. Differences in framework versions, pretrained encoders, input resolution, and hardware can produce different results even when the same repository is used.

The repositories were verified as supporting implementation sources by matching the repository README with the method title, authors, publication venue, and architectural description. Repository-name or acronym similarity alone was not accepted as evidence because unrelated projects may use the same method abbreviation. Only method-specific repositories are retained in Table 9. Publication years follow the peer-reviewed papers reported in Table 5 and are not determined by repository creation dates, recent commits, or GitHub activity.

Table 9. Verified Repository Availability and Reproducibility Materials for The Reviewed Methods

Method	Verified repository and status	Training	Inference	Checkpoint	Environment	Dataset guidance	Reproducibility
AdaIN	Official implementation https://github.com/xunhuang1995/AdaIN-style	Yes	Yes	Yes	Partial	Yes	Moderate
SANet	Official project page and unofficial implementation https://dypark86.github.io/SANET/ https://github.com/GlebSBrykin/SANET	Partial	Partial	Yes	No	Partial	Low

Method	Verified repository and status	Training	Inference	Checkpoint	Environment	Dataset guidance	Reproducibility
MCCNet	Method-specific implementation https://github.com/diyiiyiii/MCCNet	Yes	Yes	Yes	Partial	Yes	Moderate
ArtFlow	Official implementation https://github.com/pkuanjie/ArtFlow	Yes	Yes	Yes	Yes	Yes	High
AdaAttN	Supporting PyTorch reimplementation https://github.com/huage001/adaattn	Yes	Yes	Yes	Partial	Yes	Moderate
IEContraAST	Official implementation https://github.com/HalbertCH/IEContraAST	Yes	Yes	Yes	Partial	Yes	Moderate
CSBNet	Official implementation https://github.com/Josh00-Lu/CSBNet	Yes	Yes	Yes	Partial	Partial	Moderate
RAST	Official implementation https://github.com/YingnanMa/RAST	Partial	Yes	Partial	Partial	Partial	Moderate

AdaIN, ArtFlow, IEContraAST, CSBNet, and RAST explicitly identify themselves as official method implementations. The AdaAttN repository identifies itself as an unofficial PyTorch reimplementation. The SANet authors provide an official project page but state that the source code will not be released; therefore, the available unofficial PyTorch implementation is reported only as supporting material. The verification date records when repository contents were examined and does not replace the scientific publication year. Repository verification also excluded acronym-matched links that referred to unrelated projects, including a generic developer profile, a Solidity smart-contract library, a remote-sensing salient-object-detection network, and a pulmonary-nodule-detection network. These links were not used as evidence in the synthesis.

3.7. Computational Complexity, Training Requirements, and Deployment Challenges

The reviewed methods differ substantially in computational complexity, training requirements, and deployment feasibility. AdaIN performs direct feature-statistic alignment and therefore has relatively low transfer overhead [8]. SANet and AdaAttN introduce attention operations that improve local feature matching but increase memory and computation [9], [10]. ArtFlow uses reversible flow modules [28], while RAST uses restoration operations that increase architectural complexity [11], [29].

Table 10. Computational and training characteristics of the core methods.

Method	Main computation	Training requirement	Deployment implication
AdaIN	Channel-wise mean and variance alignment.	Encoder-decoder training with perceptual objectives.	Relatively suitable for real-time inference and baseline deployment.
SANet	Multi-level style attention.	End-to-end training with attention and identity losses.	Higher memory use than simple normalization methods.
MCCNet	Multi-channel correlation and feature alignment.	Image and video-oriented training with temporal considerations.	Requires temporal processing and frame-rate evaluation.
ArtFlow	Reversible neural flows and feature transfer.	Training of invertible transformation modules.	More complex architecture and memory behavior than AdaIN.
AdaAttN	Per-point adaptive attention normalization.	Multi-level attentive normalization and local feature objectives.	Attention operations increase computation and memory.
IEContraAST	Internal-external representation learning and contrastive losses.	Multiple contrastive objectives and access to broader style information.	Higher training complexity; inference cost should be reported separately.
CSBNet	Crystallization, separation, and blending modules.	Video-oriented stability-aware training.	Temporal modules increase deployment and memory requirements.
RAST	Multi-restoration operations.	Restoration-based training with dedicated losses.	Real-time feasibility requires standardized latency and memory testing.

Training complexity also depends on the learning objectives. IEContraAST uses internal-external learning and multiple contrastive losses [13]. MCCNet and CSBNet address video transfer and therefore require temporal processing or stability-aware components [14], [15]. Transformer-oriented approaches introduce long-range attention, local-global learning, or multi-level fusion [2], [3], [12], [27]. Diffusion-based approaches generally impose the greatest inference burden because outputs are generated through multiple iterative steps [17], [18].

Deployment on mobile, edge, or real-time systems remains challenging for models that rely on multi-level attention, invertible transformations, temporal feature processing, or iterative generation. Potential

strategies include pruning, quantization, knowledge distillation, lightweight encoders, reduced attention resolution, and fewer diffusion steps. Computational comparison must use the same hardware, software environment, input resolution, and batch size. Results reported under different conditions should not be ranked directly.

Exact parameter counts; FLOPs, runtime, and memory usage should be reported only when provided in the original papers or measured through a controlled reimplementation. NR should be used when information is not reported. Runtime values from different studies should not be interpreted as directly comparable when hardware, image resolution, batch size, and software configurations differ.

3.8. Research Gaps and Future Research Recommendations

The principal research gap is the absence of a widely accepted benchmark for arbitrary neural style transfer. Existing studies emphasize different objectives and use different experimental settings, which prevents reliable cross-method ranking [6], [19]. Future research should establish separate image and video benchmark tracks because the two tasks involve different evaluation requirements.

For image arbitrary style transfer, all methods should be tested using identical content-style pairs, dataset splits, preprocessing procedures, input resolutions, random seeds, and hardware. Recommended measurements include SSIM, LPIPS, content loss, style loss, ArtFID, inference latency, model size, and peak memory [19], [20]. Blind and randomized pairwise user studies should be used to validate aesthetic preference and visual acceptance.

For video arbitrary style transfer, the protocol should additionally measure temporal loss, warping error, flickering, inter-frame consistency, frames per second, and memory consumption [16], [14], [13]. Reproducibility requirements should include source code, pretrained checkpoints, environment files, evaluation scripts, dataset splits, random seeds, and hardware specifications [19], [20]. Transformer-based and diffusion-based methods should be incorporated into future comparisons because they improve global modeling and controllability but introduce additional computational and deployment challenges.

Table 11. Research Gaps and Future Research Recommendations

Research gap	Synthesis finding	Recommendation
Fair comparison	Datasets, resolutions, baselines, hardware, and metrics differ across studies.	Re-benchmark methods using identical content-style pairs, preprocessing, resolutions, devices, and evaluation scripts.
Content leakage	Strong style transfer can alter or weaken the original content structure.	Investigate reversible flows, restoration, structural constraints, and semantic preservation losses.
Temporal consistency	Image-only methods do not guarantee stable video output.	Use a separate video track with temporal loss, warping error, flickering, and inter-frame consistency.
Human preference	Objective metrics do not fully represent visual harmony or aesthetic quality.	Combine objective metrics with blind, randomized, pairwise user preference studies.
Computational reporting	Runtime is often reported without standardized hardware or resolution.	Report latency, FPS, parameters, FLOPs, model size, peak memory, hardware, and software versions.
Reproducibility	Public repositories may lack checkpoints, dependencies, or evaluation scripts.	Publish complete code, checkpoints, environment files, dataset splits, seeds, and evaluation procedures.
Emerging architectures	Transformer and diffusion methods improve modeling and controllability but increase cost.	Compare CNN, transformer, and diffusion approaches under the same benchmark and deployment constraints.
Domain-specific use	Most evaluations rely on general photographs and artistic paintings.	Test anime, promotional design, MSME visual branding, educational media, and culturally specific artwork.

3.9. Review Limitations and Threats to Validity

This review has several limitations. First, the eight core methods were determined in advance to represent major architectural directions. This directed design supports focused comparison but may exclude other relevant methods. Second, the reviewed studies use heterogeneous datasets, image resolutions, hardware, and evaluation metrics, which prevents a statistically valid meta-analysis or direct numerical ranking.

Third, computational information such as parameter count, FLOPs, peak memory, and standardized latency is not consistently reported. Repository availability was treated as supporting evidence, but code access does not guarantee that experiments can be reproduced under the same environment. Fourth, transformer-based and diffusion-based studies were discussed as supporting developments rather than included as additional core methods. Future reviews may expand the core set and conduct controlled reimplementation experiments using a unified benchmark.

4. CONCLUSION

This study presents a directed systematic literature review of eight arbitrary neural style transfer methods, namely AdaIN, AdaAttN, ArtFlow, CSBNet, IEContraAST, MCCNet, RAST, and SANet. The review shows that the development of arbitrary neural style transfer has moved from efficient feature-statistic alignment toward more advanced architectures that can better handle local structure, content preservation, and temporal stability. This evolution can be seen in the shift from normalization-based approaches to attention mechanisms, reversible flow models, contrastive learning, video temporal modeling, and restoration-based stylization, reflecting the field's effort to improve both visual quality and structural consistency.

AdaIN remains important as an efficient baseline for arbitrary style transfer, while SANet and AdaAttN demonstrate the value of local adaptation through attention-based feature mapping. ArtFlow and RAST emphasize the need to address content leakage, IEContraAST strengthens style representation through contrastive learning, and MCCNet and CSBNet expand the discussion toward video style transfer by considering temporal consistency across frames. From the evaluation perspective, this review indicates that content loss, style loss, SSIM, LPIPS, FID, ArtFID, temporal loss, and user studies should be treated as complementary metrics rather than interchangeable measures, because each metric captures a different dimension of stylization quality.

REFERENCES

[1] Q. Cai, M. Ma, C. Wang, and H. Li, "Image neural style transfer: A review," *Computers and Electrical Engineering*, vol. 108, p. 108723, May 2023, doi: 10.1016/J.COMPELECENG.2023.108723.

[2] Z. Zhang *et al.*, "Rethink arbitrary style transfer with transformer and contrastive learning," *Computer Vision and Image Understanding*, vol. 241, p. 103951, Apr. 2024, doi: 10.1016/J.CVIU.2024.103951.

[3] Z. Zhang *et al.*, "LGAST: Towards high-quality arbitrary style transfer with local–global style learning," *Neurocomputing*, vol. 623, p. 129434, Mar. 2025, doi: 10.1016/J.NEUCOM.2025.129434.

[4] S. Li, Z. Zhang, G. Sun, L. W. H. Lehman, J. Yang, and J. Li, "Creative style transfer for image stylization via learning neural permutation," *Knowl. Based. Syst.*, vol. 330, p. 114368, Nov. 2025, doi: 10.1016/J.KNOSYS.2025.114368.

[5] D. L. Yonia, I. Ariansyah, A. B. Dina, and N. Suciati, "Enhancing Image Captioning Performance with VGG16 Feature Extraction and LSTM Sequence Processing," *ICECOS 2024 - 4th International Conference on Electrical Engineering and Computer Science, Proceeding*, pp. 77–82, 2024, doi: 10.1109/ICECOS63900.2024.10791213.

[6] E. Ioannou and S. Maddock, "Evaluation in Neural Style Transfer: A Review," *Computer Graphics Forum*, vol. 43, no. 6, p. e15165, Sep. 2024, doi: 10.1111/CGF.15165;ISSUE:ISSUE:DOI.

[7] C. Ma, Z. Sun, and C. Ruan, "Style creation: multiple styles transfer with incremental learning and distillation loss," *Multimed. Tools Appl.*, vol. 83, no. 10, pp. 28341–28356, Mar. 2024, doi: 10.1007/S11042-023-15532-5/METRICS.

[8] X. Huang and S. Belongie, "Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October, pp. 1510–1519, Mar. 2017, doi: 10.1109/ICCV.2017.167.

[9] D. Y. Park and K. H. Lee, "Arbitrary Style Transfer with Style-Attentional Networks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 5873–5881, Dec. 2018, doi: 10.1109/CVPR.2019.00603.

[10] S. Liu *et al.*, "AdaAttN: Revisit Attention Mechanism in Arbitrary Neural Style Transfer," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6629–6638, Aug. 2021, doi: 10.1109/ICCV48922.2021.00658.

[11] Y. Ma, C. Zhao, B. Huang, X. Li, and A. Basu, "RAST: Restorable Arbitrary Style Transfer," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 5, Jan. 2024, doi: 10.1145/3638770;CSUBTYPE:STRING:JOURNAL;SUBPAGE:STRING:BASIC.

[12] C. Tang and X. Chang, "StyleCraft: High-quality Arbitrary Style Transfer via Unified Content-Style Fusion," *Lecture Notes in Computer Science*, vol. 15843 LNCS, pp. 231–241, 2025, doi: 10.1007/978-981-96-9866-0_20/SAVE-RESEARCH.

[13] H. Chen *et al.*, "Artistic Style Transfer with Internal-external Learning and Contrastive Learning," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 26561–26573, Dec. 2021.

[14] H. Lu and Z. Wang, "Universal Video Style Transfer via Crystallization, Separation, and Blending," *IJCAI International Joint Conference on Artificial Intelligence*, vol. 6, pp. 4957–4965, Jul. 2022, doi: 10.24963/IJCAI.2022/687.

[15] Y. Deng, F. Tang, W. Dong, H. Huang, C. Ma, and C. Xu, "Arbitrary Video Style Transfer via Multi-Channel Correlation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, pp. 1210–1217, May 2021, doi: 10.1609/AAAI.V35I2.16208.

[16] N. Z. Rizqullah, H. Fabroyir, D. L. Yonia, and J. Alekhine, "Integrating AR into School Fire Safety Training: Enhancing Knowledge, Motivation, and Self-Efficacy," *Proceedings - 2024 International Conference on Cyberworlds, CW 2024*, pp. 159–166, 2024, doi: 10.1109/CW64301.2024.00051.

[17] Y. Sun and H. Meng, "StyDiff: a refined style transfer method based on diffusion models," *Scientific Reports 2025 15:1*, vol. 15, no. 1, pp. 33521-, Sep. 2025, doi: 10.1038/s41598-025-17899-x.

[18] N. Huang, Y. Zhang, and W. Dong, "Style-A-Video: Agile Diffusion for Arbitrary Text-Based Video Style Transfer," *IEEE Signal Process. Lett.*, vol. 31, pp. 1494–1498, 2024, doi: 10.1109/LSP.2024.3398538.

[19] H. Chen *et al.*, "Quality Evaluation of Arbitrary Style Transfer: Subjective Study and Objective Metric," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 7, pp. 3055–3070, Jul. 2023, doi: 10.1109/TCSVT.2022.3231041.

[20] M. Wright and B. Ommer, "ArtFID: Quantitative Evaluation of Neural Style Transfer," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13485 LNCS, pp. 560–576, Jul. 2022, doi: 10.1007/978-3-031-16788-1_34.

[21] D. Holst, K. Moenck, J. Koch, O. Schmedemann, and T. Schüppstuhl, "Transparent Reporting of AI in Systematic Literature Reviews: Development of the PRISMA-trAIce Checklist," *JMIR AI*, vol. 4, no. 1, p. e80247, Dec. 2025, doi: 10.2196/80247.

[22] A. M. Streiber *et al.*, "Improving reproducibility of data analysis and code in medical research: 5 recommendations to get started," *BMJ Open*, vol. 15, no. 10, p. e104691, Oct. 2025, doi: 10.1136/BMJOPEN-2025-104691.

[23] M. J. Page *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, Mar. 2021, doi: 10.1136/BMJ.N71.

[24] S. O. Nduaguba and K. M. Kelly, "Multilevel factors associated with delays in screening, diagnosis, and treatment for lung cancer—A mixed methods systematic review protocol," *PLoS One*, vol. 19, no. 10, p. e0309196, Oct. 2024, doi: 10.1371/JOURNAL.PONE.0309196.

[25] D. L. Yonia and S. C. Hidayati, "Assessing Food Photo Aesthetics Through Dominant Color and Color Harmony with Adaptive Smoothing Regularization," in *2024 FORTEI-International Conference on Electrical Engineering (FORTEI-ICEE)*, IEEE, Oct. 2024, pp. 295–300. doi: 10.1109/FORTEI-ICEE64706.2024.10824632.

[26] Y. Ma, C. Zhao, B. Huang, X. Li, and A. Basu, "RAST: Restorable Arbitrary Style Transfer," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 5, Jan. 2024, doi: 10.1145/3638770;CSUBTYPE:STRING:JOURNAL;SUBPAGE:STRING:BASIC.

[27] Y. Deng *et al.*, "StyTr2: Image Style Transfer With Transformers," 2022. Accessed: Jun. 20, 2026. [Online]. Available: https://github.com/diyiiyiii/StyTR-2.

[28] J. An, S. Huang, Y. Song, D. Dou, W. Liu, and J. Luo, "ArtFlow: Unbiased Image Style Transfer via Reversible Neural Flows," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 862–871, Mar. 2021, doi: 10.1109/CVPR46437.2021.00092.

[29] Y. Ma, C. Zhao, A. Basu, and X. Li, "RAST: Restorable Arbitrary Style Transfer via Multi-restoration," *Proceedings - 2023 IEEE Winter Conference on Applications of Computer Vision, WACV 2023*, pp. 331–340, 2023, doi: 10.1109/WACV56688.2023.00041.