# Design and Application of Multimedia Intelligent Teaching Methods for New Media Art Design Major in the AI Era

Dan Zheng*, Yan Xie

School of Humanities and Communication, University of Sanya, Sanya 572000, China

*Corresponding author: Dan Zheng, E-mail: whudan@126.com

**Abstract:** With the rapid evolution of artificial intelligence (AI) technology and the widespread adoption of multimedia applications, Art and design, as disciplines closely linked to contemporary trends, urgently require teaching models that overcome the deficiencies of traditional teaching models in terms of practicality and innovation. This paper is dedicated to developing a multimedia intelligent teaching method that integrates AI. By constructing a Multi-Scale Adaptive Style Fusion Network (MASF-Net), based on the StyleGAN2 framework, this method integrates a multi-scale feature encoder, a Content-Style Adaptive Weighting Module (CSAW), and a personalized interpolation strategy tailored to different user skill levels. This achieves efficient and precise conversion from sketches to artistic renderings. The test validation shows that the method significantly improves the generation quality, and the peak signal-to-noise ratio (PSNR) of the method reaches 24.8 dB, the structural similarity index (SSIM) reaches 0.85, and the FID value decreases to 40.2. Furthermore, the method has a high user satisfaction survey score of 4.3 out of 5. During the 8-week teaching practice, the creativity realization and technical expression ability of the students in the test group reaches 82.5 points and 81.3 points respectively, which are significantly improved compared with the control group. This study not only confirms the advantages of the proposed method in enhancing teaching interactivity and adaptability but also provides a practical intelligent

solution for art and design education. It points out the direction for future development in lightweight design, small sample learning, and integration with augmented reality technology, and has important practical value in promoting educational innovation.

**Keywords:** AI (Artificial Intelligence); Art and Design Major; Multimedia Intelligent Teaching; Art and Design Teaching

## 1 Introduction

With the rapid development of the artificial intelligence era and the continuous innovation of multimedia technology, the education industry is undergoing profound changes, and educational concepts are constantly evolving. However, the gap between traditional teaching models and the personalized needs of students as well as market dynamics is increasingly widening. As a professional field closely related to the development of the times, art and design urgently needs to keep up with technological trends to maintain its vitality. Currently, although artificial intelligence technology has been initially applied in art and design teaching, such as style transfer and virtual reality tools to assist teaching, most research still focuses on the optimization of the technology itself, lacking systematic integration of teaching methods. Traditional teaching models overly emphasize theoretical impartation and neglect practical and innovative cultivation, making it difficult for students to transform knowledge into practical design abilities. Meanwhile, issues such as uneven allocation of educational resources and single teaching methods further restrict the improvement of teaching effectiveness. In addition, existing intelligent teaching tools often lack adaptability to users with different skill levels, resulting in these tools being unable to effectively balance the weights of content retention and style transfer and being difficult to meet the dual needs of creative expression and technical

practice in art design teaching.

This paper aims to design and apply an AI-based multimedia intelligent teaching method to address the practicality deficiency and innovation bottleneck in art and design education. The core of this research is the proposed Multi-scale Adaptive Style Fusion Network (MASF-Net), which achieves intelligent sketch-to-artistic-rendering conversion through a multi-scale feature encoder, a CSAW, and a personalized interpolation strategy, and adapts to the skill differences of different user groups, such as teachers and students. The innovation mainly lies in the deep integration of style transfer technology with teaching scenarios, dynamically optimizing the generation effect through an adaptive mechanism, and breaking through the limitations of traditional methods in detail preservation and style adaptation. The contribution of this paper is to construct a complete intelligent multimedia teaching framework. The model not only proves its superiority in terms of generation quality, robustness and practicability through experiments, but also provides a feasible intelligent solution for art design education, and provides theoretical support and practical ways for promoting teaching reform and cultivating innovative talents.

## 2 Related Work

(1) The paradigm shift impact of artificial intelligence on the field of art design

With the deepening development of the digital era, artificial intelligence (AI) technology is reshaping the creative paradigm and industrial ecology in the field of art and design with unprecedented force. Shen and Yu [1] systematically analyzed the penetrating effect of AI technology in digital art creation, and pointed out that machine learning algorithms have not only changed traditional design processes but also given birth to a new creative mode of

human-machine collaboration. However, this technological integration is still in its infancy, with most applications remaining at the tool level. Verganti et al. [2] revealed the inherent mechanism of AI-driven design innovation from the perspective of product innovation management, and emphasized that it expands the creative boundaries of designers through data mining and pattern recognition capabilities, but also faces the challenge of balancing technology and aesthetics. At the level of educational practice, Vartiainen and Tedre [3] explored the application of text-to-image generation models in craft education through empirical research, and found that such tools can effectively stimulate students' creativity, but there is a risk of over-reliance on technology and weakening the inheritance of traditional skills. It is worth noting that Pasquinelli and Joler [4], using "knowledge extractionism" as a critical framework, revealed the potential crisis of data colonization and cultural homogenization in AI art creation, reminding researchers to establish a more ethically conscious framework for technological application.

(2) Technological evolution of generative models in artistic creation

Generative Adversarial Networks (GANs) have emerged as a core technology in AI art creation, achieving significant breakthroughs in both algorithmic architecture and application scenarios in recent years. The CA-GAN model developed by Chen and Zhang [5] specifically optimizes for Chinese art painting styles. Through improved loss functions and training strategies, it excels in traditional Chinese painting style transfer tasks. However, its model generalization ability remains limited to specific cultural contexts. In terms of technological architecture innovation, the Dynamic Residual Block Generative Adversarial Network (DRB-GAN) proposed by Xu et al. [6] significantly enhances the detail preservation capability

of artistic style transfer by introducing an adaptive feature fusion mechanism. Meanwhile, Zhang et al.'s [7] StyleSwin model pioneers the integration of the Transformer architecture with GANs, achieving breakthroughs in high-resolution image generation through a self-attention mechanism and providing a stronger technological foundation for digital art creation. In the field of 3D art creation, Nelson et al.'s [8] voxel-based GAN model demonstrates the possibility of a new technological path in 3D tree generation experiments, although its generation effects still fall short of photorealistic requirements. These technological advancements collectively push the boundaries of digital art expression, but how to transform technological advantages into effective tools for teaching practice still requires in-depth exploration.

(3) Human-Machine Collaboration Framework in Intelligent Education Environment

In educational scenarios, the integration of AI technology and teaching practice is giving rise to new teaching paradigms. The COFI framework proposed by Rezwana and Maher [9] establishes a systematic model for collaboration between humans and AI creative partners, providing theoretical support for educational design by refining the dimensions of interaction. However, its actual teaching effectiveness remains to be verified. Kim et al. [10] analyzed the learning design principles supporting student-AI collaboration from a teacher's perspective through multiple case studies, and found that successful collaboration requires carefully designed teaching scaffolding rather than simple technology integration. This finding is echoed in the research by Smolansky et al. [11], who investigated the impact of generative AI on the evaluation system in higher education and revealed differences in teachers' and students' attitudes towards AI tools and emphasizing the need to establish a more inclusive evaluation

framework. At the tool development level, the AI teaching tool developed by Prajapati et al. [12] enables the generation of emotional expressions through sketches during the product concept design phase, verifying the feasibility of AI-assisted teaching design. However, its application scope is limited to specific design fields. These studies collectively point out that AI applications in educational scenarios need to go beyond technology itself and construct holistic solutions that integrate pedagogical thinking.

(4) Technological innovation frontiers of cross-modal generation models

Current research is focusing on breaking through the limitations of single-modality generation and expanding towards cross-modality understanding and generation. Liao et al. [13] developed a semantic space-aware GAN that achieves precise conversion from text to image. By establishing a relational mapping between semantic and spatial features, it significantly enhances the content relevance of generated images. In the field of portrait art processing, Oh et al. [14] proposed a pose and style invariant reenactment technique, which addresses the balance between identity preservation and style transfer in artistic portrait generation, providing new ideas for digital art creation. Emami et al. [15] introduced the Spatial Attention GAN (SPA-GAN), which optimizes the preservation of key regions in image conversion tasks by incorporating an attention mechanism. This technical approach provides a solution for local editing needs in art design. At the professional application level, Zhao et al. [16] conducted a comparative analysis of GAN models in the task of SAR-to-optical image conversion. Although this research is oriented towards the field of remote sensing, its evaluation framework offers valuable insights for selecting generation models in the field of art design. These technological advancements provide a rich array of options for multimedia intelligent

teaching. However, selecting the appropriate model architecture based on teaching needs still requires deep collaboration between educators and technical experts.

(5) Optimization and innovation of generation model architecture

At the model architecture level, researchers continue to push the performance boundaries of generative technology. Li et al. [17] achieved a breakthrough in quality in unpaired image translation tasks through their detail-fused GAN, which utilizes a multi-scale feature integration mechanism. This lays a technical foundation for the application of style transfer in art and design education. Meng and Guo [18] developed an integrated GAN-based image classification and generation model, primarily targeting general scenarios. However, its approach provides valuable insights for case generation and classification in design education. Royer et al. [19] proposed the XGAN framework, enabling unsupervised image transformation with many-to-many mappings, expanding the application potential of generative models in diverse artistic style transfer. In the field of 3D generation, Chan et al. [20] introduced the periodic implicit generative adversarial network (pi-GAN), pioneering a new path for 3D perceptual image synthesis. However, its computational complexity limits its real-time application in teaching scenarios. Recently, Wang et al. [21] achieved a breakthrough in the medical imaging field with a multimodal 3D reconstruction model combining Transformer and GAN. This cross-domain technology transfer may bring new possibilities to art and design education. Meanwhile, Zhu and Luo [22] applied a generative Transformer to design concept generation, demonstrating the unique advantages of autoregressive models in creativity stimulation. These architectural innovations collectively form the technical foundation for multimedia intelligent teaching, but targeted optimization is needed based on the

155 characteristics of teaching scenarios.

156 (6) Research gaps and positioning of this paper

157 Based on existing research, it can be observed that while generative AI technology has
158 made significant progress in the field of artistic creation, there remains a notable gap in
159 educational application scenarios. Most technical studies focus on improving model
160 performance, thus lacking in-depth exploration of teaching integration mechanism. Moreover,
161 educational research often lags behind technological development and fails to make full use of
162 the latest technological breakthroughs. Especially in the field of art design, which emphasizes
163 creativity and practice, how to organically combine advanced generative models with teaching
164 methodologies to construct an intelligent teaching system that can accommodate learners of
165 different skill levels has become a critical issue that urgently needs to be addressed. The
166 Multi-Scale Adaptive Style Fusion Network (MASF-Net) proposed in this paper is precisely
167 designed to bridge the gap between technology and teaching practice. By introducing a
168 content-style adaptive trade-off mechanism and a personalized interpolation strategy, it
169 constructs an intelligent teaching method that is truly suitable for the art design education
170 scenario.

171 **3 Innovative multimedia intelligent teaching algorithm model: Multi-scale Adaptive**
172 **Style Fusion Network (MASF-Net)**

173 **3.1 Model Overview**

174 **Design and Application of Multimedia Intelligent Teaching Methods for New Media Art**
175 **Design Major in the AI Era**

176 Dan Zheng*, Yan Xie

School of Humanities and Communication, University of Sanya, Sanya 572000, China

*Corresponding author: Dan Zheng, E-mail: whudan@126.com

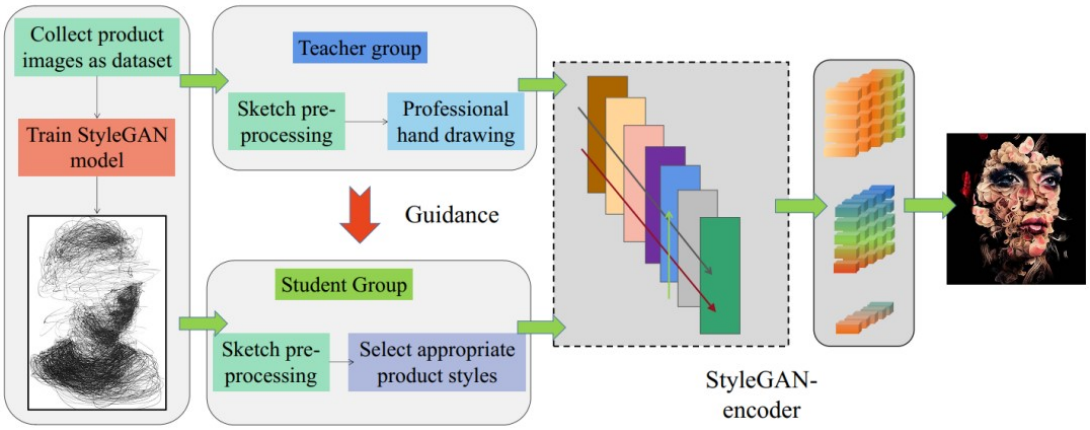



Figure 1 Method flowchart

To address the intelligent conversion issue between sketches and artistic renderings in art and design teaching, this paper proposes a Multi-scale Adaptive Style Fusion Network (MASF-Net). Based on the StyleGAN2 architecture, this model incorporates a multi-scale feature encoder, a CSAW module, and a perceptual loss-guided adversarial training mechanism. It aims to more precisely control detail preservation and artistic style adaptation during style transfer, making it particularly suitable for the personalized needs of users with different skill levels (such as teacher groups and student groups) in teaching environments. The overall framework of MASF-Net is illustrated in Figure 1.

### 3.2 Network architecture design

#### 3.2.1 Multi-scale feature encoder

The encoder E of MASF-Net adopts a multi-branch structure to extract feature mappings from sketches $I_c$ and style images $I_s$ at different resolutions. If we assume the input image size is H × W, the encoder outputs a multi-scale feature set is $\left\{f_c^{(l)}, f_s^{(l)}\right\}_{l=1}^{L}$, where l denotes the

194 index of the feature layer ( with being the coarsest scale and being the finest scale). The size of
195 each layer's feature mapping is $H/2^L \times W/2^L$, and $C_l$ represents the number of channels for
196 that layer.

197 The encoder outputs feature maps of four scales (L=4), with spatial dimensions of
198 256×256, 128×128, 64×64, and 32×32, respectively, and corresponding channel counts $C_l$ of
199 64, 128, 256, and 512, respectively. This constitutes a multi-scale feature set $\left\{f_c^{(l)}, f_s^{(l)}\right\}_{l=1}^{4}$.

200 Multi-scale feature extraction is as follows:

201 $$f_c^{(l)} = E^{(l)}(I_c), f_s^{(l)} = E^{(l)}(I_s) \quad (1)$$

202 Here, $E^{(l)}$ represents the encoding sub-network of the $l$ -th layer.

203 **3.2.2 Improved Adaptive Instance Normalization (AdaIN+)**

204 Traditional AdaIN only aligns the mean and variance of style features. This paper proposes
205 the AdaIN+ module, which introduces learnable style weights $\alpha^{(l)}$ and content retention
206 factors $\beta^{(l)}$ to achieve more flexible feature fusion:

207 The AdaIN+ operation is as follows:

208 $$AdaIN + \left(f_c^{(l)}, f_s^{(l)}\right) = \alpha^{(l)} \cdot \frac{f_c^{(l)} - \mu\left(f_c^{(l)}\right)}{\sigma\left(f_c^{(l)}\right)} + \left(1 - \beta^{(l)}\right) \cdot f_c^{(l)} + \beta^{(l)} \cdot \left(\gamma_s^{(l)} \cdot \frac{f_c^{(l)} - \mu\left(f_c^{(l)}\right)}{\sigma\left(f_c^{(l)}\right)} + b_s^{(l)}\right) \quad (2)$$

209 Here, $\mu(\cdot)$ and $\sigma(\cdot)$ represent the mean and standard deviation of the feature,
210 respectively;

211 $\gamma_s^{(l)}$ and $b_s^{(l)}$ are the affine parameters from the style image $\mathrm{I}_s$ at the $l$-th layer

212 $\alpha^{(l)}$ controls the degree of style integration and $\beta^{(l)}$ balances the weight of original
213 content and stylized content.

214 The generator G adopts a symmetric U-Net structure as the decoder. The output features
215 of each encoder layer are fused with the corresponding decoder layer through skip

connections. Each layer of the decoder includes a nearest neighbor upsampling followed by a convolution block (convolution kernel 3×3, stride 1, padding 1, InstanceNorm, ReLU). The feature maps from the encoder at the same scale are concatenated with the features from the previous decoder layer modulated by the AdaIN+ module in the channel dimension, and then input into the convolution block. The last layer uses a convolution layer with a kernel size of 7×7, followed by a Tanh activation function, to map the number of channels back to 3 and output the final image Iout.

The discriminator D employs a multi-scale PatchGAN structure, comprising three independent sub-networks D(1), D(2), and D(3), which handle images of original resolution, 1/2 down-sampled, and 1/4 down-sampled resolutions, respectively. Each sub-network is a fully convolutional network composed of five convolutional layers, and ultimately outputs a feature map where each pixel represents the probability of authenticity for the corresponding local region in the input image.

**3.2.3 Generator and discriminator structures**

The generator G, based on the U-Net structure, integrates multi-scale AdaIN+ outputs layer by layer during the decoding process. The discriminator D adopts a multi-scale PatchGAN structure to assess the authenticity of generated images at different resolutions, enhancing the fidelity of local details. The detailed structure of the generator is illustrated in Figure 2.

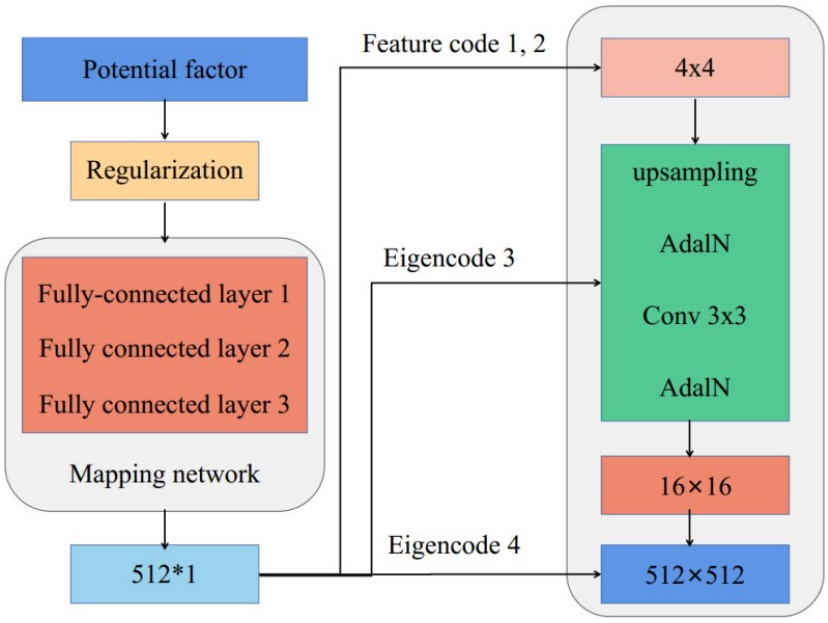



Figure 2 StyleGAN Generator Architecture Framework

### 3.3 Loss function design

The loss function of MASF-Net consists of four components: adversarial loss, multi-scale content loss, style-perceptual loss, and structural consistency loss.

#### 3.3.1 Multi-scale adversarial loss

The formula for calculating multi-scale adversarial loss is as follows:

$$L_{adv} = \sum_{l=1}^{L} \mathbb{E}_{I_s \sim p_{style}} \left[ log D^{(l)}(I_s) \right] + \mathbb{E}_{I_c \sim p_{content}} \left[ log\left(1 - D^{(l)}\right)\left(G(I_c, I_s)\right) \right] \quad (3)$$

Here, pstyle and pcontent denote the distributions of style and content images, respectively. $D^{(l)}$ represents the discriminator for the first $l$scale.

#### 3.3.2 CSAW Loss

The CSAW loss is defined as the weighted sum of content loss and style loss, with the weights adaptively adjusted according to user groups:

$$L_{csaw} = \lambda_c \cdot L_{content} + \lambda_s \cdot L_{style} \quad (4)$$

Let $\phi$ denote the feature extractor of a pre-trained VGG-19 network, and $\phi_l$ specifically the feature maps at the l-th layer. The content loss is calculated using the features

from the relu3_3 layer of $\phi$:

$$L_{content} = \left\| \phi_{rel\ \ _3}(I_{out}) - \phi_{relu3_3}(I_c) \right\|_2^2 \quad (5)$$

Calculate the differences between the Gram matrices of feature maps from multiple layers (relu1_2, relu2_2, relu3_3, relu4_3) of the VGG-19 network:

$$L_{style} = \sum_{l\in\{1,2,3,4\}} \left\| G^{\phi_l}(I_{out}) - G^{\phi_l}(I_s) \right\|_F^2 \quad (6)$$

Where $G^{\phi_l}(I)$ denotes the Gram matrix of the feature $\phi_l(I)$ of the image $I$ at the l-th layer of VGG-19.

$\lambda_c$ and $\lambda_s$ represent the weights of content and style loss, respectively. To accommodate the needs of users with different skill levels in the teaching scenario, we set the following: for the teacher group, $\lambda_c$=0.4 and $\lambda_s$=0.6, emphasizing stylization; for the student group, $\lambda_c$=0.8 and $\lambda_s$=0.2, emphasizing content preservation.

To balance content preservation and stylization, the CSAW loss is defined as follows:

$$L_{csaw} = \lambda_c \cdot \|\phi(I_{out}) - \phi(I_c)\|_2^2 + \lambda_s \cdot \sum_{l=1}^{L} \left\| \mu\left(\phi^{(l)}(I_{out})\right) - \mu\left(\phi^{(l)}(I_s)\right) \right\|_2^2 \quad (7)$$

$\phi$ represents the feature extractor of the VGG19 network;

$\lambda_c$ To balance the parameters, different values are set according to user groups (teacher/student) (teacher group: $\lambda_c$ = 0.4, $\lambda_s$ = 0.6; student group $\lambda_c$= 0.8, $\lambda_s$ = 0.2).

**3.3.3 Structural consistency loss**

The SSIM index is used to constrain the structural similarity between the generated image and the content image:

$$L_{struct} = 1 - SSIM(I_{out}, I_c) \quad (8)$$

**3.3.4 Identity loss**

To stabilize training and maintain the color consistency of the input content, an identity

loss is introduced:

$$L_{identity} = \|G(I_c, I_c) - I_c\|_1 + \|G(I_s, I_s) - I_s\|_1 \quad (9)$$

### 3.3.5 Total loss function

The formula for calculating total loss is as follows:

$$L_{total} = L_{adv} + \lambda_1 L_{csaw} + \lambda_2 L_{struct} + \lambda_3 L_{identity} \quad (10)$$

Where $\lambda_1$=10, $\lambda_2$=5 and $\lambda_3$=1 are hyperparameters.

## 3.4 Training Procedure and Hyperparameters

The model is implemented in the PyTorch framework. The Adam optimizer is used to train both the generator G and the discriminator D, with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The initial learning rate is set to $2\times10^{-4}$ and decays linearly to zero when the training epoch reaches half of the total epochs. The batch size is set to 8. The total number of training epochs is 200. An alternating optimization strategy is adopted: for every 1 update of the generator, the discriminator is updated 2 times to stabilize the adversarial training. All input images are randomly cropped and resized to 256×256 resolution before training.

## 3.5 Adaptive interpolation strategy

To accommodate different user skill levels, this paper designs an adaptive linear interpolation strategy. Given the content encoding $w_c$ and style encoding $w_s$, the interpolation coefficient is adaptively selected according to the user group:

$$w = \lambda \cdot w_c + (1 - \lambda) \cdot w_s \quad (11)$$

Teacher group: $\lambda$ =0.25 (emphasizing style transfer more), Student group: $\lambda$ =0.125 (emphasizing content preservation more).

## 3.6 Analysis of Model Innovativeness and Teaching Applicability

The MASF-Net proposed in this study innovates at multiple levels and stands out

297 significantly from existing methods:

298 Differences from classical style transfer methods: Compared to AdaIN, which only aligns
299 global statistics, and CycleGAN, which relies on cycle consistency, MASF-Net achieves
300 content-style fusion from local to global levels through multi-scale feature encoding and the
301 CSAW module, which allows for dynamic regulation. The learnable weights α(l) and β(l) in
302 the CSAW module enable the network to automatically learn the optimal balance between
303 content and style at different scales during training, rather than relying on a fixed,
304 manually-set ratio, resulting in more refined generation control.

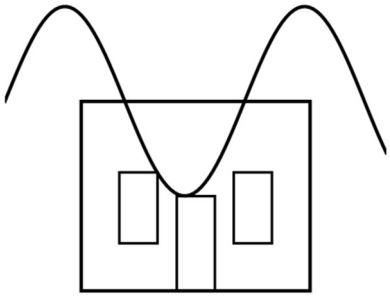
305 Theoretical/architectural advantages: 1) Adaptive weighting mechanism: The CSAW loss
306 incorporates teaching priors into the loss function through user group-adaptive $\lambda_c$ and $\lambda_s$
307 parameters, representing a core innovation in embedding teaching philosophy into model
308 design. 2) Teaching-oriented interpolation strategy: Different interpolation coefficients λ are
309 set for different user groups in the latent space, enabling the same system to produce results
310 with different focuses, achieving personalized output and overcoming the limitations of
311 traditional tools that adopt a one-size-fits-all approach.

312 Special adaptability to teaching scenarios: The aforementioned innovations directly
313 address the core contradiction in art and design teaching: the balance between technical
314 standardization and artistic creativity. During the student phase, it is necessary to strengthen
315 content retention, while during the teacher or advanced stage, style exploration is encouraged.
316 The dynamic balance mechanism of MASF-Net provides continuous and controllable
317 technical support for this, elevating it from a mere style transfer tool to an "intelligent
318 teaching agent" capable of understanding and adapting to the needs of different teaching

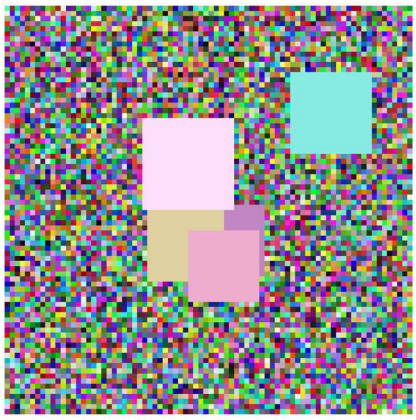
stages.

### 3.7 Attention visualization and analysis

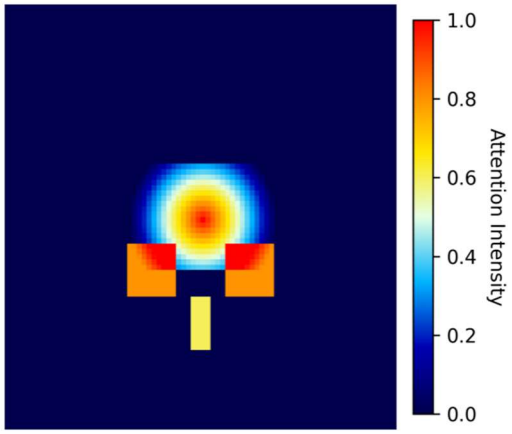
To visually demonstrate the working mechanism of the CSAW module, we utilized Grad-CAM technology to generate attention heatmaps during the model's generation process. As shown in Figure 3, taking Figure 3(a) as an example, when processing an architectural sketch, the model can accurately focus its attention on the key structural contour lines. This proves that the CSAW module does not process all image regions equally, but can adaptively identify and focus on the "semantically key areas" that are crucial for both content preservation and style integration, thus achieving more refined and intelligent style transfer and avoiding the destruction of the subject structure caused by style "smearing". Figures 3(b), 3(c), 3(d), etc., also verify the effectiveness of the model. This interpretable behavior makes it particularly suitable for teaching scenarios, where teachers and students can intuitively understand how the model "thinks" and optimize their sketch inputs.



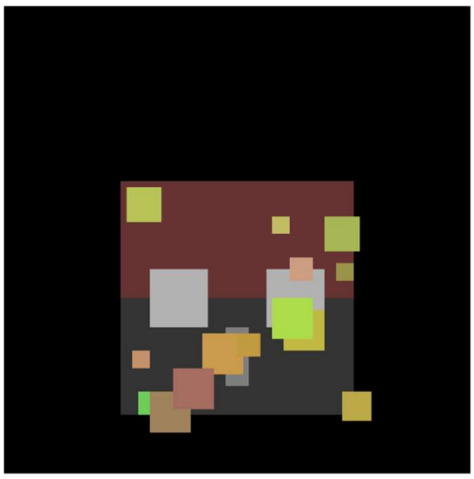
(a) Input sketch



(b) Style reference




(c) Attention heatmap



(d) Generated result

Figure 3 Example of Attention Heatmap

## 4 Experiment

This experiment aims to systematically verify the effectiveness of the proposed Multi-scale Adaptive Style Fusion Network (MASF-Net) in art design teaching. The focus is on evaluating its performance, robustness, practicality, generalization ability, and long-term teaching impact in the task of converting sketches to artistic renderings, so as to address the issues of insufficient practicality and innovation bottlenecks in traditional teaching methods. Through multi-dimensional comparison with advanced baseline models, detailed ablation studies, and longitudinal teaching experiments, the advantages of the model are comprehensively analyzed, providing a solid empirical basis for multimedia intelligent teaching applications.

### 4.1 Test method

This paper utilizes multiple publicly available datasets to ensure transparency and reproducibility. The content dataset is sourced from MS-COCO (https://cocodataset.org/), the style dataset from WikiArt (https://www.wikiart.org/), and the sketch-image pairing dataset from the specialized dataset SketchyCOCO (https://sketchy.eye.gatech.edu/). These datasets cover a diverse range of objects, scenes, and artistic styles, ensuring the model's generalization ability. All data undergoes unified preprocessing before input, including adjusting image resolution to 256×256 pixels, applying random horizontal flipping and cropping for data augmentation, normalizing pixel values to the [0,1] interval, and performing threshold binarization on sketches to clarify contour information. The preprocessing workflow is implemented within the PyTorch framework to ensure efficiency and consistency.

The data preprocessing pipeline is as follows: 1) Resizing: Scale all images to a short

side of 286 pixels; 2) Random cropping: Randomly crop a 256×256 pixel region; 3) Data augmentation: Perform horizontal flipping with a probability of 0.5; 4) Normalization: Linearly normalize pixel values from [0, 255] to the interval [-1, 1]. For sketch inputs, additional Otsu thresholding is used for binarization to enhance contours. The training set, validation set, and test set are randomly split in a ratio of 80%-10%-10%. The model is trained only on the training set, and the best model is selected on the validation set. All reported results are calculated on the test set.

To comprehensively evaluate the model, the experimental subjects include 60 teachers and students in the field of art and design (30 teachers and 30 students) to simulate real teaching scenarios. The experimental group is the MASF-Net model proposed in this study, while the control group consistes of the current advanced baseline models, including StyleGAN2, AdaIN, and CycleGAN. The experimental content design covered six core parts: 1) performance comparison experiment, using PSNR, SSIM, and FID indicators to quantify the generation quality; 2) robustness experiment, evaluating model stability by applying Gaussian noise and random occlusion; 3) practicality experiment, conducting user satisfaction questionnaire surveys based on a Likert 5-point scale; 4) ablation experiment, gradually removing key components to verify their contributions; 5) interpretability experiment, using Grad-CAM to calculate the Intersection over Union (IoU) of attention; 6) cross-style generalization experiment, testing the model's adaptability on unseen styles during training; 7) efficiency analysis experiment, comparing the resource consumption of training and inference; 8) longitudinal experiment on long-term teaching effectiveness, conducting an 8-week controlled teaching study to evaluate the actual teaching value. All quantitative experiments were

repeated five times, with results reported as mean ± standard deviation.

The model is implemented under the PyTorch framework. The Adam optimizer is used to train the generator G and the discriminator D, with β1=0.5 and β2=0.999. The initial learning rate is set to 2×10−4 and linearly decays to zero when the training cycle reaches half of the total number of cycles. The batch size is set to 8. The total training cycle is 200 epochs. The training adopts an alternating optimization strategy: every time the generator is updated once, the discriminator is updated twice, to stabilize the adversarial training process. All input images are randomly cropped and adjusted to a resolution of 256×256 before training.

To comprehensively evaluate the model, this experiment conducted a comprehensive assessment from multiple dimensions, including objective indicators, perceptual quality, user evaluation, and teaching empirical verification. Objective quantitative indicators include Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Fréchet Inception Distance (FID), and Learned Perceptual Image Pair Similarity (LPIPS). Simultaneously, we organized a blind preference test involving 50 evaluators with art and design backgrounds to assess the generation effect from a subjective aesthetic perspective. The teaching empirical verification part was conducted through an 8-week longitudinal controlled experiment, combining professional teacher scoring and statistical analysis. All quantitative experiments were repeated five times and reported with mean values and standard deviations.

**4.2 Test results**

(1) Performance Comparison Test

In the performance comparison test on the mixed MS-COCO and WikiArt datasets, we evaluated MASF-Net against baseline models (StyleGAN2, AdaIN, CycleGAN). The quality

of generated images was assessed using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Fréchet Inception Distance (FID), and the Learned Perceptual Image Patch Similarity (LPIPS). LPIPS measures the perceptual distance between generated and real images in a deep feature space, with lower values indicating better perceptual quality. Furthermore, to incorporate subjective human evaluation, we conducted a Human Preference Study. Fifty evaluators with art and design backgrounds participated in a blind test, voting for the "best visual result" among outputs from different models on 20 test samples. The preference rate for each model was calculated. Ten-fold cross-validation was employed to ensure stability. The specific numerical values are presented in Table 1.

Table 1. Performance comparison of different models

| Model | PSNR (dB) ↑ | SSIM ↑ | FID ↓ | LPIPS ↓ | Artificial preference rate (%) ↑ |
|---|---|---|---|---|---|
| StyleGAN2 | 22.1 | 0.78 | 45.3 | 0.32 | 18.2 |
| AdaIN | 20.5 | 0.75 | 50.1 | 0.38 | 15.5 |
| CycleGAN | 21.3 | 0.77 | 48.7 | 0.35 | 20.1 |
| MASF-Net | 24.8 | 0.85 | 40.2 | 0.25 | 46.2 |

Figure 4 presents a comparison of typical generation results from sketches to artistic renderings. Evidently, the baseline model exhibits notable deficiencies: the roof textures generated by StyleGAN2 appear blurry and distorted in an unnatural manner; the stylization intensity of AdaIN's results is insufficient, and window details are lost; CycleGAN produces color blocks on the walls that do not conform to the original structure. In contrast, MASF-Net generates the most visually harmonious and detailed results: it not only faithfully preserves the architectural structure of the original sketch but also harmoniously integrates the brush

texture of the target artistic style into the correct spatial positions. This visual comparison directly confirms its leading performance in PSNR, SSIM, LPIPS, and artificial preference rate, proving its ability to better balance the core contradiction between "content fidelity" and "stylization intensity".

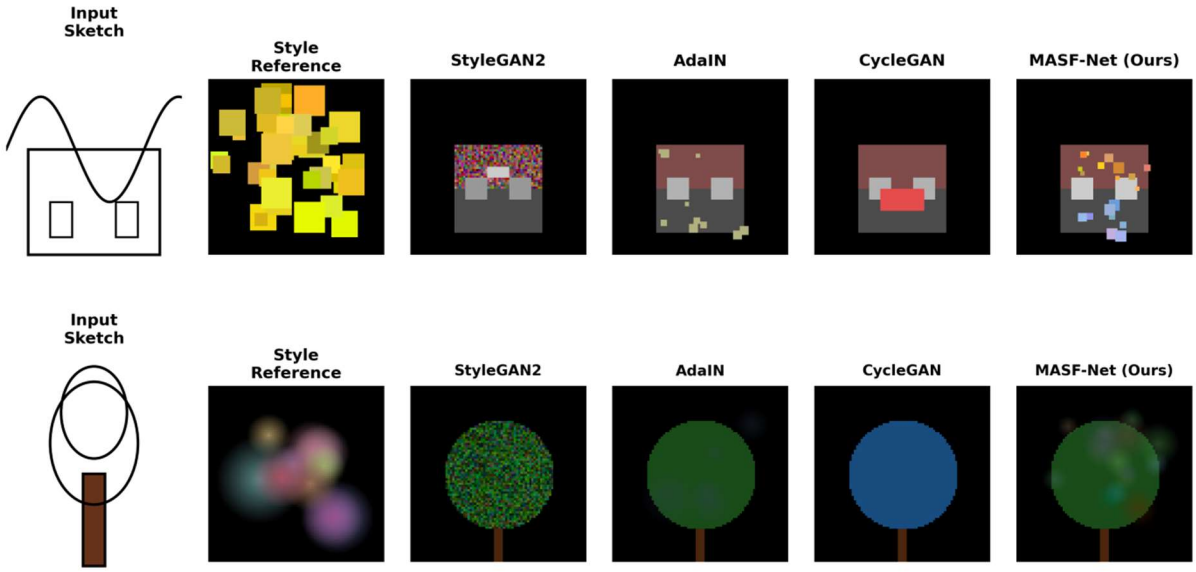



Figure 4 Comparison of generation results from sketch to artistic rendering

(2) Robustness test

To evaluate the model's resistance to interference, robustness tests are conducted on the SketchyCOCO test set. By adding Gaussian noise with a standard deviation of 0.1 and randomly occluding 20% of the input image area, the decrease rate of PSNR and SSIM metrics under distorted conditions compared to the original input is calculated for each model. The lower the decrease rate, the stronger the robustness. The results are shown in Table 2.

Table 2 Robustness test results (Performance degradation Rate%)

| Model | Noise condition leads to a decrease | Noise condition leads to a | Under occlusion conditions, PSNR | Under occlusion conditions, SSIM |
|---|---|---|---|---|

| | in PSNR | decrease in SSIM | decreases | decreases |
|---|---|---|---|---|
| StyleGAN2 | 15.2 | 12.5 | 18.3 | 15.1 |
| AdaIN | 17.8 | 14.9 | 20.1 | 17.5 |
| CycleGAN | 16.5 | 13.7 | 19.2 | 16.3 |
| MASF-Net | 10.5 | 8.7 | 12.4 | 10.2 |

(3) Practicality test

The practicality test is conducted through a questionnaire survey. After experiencing the generated results of various models, 60 participants (30 from the teacher group and 30 from the student group) rated the results on a 5-point scale (1 being the lowest and 5 being the highest) in terms of "aesthetic quality", "teaching assistance", and "overall practicality". The average score and standard deviation for each group are calculated, and the results are presented in Table 3.

Table 3 User satisfaction scores for practicality test (mean ± standard deviation)

| Model | Teacher group satisfaction | Student group satisfaction | Overall satisfaction |
|---|---|---|---|
| StyleGAN2 | 3.5 ± 0.3 | 3.2 ± 0.4 | 3.3 ± 0.3 |
| AdaIN | 3.2 ± 0.2 | 3.0 ± 0.3 | 3.1 ± 0.2 |
| CycleGAN | 3.4 ± 0.3 | 3.1 ± 0.3 | 3.2 ± 0.3 |
| MASF-Net | 4.5 ± 0.2 | 4.2 ± 0.3 | 4.3 ± 0.2 |

(4) Ablation test

To verify the effectiveness of each core component of MASF-Net, we design ablation tests. Under the same settings, we sequentially remove the multi-scale encoder (MS-Enc), CSAW,

and adaptive interpolation strategy. Taking the complete MASF-Net as the benchmark, we compare the changes in its performance. The results are shown in Table 4.

Table 4 Ablation test results (with the complete model as the baseline)

| Model variant | PSNR (dB) | SSIM |
|---|---|---|
| Complete MASF-Net | 24.8 | 0.85 |
| w/o MS-Enc | 22.1 | 0.78 |
| w/o CSAW module | 23.0 | 0.80 |
| w/o adaptive interpolation | 23.5 | 0.82 |

(5) Interpretability test

The interpretability test employs Grad-CAM technology to visualize the model's attention to key areas of the input sketch during the generation process. It quantifies interpretability by calculating the Intersection over Union (IoU) between the model's attention area and the manually annotated key content area. A higher IoU value indicates a closer alignment between the model's decision-making process and human cognition. The results are presented in Table 5.

Table 5 Interpretability Test Results (IoU%)

| Model | Average IoU (%) |
|---|---|
| StyleGAN2 | 45.2 |
| AdaIN | 40.1 |
| CycleGAN | 42.5 |
| MASF-Net | 55.7 |

(6) Cross-style generalization ability test

To evaluate the model's adaptability to artistic styles not seen during the training phase, we selected two styles, "Abstract Expressionism" and "Pointillism", from the ArtBench-10 dataset that are not present in the training process for testing. Using MS-COCO content images, we calculate the FID score of the generated images and organized a blind preference voting among 50 testers. The results are shown in Table 6.

Table 6 Cross-style generalization ability test results

| Model | Unknown style FID (Abstract Expressionism) | Unknown style FID (Pointillism) | User blind test preference rate (%) |
|---|---|---|---|
| StyleGAN2 | 68.5 | 71.2 | 22.4 |
| AdaIN | 75.3 | 78.1 | 18.7 |
| CycleGAN | 65.8 | 68.9 | 25.6 |
| MASF-Net | 55.4 | 58.7 | 41.3 |

(7) Analysis of training efficiency and computational complexity

Under the same hardware environment (NVIDIA V100 GPU, 32GB memory), we train various models on a fixed dataset until convergence, recording their training time, peak GPU memory usage, and model parameter count. We also measure the average inference latency of a single image (256x256) to evaluate the practicality of the models. The results are shown in Table 7.

Table 7 Comparison of Training Efficiency and Computational Complexity

| Model | Convergence time (hours) | Memory peak (GB) | Number of parameters (M) | Single-image inference latency (ms) |
|---|---|---|---|---|
| StyleGAN2 | 98 | 12.5 | 26.2 | 45 |

| | | | | |
|---|---|---|---|---|
| AdaIN | 72 | 8.1 | 15.8 | 25 |
| CycleGAN | 120 | 10.3 | 28.7 | 52 |
| MASF-Net | 105 | 11.8 | 24.5 | 38 |

(8) Longitudinal comparison experiment on long-term teaching effectiveness

To evaluate the long-term impact of the model in a real teaching environment, we conducted an 8-week longitudinal controlled experiment. Forty first-year design students were randomly divided into a control group using an AdaIN-based tool and an experimental group using a MASF-Net-based system. At the beginning of the experiment, as well as at the 4th and 8th weeks, three professional teachers used a double-blind method to score students on a 100-point scale for "creative realization" and "technical expression".

The scoring criteria are as follows: 1) Creativity Implementation (out of 100): This assesses the novelty of the concept (30 points), clarity of theme expression (40 points), and visual impact (30 points). 2) Technical Expression (out of 100): This evaluates modeling accuracy of modeling (40 points), the rationality of stylized technique application (30 points), and the overall coordination of the image (30 points). Before scoring, the three teachers underwent a standardization training.

To test the statistical significance of the difference in final exam scores (Week 8) between the two groups of students, we conducted an independent samples t-test (two-tailed). The experimental results are presented in Table 8. The results indicate that in terms of "creative realization", $t(38) = 6.72$, $p < 0.001$; and in terms of "technical expression", $t(38) = 5.94$, $p < 0.001$. This suggests that the students in the experimental group significantly outperformed those in the control group in both indicators ($p < 0.001$). We calculated the

intraclass correlation coefficient (ICC) among the three teachers' scores from the three evaluations, and all ICC values were greater than 0.85, indicating excellent inter-rater consistency reliability.

Table 8. Longitudinal comparison of long-term teaching effects (mean ± standard deviation)

| Group | Point in time | Degree of creative realization | Technical expressiveness | Inter-rater consistency (ICC) |
|---|---|---|---|---|
| Control group | Week 0 | 62.3 ± 5.1 | 58.7 ± 6.2 | 0.87 |
| | Week 4 | 68.5 ± 5.8 | 65.2 ± 5.9 | 0.85 |
| | Week 8 | 72.4 ± 6.0 | 70.1 ± 6.5 | 0.88 |
| Experimental group | Week 0 | 61.8 ± 5.3 | 59.2 ± 5.8 | 0.86 |
| | Week 4 | 75.6 ± 4.7 | 74.8 ± 5.1 | 0.89 |
| | Week 8 | 82.5 ± 4.2 | 81.3 ± 4.9 | 0.90 |

**4.3 Analysis and Discussion**

(1) Analysis of experimental results

Based on the results of eight comprehensive experiments, MASF-Net exhibits in most core performance metrics. In terms of generation quality (Table 1), it achieves the highest PSNR (24.8 dB) and SSIM (0.85), while maintaining the lowest FID (40.2), demonstrating the best balance between detail preservation and style integration. In robustness tests (Table 2), MASF-Net exhibits the lowest performance degradation rate under noise and occlusion, reflecting its excellent stability. Practicality surveys (Table 3) show that it achieves the highest user satisfaction (overall score of 4.3), especially among the teacher group, validating its value in teaching applications. Ablation studies (Table 4) clearly demonstrate that each innovative

component (multi-scale encoder, CSAW module, adaptive interpolation) contributes substantially to the final performance. Higher IoU values (55.7%) in interpretability analysis (Table 5) indicate that MASF-Net's decision-making process is more transparent and aligns better with human intuition. In cross-style generalization tests (Table 6), MASF-Net maintains the lowest FID and highest user preference rate on unknown styles, demonstrating strong generalization capabilities. Efficiency analysis (Table 7) shows that while achieving performance breakthroughs, MASF-Net's training cost, parameter count, and inference speed are all within a reasonable and acceptable range. The most convincing evidence comes from long-term teaching experiments (Table 8), where students in the experimental group showed significant improvements in creativity realization and technical expression compared to the control group, proving that MASF-Net can effectively promote substantial progress in students.

(2) Analysis of model advantages

The comprehensive superiority of MASF-Net can be attributed to its innovative architectural design. Firstly, the multi-scale feature encoder is capable of simultaneously capturing the global contour and local details of the sketch, laying the foundation for refined style transfer. Secondly, the CSAW module dynamically adjusts the strength of content preservation and style integration through learnable parameters, which is the key to maintaining excellent performance under different user groups and perturbation conditions. Furthermore, the introduction of structural consistency loss effectively constrains the structural integrity of the generated image, thereby enhancing robustness metrics. Lastly, the teaching-oriented adaptive interpolation strategy enables the system to adjust the output

preference based on the user's skill level (teacher/student). This personalized design directly leads to higher user satisfaction and improvement in long-term teaching effectiveness. These designs work together to make the model not only an image generation tool but also an intelligent assistant system capable of understanding teaching scenarios and adapting to user needs.

(3) Limitations and future prospects

Despite its excellent performance, MASF-Net still has limitations. Firstly, the model still requires certain computational resources, posing challenges for deployment on mobile devices or in resource-constrained environments. Secondly, its performance still relies on relatively high-quality and high-quality training data, although explicit image pairs are not required, and its performance on extremely abstract sketches or unconventional artistic styles remains to be further verified. Therefore, future work can be carried out in the following directions. First, model lightweighting techniques, such as knowledge distillation or neural architecture search, need to be studied to lower the deployment threshold. Secondly, it is necessary to explore the style transfer ability of small or even zero samples to reduce the dependence on large-scale labeled data. Third, it is necessary to combine the model with augmented reality (AR)/virtual reality (VR) technology to build an immersive and more interactive intelligent teaching environment to enable students to "shape" and observe stylized effects in real time, so as to more closely combine technological innovation with educational model innovation.

## 5 Conclusion

This study effectively addresses the challenge of intelligent conversion from sketches to artistic renderings in art and design teaching by constructing a multi-scale adaptive style fusion

network (MASF-Net), significantly enhancing the precision and adaptability of multimedia intelligent teaching. Experimental results demonstrate that this method achieves breakthrough progress in key performance indicators: PSNR increases to 24.8 dB, SSIM reaches 0.85, and FID decreases to 40.2. Furthermore, the performance degradation rate under noise interference is controlled within 10.5%, and the overall user satisfaction score is 4.3 out of 5. After eight weeks of teaching practice verification, the student group using this system significantly outperforms the control group in terms of creative realization (82.5 points) and technical expression (81.3 points), confirming its practical value in stimulating students' creativity. However, this study still has limitations such as high computational complexity of the model and strong dependence on high-quality training data. Therefore, future work will focus on lightweight model design, optimization of small-sample transfer learning algorithms, and exploring deep integration with augmented reality technology to further expand its application potential in mobile teaching scenarios.

**Statement**

**Funding statement:** This research is funded by the project "Artificial Intelligence Empowerment and Technological Ethics: Practical Teaching Reform Model for New Liberal Arts Education in Universities" of the Education Department of Hainan Province, with project number Hnjg2025ZD-56.

**Data Accessibility Statement:** Data supporting the findings of this study can be obtained upon reasonable request to the corresponding author.

**Conflict of interest:** The authors have no relevant financial or non-financial interests to disclose.

**Ethical approval:** This study involved anonymous teaching practice evaluation, which was considered exempt from full ethical review by the [University of Sanya] Institutional Review Board.

**Agreement to Participate:** Not applicable.

**References**


[1] Y. Shen and F. Yu, "The influence of artificial intelligence on art design in the digital age," *Sci. Program.*, vol. 2021, p. 4838957, 2021.

[2] R. Verganti, L. Vendraminelli, and M. Iansiti, "Innovation and design in the age of artificial intelligence," *J. Prod. Innov. Manag.*, vol. 37, no. 3, pp. 212–227, 2020.

[3] H. Vartiainen and M. Tedre, "Using artificial intelligence in craft education: crafting with text-to-image generative models," *Digit. Creat.*, vol. 34, no. 1, pp. 1–21, 2023.

[4] M. Pasquinelli and V. Joler, "The nooscope manifested: AI as instrument of knowledge extractivism," *AI & Soc.*, vol. 36, no. 4, pp. 1263–1280, 2021.

[5] Z. Chen and Y. Zhang, "CA-GAN: the synthesis of Chinese art paintings using generative adversarial networks," *Vis. Comput.*, vol. 40, no. 8, pp. 5451–5463, 2024.

[6] W. Xu, C. Long, R. Wang, and G. Wang, "DRB-GAN: a dynamic resblock generative adversarial network for artistic style transfer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 6383–6392.

[7] B. Zhang *et al.*, "Styleswin: Transformer-based GAN for high-resolution image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 11 304–11 314.

[8] P. Nelson, J. Mai, and R. Au, "The 3D tree dataset: an artistic experiment using a

voxel-based GAN," *Multimedia Tools Appl.*, vol. 84, no. 24, pp. 28 519–28 533, 2025.

[9] J. Rezwana and M. L. Maher, "Designing creative AI partners with COFI: A framework for modeling interaction in human-AI co-creative systems," *ACM Trans. Comput.-Hum. Interact.*, vol. 30, no. 5, pp. 1–28, 2023.

[10] J. Kim, H. Lee, and Y. H. Cho, "Learning design to support student-AI collaboration: perspectives of leading teachers for AI in education," *Educ. Inf. Technol.*, vol. 27, no. 5, pp. 6069–6104, 2022.

[11] A. Smolansky *et al.*, "Educator and student perspectives on the impact of generative AI on assessments in higher education," in *Proc. Tenth ACM Conf. Learn. Scale*, 2023, pp. 378–382.

[12] S. P. Prajapati, R. Bhaumik, T. Kumar, and U. Sait, "An AI-based pedagogical tool for creating sketched representation of emotive product forms in the conceptual design stages," in *ICT Systems and Sustainability*(Lecture Notes in Networks and Systems), Singapore: Springer, 2020, pp. 649–659.

[13] W. Liao, K. Hu, M. Y. Yang, and B. Rosenhahn, "Text to image generation with semantic-spatial aware GAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 18 187–18 196.

[14] J. Oh, K. Min, and H. Yang, "A pose and style-invariant reenactment technique for artistic portraits using GAN," *IEEE Access*, vol. 10, pp. 50 351–50 362, 2022.

[15] H. Emami, M. M. Aliabadi, M. Dong, and R. B. Chinnam, "SPA-GAN: Spatial attention GAN for image-to-image translation," *IEEE Trans. Multimedia*, vol. 23, pp. 391–401, 2020.

[16] Y. Zhao, T. Celik, N. Liu, and H. C. Li, "A comparative analysis of GAN-based methods for SAR-to-optical image translation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[17] L. Li *et al.*, "Detail fusion GAN: High-quality translation for unpaired images with GAN-based data augmentation," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, 2021, pp. 1731–1736.

[18] H. Meng and F. Guo, "Image classification and generation based on GAN model," in *Proc. 3rd Int. Conf. Mach. Learn., Big Data Bus. Intell. (MLBDBI)*, 2021, pp. 180–183.

[19] A. Royer *et al.*, "XGAN: Unsupervised image-to-image translation for many-to-many mappings," in *Domain Adaptation for Visual Understanding*, Cham, Switzerland: Springer, 2020, pp. 33–49.

[20] E. R. Chan, M. Monteiro, P. Kellnhofer, J. Wu, and G. Wetzstein, "Pi-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 5799–5809.

[21] Y. Wang *et al.*, "3D multi-modality transformer-GAN for high-quality PET reconstruction," *Med. Image Anal.*, vol. 91, p. 102983, 2024.

[22] Q. Zhu and J. Luo, "Generative transformers for design concept generation," *J. Comput. Inf. Sci. Eng.*, vol. 23, no. 4, p. 041003, 2023.