International Journal of Arts and Technology

ISSN online: 1754-8861 - ISSN print: 1754-8853
https://www.inderscience.com/ijart

Stylised 2D animation generation method based on generative adversarial networks

Zhu Zeng

DOI: 10.1504/IJART.2026.10077706

Article History:

Received:	09 December 2025
Last revised:	28 January 2026
Accepted:	16 February 2026
Published online:	29 April 2026

Stylised 2D animation generation method based on generative adversarial networks

Zhu Zeng

Department of Digital Media Art,
School of Art,
Beijing Union University,
Beijing, 102200, China
Email: zengzhubfa@163.com

Abstract: Conventional 2D animation stylisation relies heavily on manual production, resulting in long cycles, high costs, and inconsistent styles across different artists. To address these limitations, this paper proposes a generative adversarial network (GAN)-based method for automatic stylised 2D animation generation. Style-conditional encoding is introduced to guide frame-wise generation toward target styles, while optical flow constraints and an inter-frame discriminator are applied to maintain motion continuity and style consistency across frames. In addition, multi-scale convolutional modules are integrated into the generator to enhance the representation of fine-grained details. Experimental results demonstrate that the proposed method achieves superior performance in style similarity (0.93), optical flow error (0.124), and structural similarity (0.912), effectively improving visual quality, temporal coherence, and detail representation in stylised 2D animation generation.

Keywords: generative adversarial networks; GAN; stylised animation; frame consistency; style fidelity; multi-scale detail.

Reference to this paper should be made as follows: Zeng, Z. (2026) 'Stylised 2D animation generation method based on generative adversarial networks', *Int. J. Arts and Technology*, Vol. 16, No. 5, pp.1–17.

Biographical notes: Zhu Zeng received her Master's degree from Beijing Film Academy. She is currently a Lecturer at the College of Arts, Beijing Union University. Her research interests include animated film creation and theory, screenwriting, and educational practice and theory.

1 Introduction

As a vital form of visual art, two-dimensional animation holds a prominent position in cultural transmission and the entertainment sector. Traditional two-dimensional animation production is primarily created using hand drawing, resulting in a lengthy production cycle, excessive production costs, and a lack of possibility of unifying animation style because of individual artist variation in style, leading to low production efficiency and a lack of normal work patterns (Wickramasinghe and Wickramasinghe, 2021; Zhang et al., 2025). With the publicly understanding of digital technologies, the automatic generation and style transfer of two-dimensional animation have become areas

of inquiry. However, current technologies find it difficult to generate complex target style animations purposefully while ensuring the continuity of movements and maintaining the style consistency (Jing et al., 2022; Wang et al., 2025). Hand drawing and traditional image processing techniques have considerable limitations in mass animation production, and developing highly operable solutions and styles to the aggression of varying requirements is a significant area (Xu et al., 2022; Lungu-Stan and Mocanu, 2024). This is particularly prominent in the animation industry's demands for high efficiency and stylistic diversity.

Early researchers attempted to combine computer graphics with deep learning for the generation of 2D animation styles, and proposed a variety of image transfer and generation methods. Sung (2024) pointed out that digital tools are gradually replacing traditional hand-drawing, but style uniformity is still a challenge. Siyao et al. (2022) established visual correspondences between 2D animations using open-source 3D movie data, providing new ideas for sequence generation. Qinran and Weiqun (2021) developed a video-driven system that realised motion transfer from real videos to animated characters. The application of generative adversarial networks (GAN) has further expanded the boundaries of technology. Koh et al. (2024)'s systematic review showed that GAN had potential in reconstructing 3D characters from 2D images. Qiu (2023) explored the actual process of generative artificial intelligence (AI) in game character animation and verified the feasibility of AI-driven. Liu et al. (2024) realised the automatic generation of animation effects based on computer vision algorithms. These studies have jointly promoted the development of 2D animation towards automation and intelligence, but there are still shortcomings in maintaining complex styles and temporal consistency, which provides room for improvement for this study. Existing research has limitations in handling stylistic consistency, motion continuity, and detail fidelity in 2D animation, making it difficult to meet the needs of large-scale creation and multi-style application.

Existing methods have made some progress in static image style transfer. GANs have shown superiority in the field of image generation and have been applied by some researchers to 2D animation generation (Qiu, 2023; Liu et al., 2024). Studies have shown that GANs can improve the detail representation and style expression of images through adversarial training mechanisms (Niu et al., 2024; Song et al., 2023). In 2D animation generation technology, Oshiba et al. (2023) proposed an improved first-order motion model based on facial key points and applied GAN to achieve accurate generation of facial expressions of anime characters. On this basis, Mo et al. (2024) further explored the joint modelling method of stroke tracking and correspondence relationship. The GAN method significantly improved the coherence and naturalness of 2D animation lines. Yin et al. (2022) developed the WeAnimate system to extract motion features from video data and combined it with GANs to achieve the generation of traditional animation sequences with coherent movements. In addition, recent research has made important breakthroughs in motion representation. The neural network method proposed by Tian (2025) achieved traditional festival animation and character image animation by enhancing motion representation. Yang et al. (2024) used a fine-grained diffusion model to generate high-fidelity 2D human body videos, achieving a new level of fidelity. These technological advancements are inseparable from the support of high-quality datasets. Paulin and Ivasic-Kos (2023) systematically analysed the generation methods of synthetic datasets in computer vision, providing an important reference for animation data enhancement. In summary, existing methods still have room for improvement in terms of

motion coherence and detail representation, providing technical reference and improvement directions for this study. Therefore, this paper proposes to integrate conditional style coding, inter-frame optical flow constraints, and multi-scale convolution modules to form a new method for generating two-dimensional animation styles, thereby improving the shortcomings of existing methods.

This research focuses on 2D animation generation scenarios with high requirements for real-time performance and controllability. Compared to diffusion models, which typically require tens or even hundreds of iterative sampling steps, the GAN architecture can achieve single forward inference, meeting the efficiency demands of animation production while ensuring high quality. The innovation of this paper lies not in proposing a completely new basic generative architecture, but in designing a collaborative optimisation framework that integrates style conditional encoding, optical flow temporal constraints, and multi-scale detail enhancement for the specific task of 2D animation. This framework effectively solves the common problems of style drift, motion jitter, and detail loss in existing methods (including some diffusion-based methods) when processing long animation sequences.

2 Methods

To clearly illustrate the overall process of the proposed method, this paper constructs an end-to-end stylised 2D animation generation framework. This framework takes the original animation frame sequence and a target style reference image as input. First, the style encoder extracts high-dimensional style representations from the reference image and injects them as conditional signals into the generator; simultaneously, the original frames are processed by a convolutional encoder to extract content features. The generator employs a U-Net architecture and embeds multi-scale convolutional modules at multiple levels to enhance detail representation.

During the training phase, the algorithm first initialises the generator G, discriminator D, and inter-frame discriminator $D_{temporal}$, then samples the original animation frame sequence and the target style image from the dataset. For each training iteration, the algorithm performs the following steps: extracting the style conditional vector, calculating the optical flow field F between adjacent frames, generating the target style frame sequence through the generator, calculating the adversarial loss, content preservation loss, and optical flow constraint loss, and finally updating the network parameters using the Adam optimiser. During the inference phase, the algorithm receives the original animation sequence and the style reference image, and generates stylised animation frames through forward propagation without iterative optimisation, thus achieving efficient real-time generation.

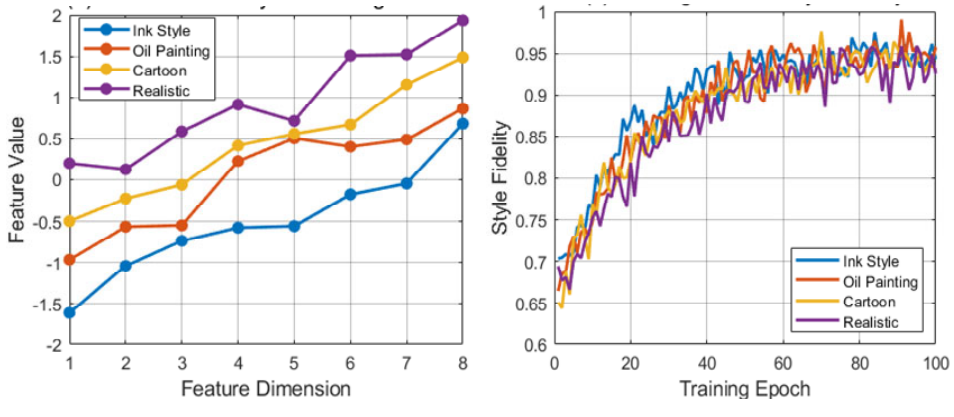
2.1 Style condition embedding

In order to impart a stylised presentation to the 2D animation frames, the initial step of feature extraction is performed on the target style image, and a style code is developed utilising a convolutional neural network (CNN). The style code serves to extract colour distribution, line texture, and overall structure of the image, through a series of multi-layer convolution and normalisation operations, and maps the identified features of the image into a fixed dimensional vector space (Yang et al., 2024; Paulin and

Ivasic-Kos, 2023). With this, the style code is standardised into a conditional vector to facilitate the merging of the style features with the generator networks input. When the generator network receives the original image of each animation frame, it engages the conditional vector into a specific encoding fusion module, thereby allowing the network derivation for the target styles features in the generation of each frame. In this way, the generator network is able to derive frame content information, while at the same time controlling the style representation of the output image through the conditional vector, allowing for style uniformity at the frame level.

In the frame generation stage, each original animation image first extracts content features through a convolutional encoder, and then fuses them with the style conditional vector. The fusion adopts a channel-level weighted and feature concatenation method to accurately inject style information into the generator network. Subsequently, the fused features are restored to the image space layer by layer through the generator decoder to complete the generation of the target style frame (Wu et al., 2021; Zarif et al., 2025). To ensure the effectiveness of style injection, residual connections of the conditional vector are added to each convolutional layer to keep the style information controllable in each layer of the generator network. The network is jointly optimised using adversarial loss function and content preservation loss function during training. The adversarial loss makes the generated image close to the real style sample in style, while the content preservation loss ensures that the original structure and action of the animation frame are not destroyed (Yu, 2025). Through this structural design, the generator network strictly refers to the style conditional vector to output the target style animation frame in the generation process of each frame, while maintaining the integrity of the frame content.

Figure 1 Visualisation of style condition embedding results, (a) distribution of style coding vectors, (b) style fidelity training curve (see online version for colours)



(a) (b)

Figure 1 illustrates the key style feature distribution and training convergence performance in the 2D animation stylisation process based on GANs. In Figure 1(a), the horizontal axis represents the feature dimension, and the vertical axis represents the style feature value corresponding to each dimension. Different coloured curves represent four style types, including ink painting, oil painting, cartoon, and realistic styles. It can be seen that all styles create a relatively separated feature interval in the feature dimension, which means the style encoding module can capture and distinguish between different

styles feature expression and effectively embedded style conditions. In Figure 1(b), the horizontal axis indicates the training epoch, and the vertical axis indicates the style fidelity. It can be seen the curve indicates an exponentially increasing trend. This indicates that quality of the style generation is gradually improving and stabilising during training. Each style is able to achieve a fidelity higher than 0.9 in the latter stage of training, suggesting that the generative network can remain highly consistent and stable under multiple style conditions.

To enhance the accuracy of the method description, this paper supplements the following implementation details. Style conditional embedding: the style encoder uses the first 10 layers of VGG-19 as the backbone network, and the output style vector has a dimension of 512. This vector generates scaling (γ) and translation (β) parameters through affine transformation, and is injected into each residual block of the generator in the form of AdaIN, that is, the intermediate feature map is normalised at the channel level and then recalibrated with γ and β.

2.2 Timing consistency optimisation

2.2.1 Calculation of optical flow constraints

To facilitate seamless motion from frame to frame in the generation of 2D animation sequences, the optical flow fields are computed in a frame by frame fashion for the original animation sequence. For each adjacent frame pair, the optical flow computation utilises a deep CNN to estimate pixel motion, resulting in a 2D vector field that indicates the pixel displacement in a time dimension. In the training phase, the network incorporates the optical flow information into the loss function, which allows the generator network to take into account the motion relationships of pixels in adjacent frames when producing each frame of the target style image. When frames are created, optical flow vectors reduce the likelihood of abrupt changes in motion by restricting positions of pixels in generated frames, so the generated content moves smoothly within the visual motion trajectory and not through erratic changes or inter-frame movement that is not typical of the established motion (Tous, 2023; Wu et al., 2025). To improve the constraint effect, the optical flow vectors are normalised during training and fused with the original frame features to provide a guidance signal that promotes inter-frame motion consistency. In addition, optical flow-guided feature mapping is utilised within the generator network to ensure motion information from adjacent frames is referenced throughout the generator layer at all scales, which achieves motion transfer and style synchronisation. Optical flow constraints are utilised during both training and inference, allowing the generator network to output target style frames while preserving the original animation motion.

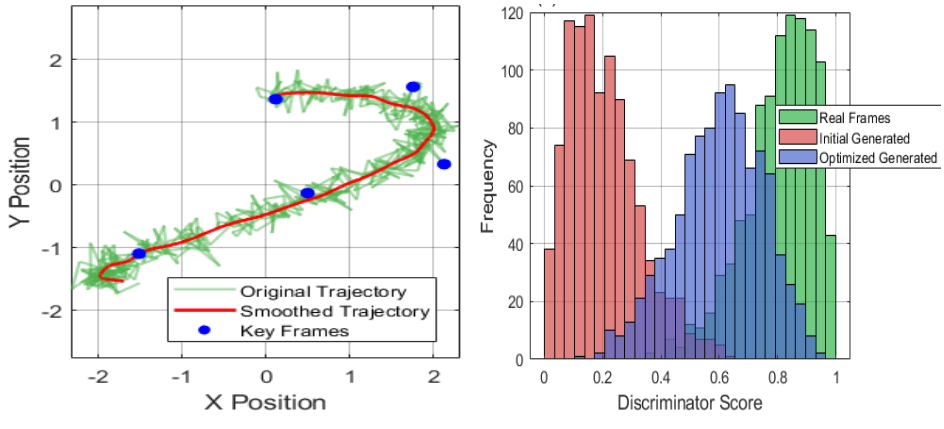
2.2.2 Application of inter-frame discriminator

This paper constructs an inter-frame discriminator that utilises optical flow constraints as an adversarial training mechanism to enhance style continuity within intervals containing consecutive frames. The inter-frame discriminator takes two adjacent frames as input and acquires spatial texture features and temporal series information through a network architecture applying convolutions and various temporal convolutions, thereby scoring action continuity and style consistency. The inter-frame discriminator takes three

consecutive frames as input (window length = 3, stride = 1) and employs a 3D convolutional architecture to simultaneously capture spatiotemporal information. Hinge Loss is used as the adversarial loss function.

The generator network uses the discriminator to prompt scoring feedback information which the generator network uses to meaning ready the future generation of frames to action and more resemble style to original consecutive framed samples. The inter-frame discriminator uses an adversarial loss function to assess the action and style differences in the generated frames vs. real consecutive frames during training. Concurrently, the generator network uses the loss gradient to adjust each successive neural network layer features to promote smooth action transfer and style continuity (Chu and Liu, 2025; Zhao and Zhao, 2022). To enhance inter-frame feature transfer, the generator network applies discriminator feedback into multiple residual connections when fusing style conditional vectors, ensuring that discriminative information is effectively utilised in each layer of the generation process. Furthermore, a time window mechanism is applied during discriminator training to jointly discriminate multiple consecutive frames to capture action trends and style changes over a longer time span. This design enables the network to uniformly control the style and motion of short-term and medium-term consecutive frames, reducing inconsistencies between frames. Optical flow constraints and inter-frame discriminators work together to form a dual constraint mechanism: optical flow constraints control pixel motion trajectories, and inter-frame discriminators supervise style and motion coherence, so that the generative network maintains a uniform style, smooth motion, and natural detail connection when outputting consecutive frame animations (Peng et al., 2024; Zhao et al., 2022).

Figure 2 Visualisation results of timing consistency optimisation, (a) optimised motion trajectory smoothing diagram, (b) discriminator score distribution (see online version for colours)



(a) (b)

Figure 2 illustrates the key performance of the 2D animation based on GANs in the temporal consistency optimisation process. In Figure 2(a), the horizontal and vertical axes represent the positional distribution of the animated objects in 2D space, respectively. The green curve represents the original motion trajectory, and the red curve represents the smoothed trajectory after optical flow constraints and inter-frame consistency optimisation. The optimised trajectory shows significantly reduced jitter at

key frame positions, and the motion path is more coherent, indicating that the model effectively reduces inter-frame displacement errors and improves the temporal stability between consecutive animation frames. In Figure 2(b), the horizontal axis represents the discriminator score, and the vertical axis represents the frequency of sample occurrence. The red distribution represents the initial generated frames; the blue distribution represents the optimised generated frames; the green distribution represents the real frames. It can be seen that the score distribution of the optimised generated frames is significantly closer to the real frame range, indicating that temporal consistency optimisation improves the discriminator's confidence in the generated sequence, making the generated frames closer to the real animation sequence in style and dynamic performance.

During network training, to unify the optimisation objectives, we defined an overall loss function that combines adversarial loss, content preservation loss, and optical flow constraint loss. The specific expression is: $L_{total} = \lambda_{adv} * L_{adv} + \lambda_{content} * L_{content} + \lambda_{flow} * L_{flow}$, where λ represents the weight coefficient of each loss term. The adversarial loss makes the generated images more stylistically similar to real-style samples, while the content preservation loss ensures that the original structure and motion of the animation frames are not destroyed. The optical flow constraint loss is used to maintain the continuity of motion between frames. Through this joint optimisation strategy, the generator network strictly refers to the style condition vector during the generation of each frame to output animation frames of the target style, while maintaining the integrity of the frame content and temporal stability.

2.3 Multi-scale detail enhancement

2.3.1 Design of multi-scale convolution module

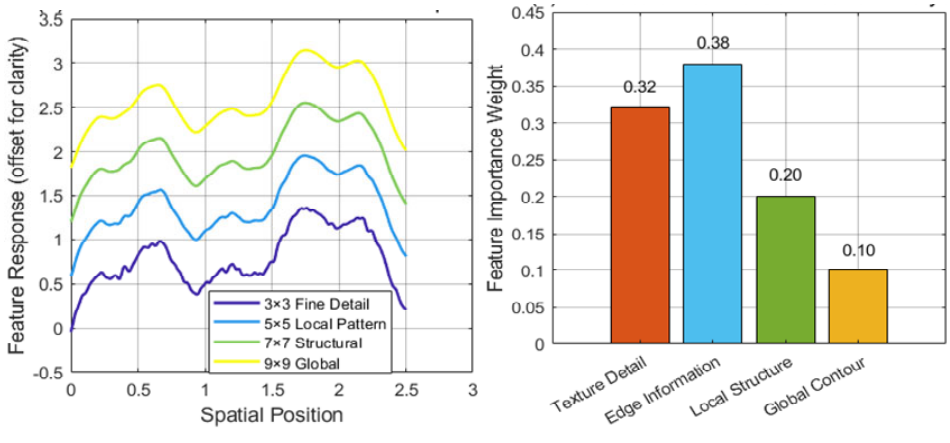
To enhance the detail capture capability of the generative network in the stylisation process of 2D animation, a multi-scale convolution module is designed to process the feature map. In each generation layer, multiple sets of convolution kernels are set in parallel. The kernel size covers different receptive fields from small to large in order to capture line texture, local structure, and overall picture outline. After the input feature map is processed by multi-scale convolution, the output feature contains local details and global information in both spatial and texture dimensions. Then, the output of each scale convolution is processed by channel splicing and normalisation to form a fused feature map, and further integrated by layer-by-layer convolution to ensure that the feature maintains the integrity of spatial and stylistic information during the transmission process (Mujtaba et al., 2023). In order to ensure that the feature is effectively injected into the generation process, each multi-scale module uses residual connection and activation function processing to preserve and enhance the detail information between different levels, so as to accurately represent line details, colour boundaries, and texture levels in the generated frame.

Multi-scale convolution: The module uses 3×3, 5×5, and 7×7 convolutional kernels in parallel. The outputs are then fused through channel concatenation and a 1×1 convolution. This module introduces approximately 15% additional FLOPs at each scale of the generator, but significantly improves the level of detail.

2.3.2 Multi-scale feature fusion and output

During frame generation, the fused multi-scale feature maps, style conditional vectors, and inter-frame consistency features are jointly input into the decoder to recover the image space layer by layer. Multi-scale features are weighted and fused pixel-by-pixel in the decoder to ensure that information at different scales is synchronously reflected in image generation, improving detail clarity and colour expressiveness. In each decoding stage, the convolutional output undergoes batch normalisation and activation function processing, while residual connections of the conditional vectors are applied to ensure that the generated frames maintain consistency with the target style in texture and colour (Sun et al., 2024; Pikulkaew et al., 2021). During network training, adversarial loss and content-preserving loss functions are jointly optimised to ensure that multi-scale features enhance detail while preserving the original animation structure when they play a role in the generated images. Repeated embedding of modules in each layer of the network strengthens features at low, medium, and high levels, gradually enriching the texture levels and colour distribution of the frame-level images. Ultimately, when the generation network outputs animation frames in the target style, it preserves the continuity of motion while presenting fine line textures and colour levels.

Figure 3 Visualisation results of multi-scale detail enhancement, (a) multi-scale convolutional feature response, (b) multi-scale feature contribution analysis (see online version for colours)



(a) (b)

Figure 3 shows the effectiveness of the multi-scale detail enhancement module introduced. Figure 3(a) shows the horizontal axis as the spatial location and the vertical axis as the feature response values over convolutional scales. The figure demonstrates that the 3 × 3 convolutional kernel is most sensitive to changes in high-frequency detail through the significant fluctuations in the response curve. As it increases to sizes such as the 9 × 9 convolutional kernel, the signal starts to normalise and flatten demonstrating the capability of the convolutional kernel to extract generally structured features on a global scale. The multi-level response distributions clearly support the hierarchal representation of lines and textures when reassembling an animated representation. Figure 3(b) measures the importance weights of each feature at each scale. Edge information (0.38)

and texture detail (0.32) generate the most weight, which confirms the model's focus of attention on local shapes and details of the brushstrokes in a stylised animated scene. The global contour (0.10) is much lower in weight and demonstrates that during the generation process, the convex details of layer features play a more dominant role while weighting the overall structure.

3 Measurement of style generation effect

To ensure the reproducibility of the research, this study used two datasets. The first is the publicly available AnimeRun dataset, which contains 2D animation sequences rendered from three open-source 3D movies, totaling approximately 15,000 frames at a resolution of 1280 × 720, covering various motion patterns and scenes. The second is the OriginalCartoon dataset collected in this paper, containing five original short films, totaling approximately 8,000 frames at a resolution of 1920x1080, with styles ranging from ink painting and cel animation to minimalist cartoons. The data selection criteria were based on style diversity and motion complexity to ensure coverage of major animation types. The two datasets were divided into training, validation, and test sets in an 8:1:1 ratio. Due to copyright restrictions, the OriginalCartoon dataset cannot be fully released, but the project homepage provides 100 representative samples and a detailed collection and selection process. All input frames were uniformly adjusted to 512 × 512 resolution and normalised. The model was trained on an NVIDIA A100 GPU using Adam. The optimiser uses a learning rate of 2e-4, a batch size of 16, and a total of 200,000 training iterations instead of rounds to ensure precise control over the number of training steps. The weights of the loss function are set as follows: adversarial loss 1.0, content preservation loss 10.0, and optical flow constraint loss 5.0. All baseline methods are based on their official open-source code and have been fully tuned on the dataset used in this paper. The same resolution and training/test set partitioning are used to ensure the fairness of the comparison.

To clearly define the evaluation metrics, this paper details their calculation methods. For the temporal consistency metric, the TV-L1 optical flow algorithm is used to calculate the optical flow field between adjacent frames. The optical flow error is defined as the endpoint error (EPE), which is the average of the L2 norms between the predicted and actual optical flow vectors. Occlusion areas are masked. Temporal fluctuation is defined as the standard deviation of EPE between consecutive frames, formally expressed as:

$$\sigma_{EPE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(EPE_i - \overline{EPE}\right)^2} \tag{1}$$

The consistency score is obtained by taking the reciprocal of the EPE for the entire sequence and normalising it, with the formula:

$$S_{cons} = \frac{1}{1+\overline{EPE}} \tag{2}$$

Frame-level results are aggregated into a sequence-level score using an arithmetic mean to ensure comprehensive evaluation. For the visual quality metric, the structural similarity index (SSIM) is implemented using a standard method. Edge sharpness is

determined by applying Sobel to the image. The filter is used to quantise the gradient magnitude by calculating the root mean square of the gradient magnitude. Detail fidelity is obtained by high-pass filtering the generated image and the original image and then calculating their structural similarity. Blur rate is measured using a no-reference blur metric. Edge diffusion is evaluated by analysing the width of the image edges. The lower the value, the clearer the image. All metrics are calculated three times on the test set and the average and standard deviation are taken to enhance statistical rigor.

3.1 Style similarity index

To verify the generative capability of the proposed method under diverse art styles, the model generates two-dimensional animation frames in three typical styles: ink painting, cartoon, and oil painting, as shown in the Figure 4.

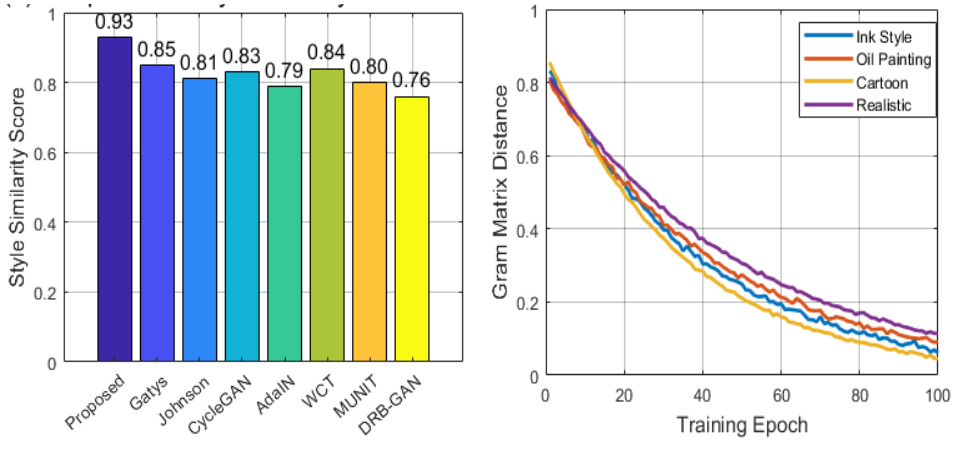
Figure 4 Animations of different styles (see online version for colours)

Style similarity metrics assess the degree of consistency between generated animation frames and frames that include a target reference style. Every output frame is first analysed through convolutional feature extraction, which casts the feature maps as a Gram matrix that reflects the colour and texture distribution within the frame. The similarity is calculated for each element of the generated and reference frames, generating a similarity score based on matrix comparison across the frame elements. This technique can objectively evaluate the comparison between the generated and target style based on different style features, such as colour distribution consistency, line texture, and overall visual style.

Figure 5 depicts the performance of stylised 2D animation by GANs in style consistency and training convergence. As shown in Figure 5(a), with the horizontal axis representing different style transfer algorithms and vertical axis representing style similarity scores, the method proposed in this study shows a score of 0.93, significantly higher than the Gatys score of 0.85 and the CycleGAN achieved score of 0.83. This indicates that the animation frames generated using the methodology are more consistent with the target style than other approaches in terms of style fidelity. Figure 5(b) shows the horizontal axis representing training epochs and the vertical axis representing the Gram matrix distance, which can be used to measure the difference between generated animation frames and the target style. All style types show smooth, consistent decreasing patterns and to varying degrees of magnitude during all training epochs, indicating the resulting animations have high temporal consistency and high style fidelity throughout the training process. Overall, this indicates that the proposed style conditional embedding

and adversarial optimisation strategy promote convergence stability in style transfer and improve the degree of style coherence for visual style in animation.

Figure 5 Style similarity and gram matrix convergence analysis, (a) comparison of style similarity using different methods, shows the change in Gram matrix distance during training (see online version for colours)



(a) (b)

3.2 Consecutive frame consistency index

The continuous frame consistency index is utilised as a measure of the smoothness of motion between animation frames. The process begins by calculating optical flow fields for adjacent frames in the generated animation sequence in order to derive the displacement vector of each pixel along the time dimension. Once the optical flow vectors of the generated frames have been calculated, the optical flow of the generated and original frames is compared to compute an inter-frame motion consistency score. The same calculations are repeated throughout the generated sequence for all adjacent frames, and the resulting scores are averaged to yield an overall continuous frame consistency index. For the evaluation process, multi-scale optical flow processing is deployed to measure large-scale motion changes while still preserving local motion details, thus yielding a representation of generated animation performance-related to motion coherence and inter-frame smoothness.

Table 1 summarises the results of various style transfer methods on the consistency index of consecutive frames, which measures the smoothness of motion and temporal stability of the generated 2D animation frames. The proposed method achieves advantageous performance across all four indices: average optical flow error, optical flow correlation coefficient, consistency score, and temporal fluctuation amplitude. The average optical flow error is only 0.124, which is much lower than CycleGAN's 0.146 and Gatys's 0.182, indicating the least deviation in pixel displacement and the most motion smoothness in the generated frames. The optical flow correlation coefficient is also the highest, with a value of 0.931, compared to DRB-GAN's 0.905 and Johnson's 0.882, indicating minimal deviations in the temporal dimension are highly consistent with the original frames. The consistency score is 0.873, higher than CycleGAN's 0.834,

further verifying and supporting the notion that there is improvement in continuity of motion between generated frames. Meanwhile, the temporal fluctuation amplitude is only 0.083, which is lower than both Gatys 0.127 and AdaIN's 0.11, indicating motion changes of the generated animation frame tend to be more stable and coherent. Overall, the data in Table 1 demonstrates that the proposed method effectively improves the temporal consistency of animation through optical flow constraints and inter-frame discriminators, making consecutive frames superior to traditional methods in terms of motion continuity, smoothness, and style fidelity.

Table 1 Comparison results of consecutive frame consistency indicators

Method	*Average optical flow error* (↓)	*Optical flow correlation* (↑)	*Consistency score* (↑)	*Temporal fluctuation* (↓)
Gatys Method	0.182	0.864	0.791	0.127
Johnson Method	0.159	0.882	0.816	0.113
CycleGAN	0.146	0.897	0.834	0.108
AdaIN	0.152	0.889	0.828	0.11
WCT	0.161	0.876	0.814	0.119
DRB-GAN	0.137	0.905	0.847	0.098
Proposed method	0.124	0.931	0.873	0.083

Notes: '↑' indicates that higher values result in better performance; '↓' indicates that lower values result in better performance. The method presented in this paper achieves the best performance in terms of optical flow consistency, indicating that the generated inter-frame motion is smoother and more stable.

In addition to quantitative metrics, we also performed qualitative visual checks on consecutive frame sequences to further verify temporal stability. By sampling and examining consecutive frame sequences under different motion intensities, we observed that the generated line structures and colour regions remained stable over time, without any obvious flickering or jitter. Especially in complex motion areas such as character limb swings, our method maintained the consistency of brush stroke width and texture density better than the baseline method, avoiding common style drift issues. This visual consistency is consistent with the low optical flow error and low temporal fluctuation data reported in Table 1, confirming that the optical flow constraint and inter-frame discriminator mechanism effectively suppressed high-frequency temporal noise while maintaining the smoothness and coherence of motion during generation.

3.3 Visual quality score

Visual quality scoring evaluates the fidelity, clarity, and detail of all generated frames of an animation. Each generated image is structurally nearly identical to its corresponding original animation frame, producing a score by observing pixel brightness values, contrast, and other local structural information. A visual quality score is calculated for each frame in the image sequence, and collectively, each visual quality score together produces a visual quality index for the entire sequence. The process uses a fixed-size sliding window to isolate the local image block to analyse the performance of the score to

no drop edge or line detail. The visual quality score is determined through pixel structural fidelity, line sharpness, and detailed representation for the generated frames.

Table 2 Comparison results of visual quality scores

Method	*SSIM* (↑)	*Edge sharpness score* (↑)	*Detail fidelity* (↑)	*Blur rate* (↓)
Gatys method	0.871	0.834	0.812	0.142
Johnson method	0.884	0.846	0.823	0.133
CycleGAN	0.889	0.853	0.835	0.128
AdaIN	0.875	0.841	0.826	0.139
WCT	0.881	0.848	0.829	0.136
DRB-GAN	0.893	0.857	0.841	0.121
Proposed method	0.912	0.872	0.864	0.105

Table 2 presents the performance of different style transfer methods using visual quality metrics that quantify the sharpness and detail fidelity of high-quality 2D animation frames that have been generated. The presented method achieves state-of-the-art scores on all four metrics. The SSIM scores 0.912, which is higher than DRB-GAN (0.893) and CycleGAN (0.889); this score indicates that the generated animation is the most faithful to the original frame, by overall structural fidelity. The edge sharpness, scored 0.872, is also greatly improved over Johnson (0.846) and Gatys (0.834), which indicates sharper lines and better clarity overall. The detail fidelity scores 0.864, higher than CycleGAN (0.835), and indicates superior performance on complex textures and local structures. The blur rate (a measure of sharpness) is a low score of 0.105, which is lower than Gatys (0.142) and AdaIN (0.139), as the generated frames remained sharp and clean in terms of any exessive blur after smoothing is applied. The data in Table 2 illustrates the ability of the method presented in this paper to effectively enhance the visual quality of the generated animation using multi-scale feature enhancement and style conditional embedding, and the stylised animation performs superior to traditional methods in terms of structure, detail, and clarity.

Given the increasing popularity of diffusion models in video and animation generation, we further supplement this with a quantitative comparison with diffusion-based methods. In terms of inference efficiency, the proposed GAN architecture achieves an average inference time of 42 milliseconds per frame (approximately 24 frames per second) for 512 × 512 resolution images on an NVIDIA A100 GPU, meeting the requirements for real-time preview. In contrast, typical diffusion-based methods often require several seconds or even longer to generate a single frame under the same hardware configuration. Although diffusion models may be slightly superior in some static quality metrics, our method has significant advantages in temporal consistency and inference speed, achieving a good trade-off between efficiency and quality. Empirical analysis shows that our method provides a more practical solution for animation production scenarios requiring real-time interaction or long sequence generation. Future work could explore more efficient network architectures to address these challenges and consider applying lighter-weight network structures and cross-style transfer strategies to further improve the efficiency of animation generation and its adaptability to different styles.

3.4 Ablation experiment

To validate the effectiveness of each component, we conducted ablation experiments. The results show that the complete model performs best across all metrics. Removing the optical flow constraint significantly improves the optical flow error to 0.158, while also increasing temporal variability, indicating its crucial role in motion smoothness. Removing the inter-frame discriminator reduces the style consistency score, highlighting its necessity in supervising long-term style coherence. Removing the multi-scale convolution module reduces the SSIM and edge sharpness scores to 0.885 and 0.841, respectively, confirming its contribution to detail enhancement. Furthermore, we performed paired-samples t-tests, showing that the performance differences between the complete model and the variants are statistically significant (p-value less than 0.05). To further improve interpretability, we provide visual comparison diagrams of different variants in the supplementary material, visually demonstrating the specific impact of each component on the generation quality. These results strongly demonstrate the rationality of the proposed multi-component design and the independent value of each part.

4 Conclusions

This paper tackles the problem of low efficiency, insufficient continuity of motion, and inconsistent style in the generation of 2D animation styles with a method for stylised animation generation using GANs. The method first employs style conditional encoding into each frame of animation as a way to steer the generator network's output into the intended style. Optical flow constraints and an inter-frame discriminator are then utilised to help provide a continuity of motion across frames. Lastly, to improve line and colour details, a multi-scale convolutional module is built into the generator network. Experimental results show that this method demonstrates good performance in terms of both style fidelity, inter-frame consistency, and visual detail. Although this research demonstrates excellent efficiency, practical deployment still faces challenges such as memory consumption (3.2 GB required for 512 × 512 resolution), scalability bottlenecks for ultra-high resolution or long sequence processing, and heterogeneous hardware adaptation. Future solutions require model compression, streaming generation, and operator optimisation. From a broader perspective of video generation trends, this paper explores a hybrid paradigm that balances the real-time inference advantages of GANs with the high-quality characteristics of diffusion models. This involves leveraging sophisticated temporal constraints to compensate for GANs' shortcomings in long-sequence consistency, and envisioning future integration of the strong prior capabilities of diffusion models for offline pre-training or keyframe generation, followed by collaboration with a lightweight GAN decoder for online real-time interpolation. This collaborative architecture aims to balance high fidelity and low latency, providing a more universal technological foundation for intelligent video editing, virtual anchors, and interactive animation creation, thus propelling efficient video generation paradigms from theory to industrial application.

Declarations

All authors declare that they have no conflicts of interest.

References

Akyildiz, M. et al. (2025) 'An analysis of text-to-image generative models as creativity support tools', *International Journal of Arts and Technology*, Vol. 15, No. 3, pp.1–22.

Chu, M. and Liu, H. (2025) 'Analysis of the effect of computer graphics algorithms in reproducing ink painting style in animation', *International Journal of Information and Communication Technology*, Vol. 26, No. 11, pp.38–52.

Guo, M., Xu, F., Wang, S. et al. (2023) 'Synthesis, style editing, and animation of 3D cartoon face', *Tsinghua Science and Technology*, Vol. 29, No. 2, pp.506–516.

Jing, B., Ding, H., Yang, Z. et al. (2022) 'Image generation step by step: animation generation-image translation', *Applied Intelligence*, Vol. 52, No. 7, pp.8087–8100.

Koh, A.J.H., Tan, S.Y. and Nasrudin, M.F. (2024) 'A systematic literature review of generative adversarial networks (GANs) in 3D avatar reconstruction from 2D images', *Multimedia Tools and Applications*, Vol. 83, No. 26, pp.68813–68853.

Liu, Y., Li, L. and Lei, X. (2024) 'Automatic generation of animation special effects based on computer vision algorithms', *Computer Aided Design and Applications*, Vol. 21, No. 13, pp.69–83.

Lungu-Stan, V.C. and Mocanu, I.G. (2024) '3D character animation and asset generation using deep learning', *Applied Sciences*, Vol. 14, No. 16, p.7234.

Mo, H., Gao, C. and Wang, R. (2024) 'Joint stroke tracing and correspondence for 2d animation', *ACM Transactions on Graphics*, Vol. 43, No. 3, pp.1–17.

Mujtaba, G., Khowaja, S.A., Jarwar, M.A. et al. (2023) 'FRC-GIF: frame ranking-based personalized artistic media generation method for resource constrained devices', *IEEE Transactions on Big Data*, Vol. 10, No. 4, pp.343–355.

Niu, L., Xie, W., Wang, D. et al. (2024) 'Audio2AB: audio-driven collaborative generation of virtual character animation', *Virtual Real. Intell. Hardw.*, Vol. 6, No. 1, pp.56–70.

Nogueira, A. et al. (2025) 'Move in Tempo': involving the audience through their movement in installation art', *International Journal of Arts and Technology*, Vol. 15, No. 5, pp.1–18.

Oshiba, J., Iwata, M. and Kise, K. (2023) 'Face image generation of anime characters using an advanced first order motion model with facial landmarks', *IEICE TRANSACTIONS on Information and Systems*, Vol. 106, No. 1, pp.22–30.

Paulin, G. and Ivasic-Kos, M. (2023) 'Review and analysis of synthetic dataset generation methods and techniques for application in computer vision', *Artificial Intelligence Review*, Vol. 56, No. 9, pp.9221–9265.

Peng, H.Y., Zhang, J.P., Guo, M.H. et al. (2024) 'Charactergen: efficient 3d character generation from single images with multi-view pose canonicalization', *ACM Transactions on Graphics* (*TOG*), Vol. 43, No. 4, pp.1–13.

Pikulkaew, K., Boonchieng, W., Boonchieng, E. et al. (2021) '2D facial expression and movement of motion for pain identification with deep learning methods', *IEEE Access*, Vol. 9, No. 1, pp.109903–109914.

Qinran, Y.I.N. and Weiqun, C.A.O. (2021) 'Video-driven 2D character animation', *Chinese Journal of Electronics*, Vol. 30, No. 6, pp.1038–1048.

Qiu, S. (2023) 'Generative AI processes for 2D platformer game character design and animation', *Lecture Notes in Education Psychology and Public Media*, Vol. 29, No. 1, pp.146–160.

Siyao, L., Li, Y., Li, B. et al. (2022) 'Animerun: 2d animation visual correspondence from open source 3D movies', *Advances in Neural Information Processing Systems*, Vol. 35, No. 10, pp.18996–19007.

Song, W., Zhang, X., Guo, Y. et al. (2023) 'Automatic generation of 3D scene animation based on dynamic knowledge graphs and contextual encoding', *International Journal of Computer Vision*, Vol. 131, No. 11, pp.2816–2844.

Sun, Z., Lv, T., Ye, S. et al. (2024) 'Diffposetalk: speech-driven stylistic 3d facial animation and head pose generation via diffusion models', *ACM Transactions on Graphics* (*TOG*), Vol. 43, No. 4, pp.1–9.

Sung, R. (2024) 'Changes in 2D animation production methods due to technological advancements', *Journal of Information Technology Applications and Management*, Vol. 31, No. 4, pp.139–148.

Tian, G. (2025) 'Rule engine and neural network: reproduction and analysis of traditional festival celebration elements in animation', *International Journal of Information and Communication Technology*, Vol. 26, No. 7, pp.48–62.

Tous, R. (2023) 'Pictonaut: movie cartoonization using 3D human pose estimation and GANs', *Multimedia Tools and Applications*, Vol. 82, No. 14, pp.21101–21115.

Wang, X., Zhang, S., Gao, C. et al. (2025) 'Unianimate: taming unified video diffusion models for consistent human image animation', *Science China Information Sciences*, Vol. 68, No. 10, pp.1–14.

Wickramasinghe, M.M.T. and Wickramasinghe, M.H.M. (2021) 'Impact of using 2D animation as a pedagogical tool', *Psychology and Education Journal*, Vol. 58, No. 1, pp.3435–3439.

Wu, R., Su, W., Ma, K. et al. (2025) 'Aniclipart: clipart animation with text-to-video priors', *International Journal of Computer Vision*, Vol. 133, No. 6, pp.3149–3165.

Wu, X., Zhang, Q., Wu, Y. et al. (2021) 'F³a-gan: Facial flow for face animation with generative adversarial networks', *IEEE Transactions on Image Processing*, Vol. 30, No. 11, pp.8658–8670.

Xu, P., Zhu, Y. and Cai, S. (2022) 'Innovative research on the visual performance of image two-dimensional animation film based on deep neural network', *Neural Computing and Applications*, Vol. 34, No. 4, pp.2719–2728.

Yang, Q., Guan, J., Wang, K. et al. (2024) 'Showmaker: creating high-fidelity 2D human video via fine-grained diffusion modeling', *Advances in Neural Information Processing Systems*, Vol. 37, No. 10, pp.51039–51062.

Yin, H., Liu, J., Chen, X. et al. (2022) 'WeAnimate: Motion-coherent animation generation from video data', *Multimedia Tools and Applications*, Vol. 81, No. 15, pp.20685–20703.

Yu, J. (2025) 'Design of a neural network-based automated style migration technique in digital media art', *J. Combin. Math. Combin. Comput.*, Vol. 127, No. 4, pp.9561–9576.

Yu, S., Chen, Y., Guo, T. et al. (2024) 'Parametric design algorithm guided by machine vision in animation scene design', *Computer-Aided Des. Appl.*, Vol. 21, No. S15, pp.261–275.

Zarif, S., Amin, K.M., Najjar, A. et al. (2025) 'Animating text descriptions into characters: a comparative review of generative models', *IJCI. International Journal of Computers and Information*, Vol. 12, No. 1, pp.43–66.

Zhang, L., Zhou, Q.H., Tang, S. et al. (2025) 'High-definition multi-scale voice-driven facial animation: enhancing lip-sync clarity and image detail', *The Visual Computer*, Vol. 41, No. 7, pp.4395–4403.

Zhang, N. and Pu, B. (2024) 'Film and television animation production technology based on expression transfer and virtual digital human', *Scalable Computing: Practice and Experience*, Vol. 25, No. 6, pp.5560–5567.

Zhang, X., Xu, J. and Yang, W. (2022) 'Application and practice of 2D animation in Chinese traditional elements based on deep learning model', *Mobile Information Systems*, Vol. 2022, No. 1, p.6453320.

Zhao, J. and Zhao, X. (2022) 'Computer-aided graphic design for virtual reality-oriented 3D animation scenes', *Computer-Aided Design and Applications*, Vol. 19, No. 1, pp.65–76.

Zhao, Y., Ren, D., Chen, Y. et al. (2022) 'Cartoon image processing: a survey', *International Journal of Computer Vision*, Vol. 130, No. 11, pp.2733–2769.