Computational Visual Media
https://doi.org/10.1007/s41095-023-0356-2 Vol. 10, No. 2, April 2024, 355–373

Research Article

# Controllable multi-domain semantic artwork synthesis

**Yuantian Huang**[1] (✉), **Satoshi Iizuka**[1], **Edgar Simo-Serra**[2], **and Kazuhiro Fukui**[1]



**Abstract** We present a novel framework for the multi-domain synthesis of artworks from semantic layouts. One of the main limitations of this challenging task is the lack of publicly available segmentation datasets for art synthesis. To address this problem, we propose a dataset called *ArtSem* that contains 40,000 images of artwork from four different domains, with their corresponding semantic label maps. We first extracted semantic maps from landscape photography and used a conditional generative adversarial network (GAN)-based approach for generating high-quality artwork from semantic maps without requiring paired training data. Furthermore, we propose an artwork-synthesis model using domain-dependent variational encoders for high-quality multi-domain synthesis. Subsequently, the model was improved and complemented with a simple but effective normalization method based on jointly normalizing semantics and style, which we call spatially style-adaptive normalization (SSTAN). Compared to the previous methods, which only take semantic layout as the input, our model jointly learns style and semantic information representation, improving the generation quality of artistic images. These results indicate that our model learned to separate the domains in the latent space. Thus, we can perform fine-grained control of the synthesized artwork by identifying hyperplanes that separate the different domains. Moreover, by combining the proposed dataset and approach, we generated user-controllable artworks of higher quality than that of existing approaches, as corroborated by quantitative metrics and a user study.

1 Department of Computer Science, University of Tsukuba, Tsukuba 305-8577, Japan. E-mail: Y. Huang, huang_yuantian@cvlab.cs.tsukuba.ac.jp (✉); S. Iizuka, iizuka@cs.tsukuba.ac.jp; K. Fukui, kfukui@cs.tsukuba.ac.jp.

2 Department of Computer Science and Engineering, Waseda University, Tokyo 169-8050, Japan. E-mail: ess@waseda.jp.

Manuscript received: 2023-02-01; accepted: 2023-05-08

**Keywords** semantic artwork synthesis; generative adversarial network (GAN); datasets; non-photorealistic rendering

## 1 Introduction

Image synthesis consists of generating new images using existing data, whereas artwork synthesis focuses on generating images in the art domain. While there is no consensus on whether computer-synthesized images are art [1], it has applications such as being used as an art teaching tool, inspiring other artists, and providing different perspectives when understanding the artwork. Furthermore, conditional artwork synthesis has emerged as an important tool for artists to generate new art with a notable examples such as image stylization [2] which can transfer artistic styles to photography while preserving content.

Existing approaches to artwork generation have focused on unconditional generation [3, 4], image-to-image translation [5], or generation from sketches [6, 7]. However, these approaches yielded limited or no output control, making it difficult to determine whether, for example, a sketch is of a mountain or mountain-shaped rock. We proposed using semantic maps that provide high control over the generated image content. Some existing approaches can only partially achieve the goals [8, 9] and require additional user inputs to generate the images. For more general image generation, Park et al. [10] proposed spatially-adaptive denormalization (SPADE) for semantic image synthesis allowing users to control the semantics and styles when synthesizing a photorealistic image. However, it cannot exploit multi-domain data and require a paired training dataset.

This study proposes solving the controllability and data problems by introducing a new semantic map and artwork paired dataset, named *ArtSem*,

and a multi-domain high-quality artwork synthesis model. The dataset is created from landscape photographs by first computing semantic maps using an off-the-shelf semantic segmentation model with graph cut-based post-processing to create human-like maps. Subsequently, we train landscape photography with an artwork generation model using unpaired training data to create high-quality multi-domain paired training data. Finally, the refinement process ensured high-quality data suitable for training artwork synthesis models.

We propose a controllable multi-domain semantic artwork synthesis (CMSAS) model consisting of multiple domain-specific variational encoders that convert the artwork images into latent vectors and a generator that converts semantic label maps with an encoded latent vector in artwork images. Motivated by the observation that the images generated by SPADE are inclined to be less artistic, we based our generator on a novel spatially style-adaptive normalization (SSTAN) modules. While the SPADE modules are only aware of semantic information, SSTAN modules inject latent codes that represent styles into the normalization layers, and perform feature map modulation using semantics and style information. Accordingly, the learned modulation parameters depend on the input semantic layout and latent code, effectively propagating information throughout the network and significantly improving the results. Furthermore, because multiple encoders can learn domain-specific latent vectors, we propose to separate the domain latent vectors with a hyperplane for fine-grained control of the output artwork domain. Finally, the developed CMSAS model generates artwork from a semantic map for interactive applications.

We evaluated our approach quantitatively using automatic metrics and a perceptual user study in addition to the qualitative results. Results show that our approach outperformed the existing approaches in all metrics.

This study presents the following contributions:

- A single-model semantic artwork synthesis approach that generates high-quality artworks from easy-to-manipulate semantic layout inputs in multiple domains.
- A high-quality pixel-aligned semantic artwork dataset that contains artistic images with paired segmentation masks.
- An effective normalization method that significantly improves the artwork generation quality.
- Highly controllable generation with domain and style control via latent space manipulation.
- In-depth evaluation of our method based on qualitative and quantitative comparisons with existing approaches.

## 2 Related work

This section discusses related works and compares approaches similar to our proposed method. A high-level comparison of similar approaches is presented in Table 1. Specifically, two-step neural style transfer (NST) approaches combining semantic natural image synthesis [10] and style transfer [2] are computationally expensive for real-time editing. Moreover, their practical use is limited by the restricted set of pre-defined styles. Image-conditioning approaches such as Co-ModGAN [11] cannot generate artworks in a specific domain or use a reference style image. The same limitations are shared by OASIS [12] since it relies on random noise as the style input. SMIS [13] uses a single variational auto-encoder (VAE) to transform input images into latent vectors for reference-style images. However, these latent vectors are inseparable and cannot be constrained within a specific domain, as discussed in Section 4.3. SEAN [14] can generate artwork in a specific domain using a fixed pre-computed mean style code, but the approach lacks diversity and cannot interpolate between different domains. Our results verify that the proposed approach is the most flexible of the existing approaches.

### 2.1 Artwork synthesis

The concept of artwork synthesis goes back to image analogies [15], where filters related to a painting style are automatically learned from training data based on a simple multi-scale autoregression. It is a generative art [16, 17] that attempts to generate artworks algorithmically based on heuristics. Recently, neural style transfer [2] methods that solve the problem of applying artistic styles derived from reference images to photographs while preserving their content through optimization have become popular. The neural style transfer is further discussed in Section 2.2. With the advent of generative adversarial networks (GAN)

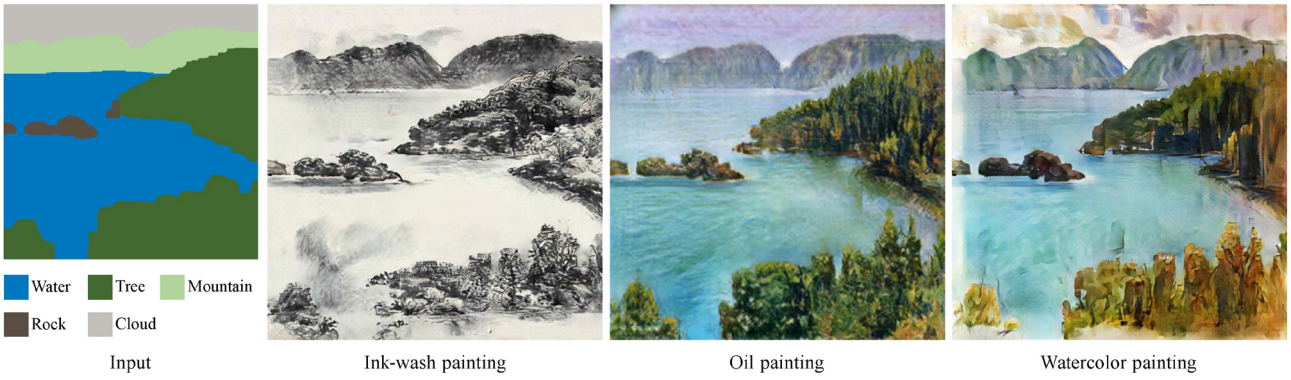



**Fig. 1** Controllable artwork synthesis. Our approach can synthesize artwork from different domains using user-provided semantic label maps. The user can freely change the semantic layout and target domain to intuitively and interactively create new artworks.

**Table 1** Comparison of interactive artwork synthesis approaches

| Method | Multiple domains in a single model | Generation in a specific domain | Cross-domain style morphing | Real-time interactive editing | Use reference |
|---|---|---|---|---|---|
| Two-step NST | | | | | ✓ |
| Co-ModGAN | ✓ | | | ✓ | |
| OASIS | ✓ | | | ✓ | |
| SMIS | ✓ | | | ✓ | ✓ |
| SEAN | ✓ | ✓ (fixed) | | ✓ | ✓ |
| CMSAS (ours) | ✓ | ✓ (diverse) | ✓ | ✓ | ✓ |

[18], data-driven generation approaches have become dominant. For GAN-based artwork generation, artGAN [3, 19] and CAN [4] extend and apply GANs to generate artworks. However, the low-resolution and limited quality of generated artworks restrict their applicability. Although some StyleGAN-based methods [20, 21] and text-conditioned diffusion models [22, 23] generated higher-quality artworks, the results are less controllable, and diffusion models incur significantly higher computational costs than GAN-based approaches, significantly limiting their application. Pix2pixGAN [6] has demonstrated its ability to generate high-quality artwork from hand-drawn sketches. However, these approaches are difficult to control because only the edge information is used. Specifically, a sketch cannot clearly indicate the differences between different objects having the same contour (e.g., mountains and rocks). Unlike previous studies, our novel semantics-based model allows users to control the shape, contour, and the semantic information of the input label map to create new artwork with higher quality and fidelity.

## 2.2 Neural style transfer

Gatys et al. [2] first used a convolutional neural network (CNN) [24] to extract style and content representations from the images and optimize the image content to match different art styles. Their approach creates new images by optimizing style and content loss, calculated by matching the Gram matrix statistics of pre-trained CNN features. Neural style transfer has many interesting applications in fields such as art, graphics, images, and video processing. There are various extensions and improvements to the original neural style transfer, focusing on different aspects, such as multiple style transfer [25], content-aware style transfer [26, 27], and other aspects [28–31]. This study adopted the concept of style loss to improve the generation quality of non-photorealistic style images in our dataset. The effectiveness of the method is verified by comparing it with state-of-the-art methods.

## 2.3 Image-to-image translation

Image-to-image translation aims to convert an image input into the desired image output and is dominated by CNN approaches. Supervised approaches [6, 32] employ paired training data and obtain impressive results. However, for the art domain, obtaining paired training data is challenging. Many approaches have started to focus on the unsupervised image-to-image translation [5, 33–42]. A notable example is CycleGAN [5], which uses cycle consistency with adversarial losses to overcome the lack of

paired images. Although limited in resolution, they demonstrated the possibility of producing realistic images for various datasets, including several Western painting styles. Based on CycleGAN, ChipGAN [43] introduced edge loss that enforces brushstroke constraints and successfully generates impressive results for ink-wash painting. However, these methods work only with photorealistic images as inputs and cannot handle abstract inputs such as semantic label maps. Therefore, we designed a two-stage image-transformation framework to produce a high-quality paired dataset of semantic labels and artwork images. Replacing photographs with semantic label maps as inputs gives users more control over the final result while preserving high performance using the existing techniques.

### 2.4 Semantic image synthesis

As a subfield of image-to-image translation, semantic image synthesis focuses on generating new images conditioned on semantic image maps, and are mainly dominated by GAN [18]. For example, pix2pixGAN [6] firstly established a common framework for mapping paired data, successfully generating realistic images and has been extended in further research [32, 45–50]. Recently, Park et al. [10] proposed SPADE, which outperformed the former approaches in terms of generating photorealistic images conditioned by semantic layout. Based on the SPADE architecture, SMIS [13] successfully produced semantically multimodal images by replacing all regular convolution layers in the generator with group convolutions. OASIS [12] surpassed SPADE in terms of diversity while maintaining similar quality by redesigning the discriminator. SEAN [14] used style input images to create spatially varying normalization parameters per semantic region. Other improvements [51] have also been made in different aspects. Another notable approach is the Co-Mod GAN [11] which achieves semantic image synthesis via the co-modulation of conditional and stochastic style representations based on unconditional generative architectures, such as StyleGAN2 [52].

Although these methods generate good results in photorealistic image generation, the results were less convincing when generating artwork. Furthermore, our approach focused on providing in-depth controllability, while most existing approaches have limited or zero controllability.

## 3 ArtSem dataset

Ideally, we would prefer using high-quality manual semantic segmentation annotations of diverse artworks for supervised semantic training of semantic artwork synthesis. However, given the nature of the art, it is challenging to create high-quality annotations because of the need for domain-specific knowledge and the inherent ambiguity of the artwork. Recent research [53] has attempted to address this problem; however, they have limitations in terms of class and precision. Instead of manual annotation, we proposed a method for the large-scale automatic generation of paired semantic maps and diverse artwork. Our approach is as follows: First, we use a semantic segmentation model to obtain segmentation maps from real-scene photographs. Next, we train an image-to-image translation model to convert semantic maps into different art styles. We base our approach on conditional unsupervised adversarial training, which can learn the mapping from semantic maps to artworks. Finally, we perform refinement based on the mimicked human judgment to generate a final dataset. An overview of the dataset construction approach is presented in Fig. 2. This study focuses on various landscape paintings, such as canyons and lake shores.

### 3.1 Data collection

First, we collected 50,000 landscape photographs of various outdoor scenes from Flickr. We then removed 15,000 samples based on generated label maps if they had irrelevant labels, such as people or animals. Of the remaining 35,000 images, 1000 were used as the validation set. We obtained images from four different domains: ink-wash, Monet oil, Van Gogh oil, and watercolor paintings. The images were obtained as follows:

- Ink-wash paintings: We collect them from search engines and manually remove images that are unrecognizable or contain too much irrelevant content. We also homogenize the style and tone by manually removing images that are different. This is repeated until 1000 paintings are obtained.
- Monet oil paintings: These are taken from Zhu et al. [5] and were originally downloaded from `wikiart.org`. There are a total of 1072 Monet oil paintings.
- Van Gogh oil paintings: These were taken from the same dataset as the Monet oil paintings. There is a total of 400 Van Gogh oil paintings.

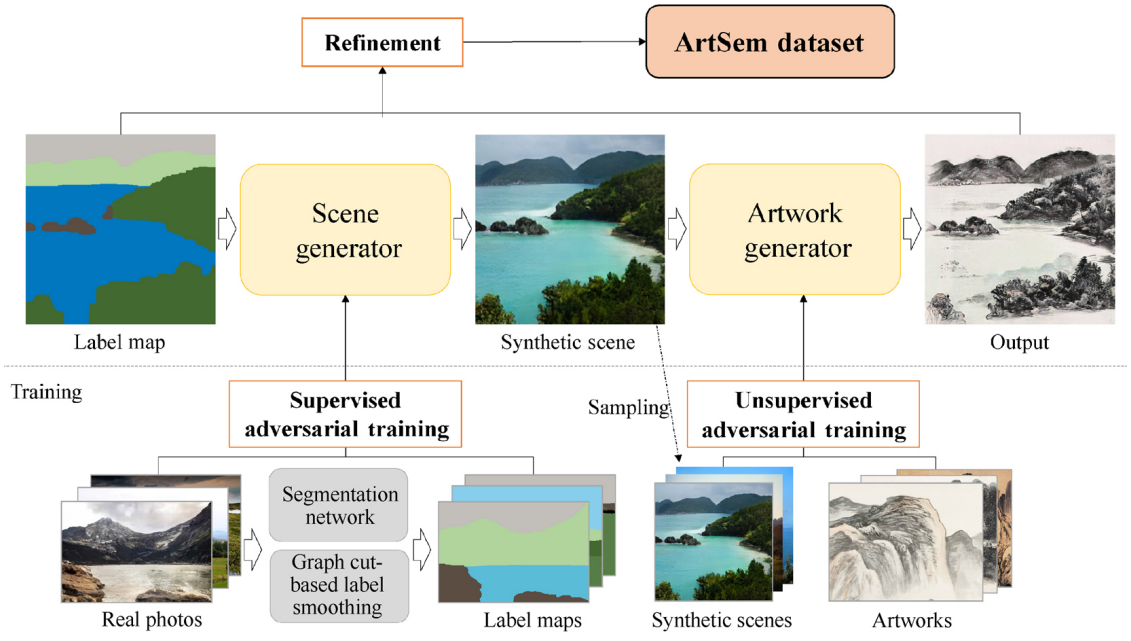



**Fig. 2** Our proposed dataset construction approach. We use a segmentation network [44] and graph cut-based smoothing to obtain semantic label maps for training the landscape scene generator. The synthesized scene images and real artworks are then used as a supervised training set for synthesizing art images. Finally, we use a refinement approach that mimics human judgment to obtain the final training dataset. After training with the dataset, our framework can generate high-quality artworks conditioned on input semantic label maps.

- Watercolor paintings: Similar to the ink-wash paintings, we collect 1000 manually curated images using search engines.

### 3.2 Semantic scene generation

In this section, we aim to learn to map from an input segmentation mask to a photorealistic image. We employ a state-of-the-art network architecture [10] for semantic image synthesis based on spatially-adaptive normalization. This model requires a dataset containing several landscape photos and uses paired semantic layouts for training. Although some datasets such as MSCOCO [54] and ADE20K [55] are widely used in semantic segmentation, most images in the dataset are not related to the landscape, which is the focus of this study. We generate training data from 35,000 landscape images using a pre-trained semantic segmentation network [44] to obtain paired training data. The model was trained on the MSCOCO dataset [54], which outputs 182 different label classes. We removed irrelevant labels and combined similar labels, such as moss and grass, to simplify the semantic map generation. Finally, we use 16 classes that commonly appear in the wild landscape. Please refer to the Electronic Supplementary Material (ESM) for more details.

The generated semantic maps exhibited irregularities and small inconsistencies that rendered them unsuitable for training high-quality generation models. We adopt a graph-cut-based label smoothing approach [56] to improve the quality of the results, given that we only require a low-frequency semantic map. An example of the smoothing effect is presented in Fig. 3.

Subsequently, we trained an SPADE model [10] to learn the mapping from semantic maps to

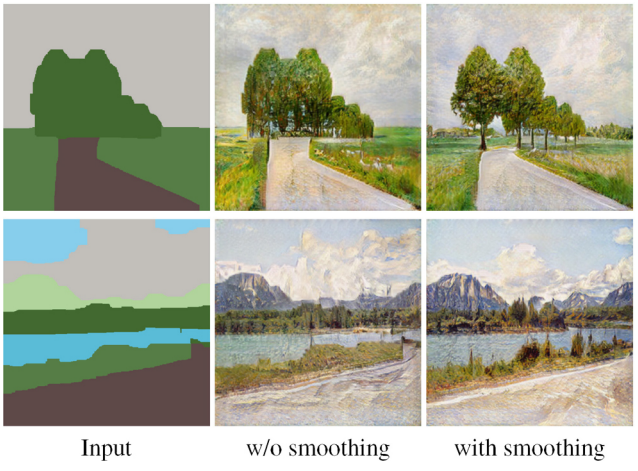



**Fig. 3** Effect of label smoothing. We compare the results from our model trained on the datasets with and without graph-cut-based label smoothing, which eliminates small irregularities allowing for more natural synthesized images. The effects are especially noticeable around the edges and outline, where the smoothing is the strongest.

photorealistic landscape images using the generated data. The model was trained with $256 \times 256$ pixel images, and bicubic up-sampling was used to increase the resolution to $512 \times 512$ pixels.

### 3.3 Weakly supervised artwork generation

Our objective at this stage is to learn a mapping function between the synthesized landscape images and real artwork without any explicit paired training data. We randomly selected 5000 synthesized images from the previous stage as landscape images and used the collected artwork images to train a generation model. The as-synthesized images, when used for inference, provided improved results than when using the synthesized images for training. Please refer to the ESM for a detailed discussion on this topic.

Our generation model is inspired by CycleGAN [5] and trains jointly a model that generates artworks from synthetic landscapes and a model that generates synthetic landscapes from the artwork. Using only adversarial and cycle-consistency losses led to poor results, given that the synthesized images have damaged parts and the generated images have high abstraction levels. To improve the results, we modified the training approach to include the style and edge loss terms. Separate models are trained for each modality. Figure 4 demonstrates that the proposed model successfully outperformed the base model with two auxiliary losses.

#### *3.3.1 Objective function*

We employ adversarial and cycle-consistency losses in our model, which simultaneously learns two mappings $G_{x\to y}: X \to Y$ and $G_{y\to x}: Y \to X$, where $X$ and $Y$ are the source and target domains, respectively. For a mapping function $G$ and its discriminator $D_Y$, the adversarial loss is given as

$$L_{\text{adv}} = \mathbb{E}_{y\sim X}\left[\log D_Y(y)\right] + \mathbb{E}_{x\sim Y}\left[\log\left(1 - D_Y(G(x))\right)\right] \tag{1}$$

Cycle-consistency loss is defined as Eq. (2):

$$\begin{aligned} L_{\text{cycle}} = {} & \mathbb{E}_{x\sim X}\left[\|G_{y\to x}\left(G_{x\to y}(x)\right) - x\|_1\right] \\ & + \mathbb{E}_{y\sim Y}\left[\|G_{x\to y}\left(G_{y\to x}(y)\right) - y\|_1\right] \end{aligned} \tag{2}$$

We introduced style loss [2] to improve the style similarity. In particular, we extracted style representations from subsets of VGG-19 [57] layers: *"conv1_1"*, *"conv2_1"*, *"conv3_1"*, *"conv4_1"*, and *"conv5_1"*. Given the computed features $F(I)^l$ from layer $l$ for a synthesized or real image $I$, we compute the Gram matrix using $M(I)^l_{ij} = \sum_k F(I)^l_{ik} F(I)^l_{jk}$. This allows style loss to be computed as the weighted sum of the difference of the computed Gram matrices as

$$L_{\text{style}} = \mathbb{E}_{x\sim X, y\sim Y} \sum_{l=0}^{L} \frac{\omega_l}{4N_l^2 M_l^2} \sum_{i,j} \left(M(G_{x\to y}(x))^l_{ij} - M(y)^l_{ij}\right)^2 \tag{3}$$

where $\omega_l$ is a weighting term for layer $l$, $N_l$ is the number of feature maps of layer $l$, and $M_l$ is the number of pixels in each feature map of layer $l$.

We further introduced edge loss [43] to emphasize the contours and lines in the generated images. Although the initial purpose of this loss was to imitate brush strokes of ink-wash paintings, it was also effective in synthesizing other artwork types. In particular, we used a holistic nested edge detector [58] $E$ to obtain edge maps of the synthesized landscape images $E(x)$ and edge maps of the synthesized artwork $E(G(x))$. The edge loss is computed as Eq. (4):

$$\begin{aligned} L_{\text{edge}} = \mathbb{E}_{x\sim X}\Big( & -\frac{1}{N}\sum_{i=1}^{N} \mu E(x)_i \log E(G(x))_i \\ & + (1-\mu)(1 - E(x)_i)\log\left(1 - E(G(x))_i\right)\Big) \end{aligned} \tag{4}$$

where $N$ is the total number of pixels in the edge map, and $\mu$ is a balancing weight. The sums of the probabilities for non-edges and edges of every pixel in $E(x)$ can be computed using $\mu = N_-/N$ and $1 - \mu = N_+/N$, $N_-$ and $N_+$, respectively.

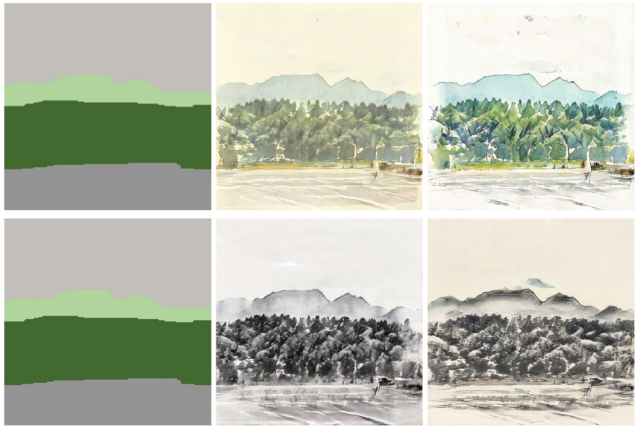



**Fig. 4** Effect of style and edge losses. The model can generate clear outlines and realistic artwork that matches the target style by employing style and edge losses.

The objective of the second stage is as Eq. (5):

$$L_{\text{total}} = L_{\text{adv}} + \alpha L_{\text{cycle}} + \beta L_{\text{style}} + \lambda L_{\text{edge}} \quad (5)$$

where $\alpha = 10$, $\beta = 0.1$, and $\gamma = 10$ control the relative importance of the individual objectives in all the experiments.

### 3.4 Training

We trained the networks from scratch in all experiments except for the pre-trained DeepLabV2 [44] model for label map generation and VGG-19 [57] model for style loss computation. The learning rate was 0.0002 in the first 100 epochs and linearly decayed to zero over the last 50 epochs for label map generation. For artwork generation, the learning rate was maintained at 0.0002 during the first 100 epochs, and then decayed to zero over the next 100 epochs. All the models were trained on two NVIDIA 1080Ti GPUs with 11 GB memory.

### 3.5 Dataset refinement

Following semantic scene generation and weakly supervised artwork generation, we obtain sets of semantic label maps, synthetic landscape images, and artwork images in the four domains. However, some synthesized artworks are less than ideal and can affect the models trained on such data, as shown in Fig. 5. To reduce data noise, we annotated 2000 images as either good or bad samples for training, and then trained a VGG classification model [57] to mimic human annotation. After training, we ran the model on the dataset and selected the top 10,000 images using the GMM-QIGA [59] quality score to obtain four different sets of paired synthetic landscapes and artworks, one for each domain. Some examples are presented in Fig. 6.

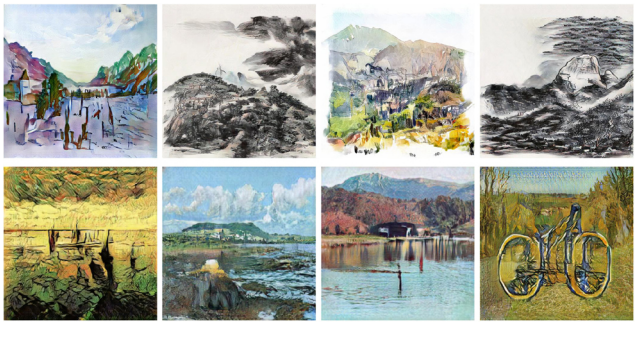


**Fig. 5** Examples of automatically removed poorly generated images. It is difficult to identify what is in the scene if confusing or partially drawn objects are present. Our dataset refinement approach removes such features when generating the training data.

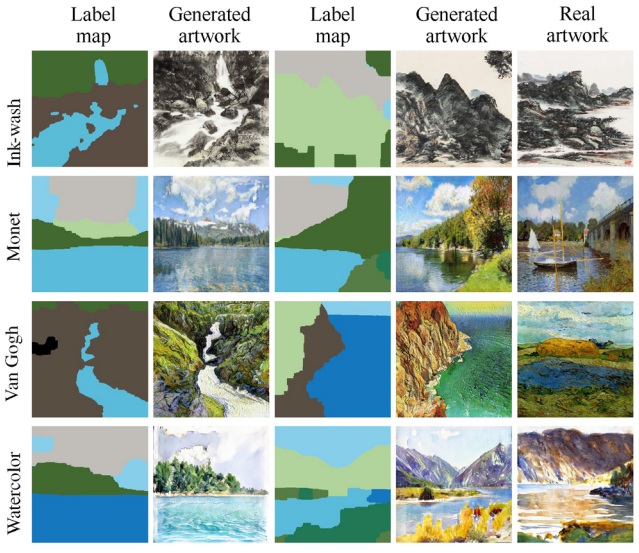



**Fig. 6** Samples from our *ArtSem* dataset and real artworks. There is diversity in the supervised label map and artwork images in the four different artistic styles.

## 4 Proposed framework

We propose CMSAS model comprising two sub-components: domain-specific variation encoders that encodes artwork images into latent vectors and a generator that generates new artwork based on an input semantic map and encoded latent vector. The generator was modified to use attention modules in each residual block with spatially style-adaptive normalization for better generation quality when synthesizing artistic images. An overview of this model is provided in Fig. 7. Furthermore, the proposed model could precisely manipulate the synthesized artwork by exploiting the latent space structure.

### 4.1 Model architecture

We divided the encoder into domain-specific and shared components. The domain-specific component consists of all the convolutional layers, while the last two fully connected layers output Gaussian distribution, characterized by the mean and standard deviation. In addition, the encoder uses the domain type as input to activate the domain-specific components. We use a shared discriminator because our experiments show no notable improvements compared to using one per domain. We also use a shared generator because it is computationally heavy compared with encoders. Inspired by Ref. [60], we introduce a dual attention module in each residual block to improve the generator performance. Each dual attention module consists of channel and spatial

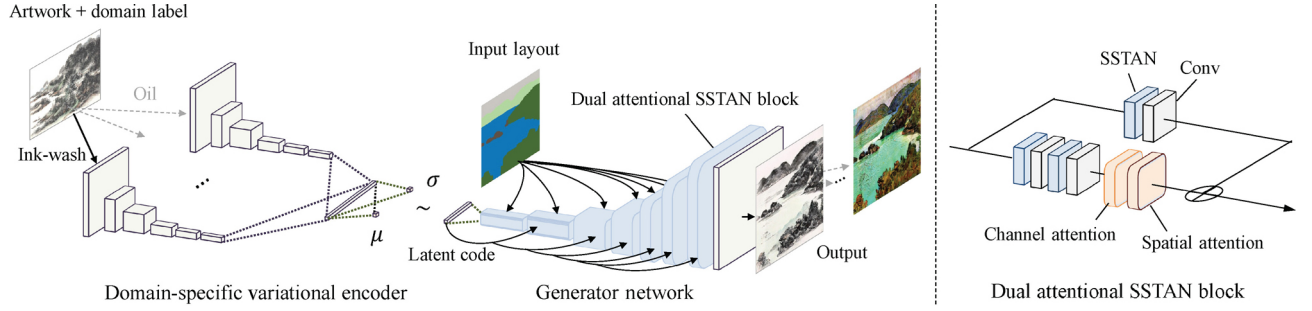



**Fig. 7** Proposed architecture. Our model consists of a domain-specific variational encoder and a generator network based on SSTAN. The encoder extracts the style from an input image as a latent code and passes it to the generator to synthesize new artworks with similar style appearance while conditioned by an input label map. The attention module and SSTAN in the generator improve the quality of the generated images.

attention, computed on the channel and spatial axes, respectively. The model has seven residual blocks, and the full layout is given in Fig. 7(right). For an in-depth overview of the full model, please refer to the ESM.

### 4.2 Spatially style-adaptive normalization

Based on the observation that the generated images from models using the SPADE module are less artistic, we strengthen the influence of the style information by passing it to the normalization layers of each residual block instead of just once at the generator input. Therefore, we propose a novel conditional normalization module called SSTAN to jointly learn semantic and style representations for higher-quality artwork generation. The modulation parameters of SSTAN are tensors with spatial dimensions, making the feature map modulations spatially adaptive. The latent code was expanded to the same size as the input semantic layout. Figure 8 illustrates the SSTAN module structure. The semantic layout and latent code that represent the style of the input artwork were processed by different convolutional neural networks to learn modulation parameters.

The input of each SSTAN block is the latent code $l$ and segmentation mask $M$. Let $h$ denote the network input activation for a batch of $N$ samples; let $C$, $H$, and $W$ be the number of channels, height, and width of the activation map, respectively. The modulated activation value at location $(n \in [1, N], c \in [1, C], y \in [1, H], x \in [1, W])$ becomes

$$\begin{aligned}&\mathrm{SSTAN}(n, c, y, x)\\ &= \gamma_{c,y,x}(l, M)\frac{h_{n,c,y,x} - \mu_c}{\sigma_c} + \beta_{c,y,x}(l, M) \qquad (6)\end{aligned}$$

where $h_{n,c,y,x}$ is the activation of the previous layer before normalization; $\mu_c$ and $\sigma_c$ denote the mean and standard deviation of the activation in channel $c$, respectively.

$$\mu_c = \frac{1}{NHW}\sum_{n,y,x} h_{n,c,y,x} \qquad (7)$$

$$\sigma_c = \sqrt{\frac{1}{NHW}\sum_{n,y,x}((h_{n,c,y,x})^2 - (\mu_c)^2)} \qquad (8)$$

The weighted sums of $\mu_c$ and $\sigma_c$ are used to modulate the activations of the generator. Two learnable parameters, $\alpha_\gamma$ and $\alpha_\beta$, which leverage the weight of each element, are trained directly to minimize the loss with backpropagation. The final modulation parameters $\gamma_{c,y,x}$ and $\beta c, y, x$ are defined as Eqs. (9) and (10):

$$\gamma_{c,y,x}(l, M) = \alpha_\gamma \gamma_{c,y,x}(l) + (1 - \alpha_\gamma)\gamma'_{c,y,x}(M) \qquad (9)$$

$$\beta_{c,y,x}(l, M) = \alpha_\beta \beta_{c,y,x}(l) + (1 - \alpha_\beta)\beta'_{c,y,x}(M) \qquad (10)$$

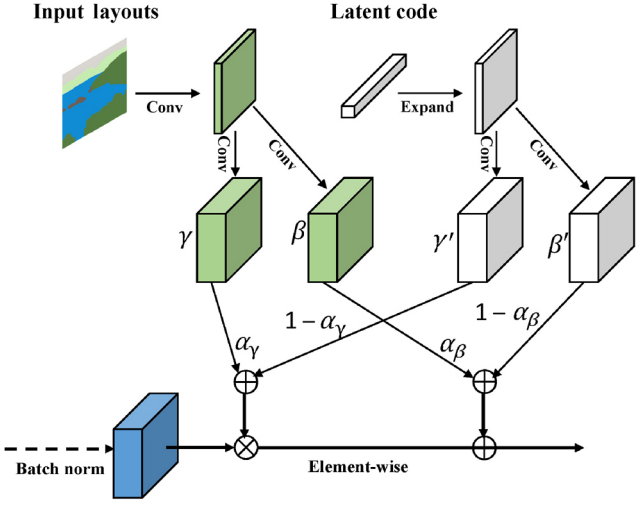



**Fig. 8** Proposed SSTAN module. The semantic layout is projected onto an embedding space and then convolved to produce the modulation values $\gamma$ and $\beta$. On the right side, the extracted input latent code is first extended to a two-dimensional space and then convolved to produce the modulation values $\gamma'$ and $\beta'$. These are used to modify the features of the neural network.

### 4.3 Domain and style control

To provide more control over the output artwork images, we assumed that the latent vectors of the different domains are separable with a hyperplanes [61]. We used a uniform manifold approximation and projection (UMAP [62]), a dimensionality reduction tool for projecting latent vectors in low-dimensional space for intuitive visualization. As shown in Fig. 9, unlike a single encoder, which learns entangled representations that are less separable in their latent space, our domain-specific encoders learn disentangled spaces that can be better separated into domains. Consequently, this allows our proposed model to achieve latent space manipulation to specify the domain of the generated artwork and perform cross-domain style morphing between different domains. Table 2 shows that the FID score of the generated artwork decreased only slightly as the number of domains increased.

For each latent code $z \in \mathcal{Z}$, our generator can be seen as a mapping $f_{\mathrm{g}}\colon \mathcal{Z} \to \mathcal{A}$, where $\mathcal{A}$ is the

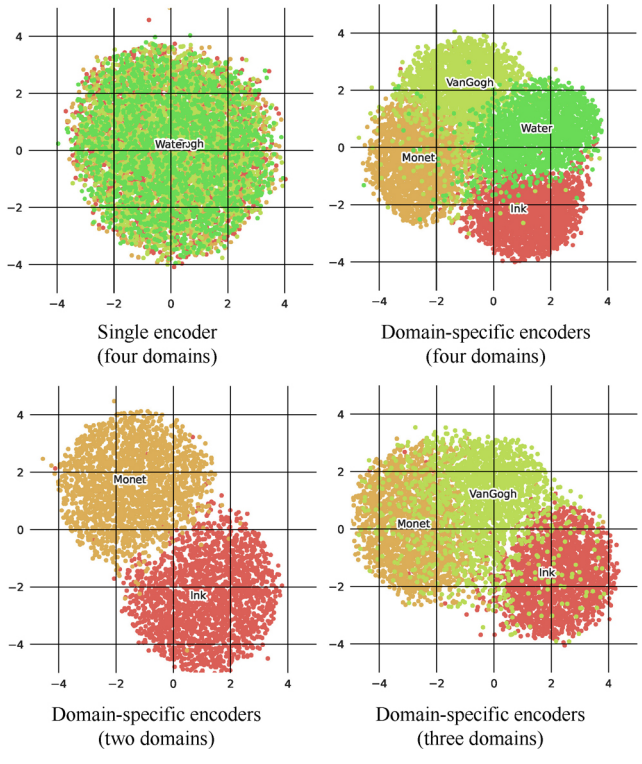



**Fig. 9** Visualization of the latent codes. By projecting the latent codes to a two-dimensional space with UMAP, we understand how the latent space is divided into different artwork domains using our domain-specific variational encoders. In contrast, the single encoder fails to separate the different domains. Comparisons of visualizations trained with different numbers of domains show that the separability remains high in all three patterns.

**Table 2** Quantitative effect of the number of domains. Although performance slightly decreases with the number of domains, joint training allows for interpolation and more flexible interactivity

| Domain | Per | Two | Three | Four |
|---|---|---|---|---|
| Ink-wash | **45.95** | 46.14 | 46.17 | 49.33 |
| Monet | **63.91** | 64.18 | 65.20 | 65.97 |
| Van Gogh | 123.90 | — | 125.87 | **115.30** |
| Watercolor | **68.04** | — | — | 68.96 |

manifold of artworks. For each domain, we defined a scoring function $f_{\mathrm{s}}\colon \mathcal{A} \to [0,1]$ that predicts whether an artwork corresponds to a particular domain. We then formulate the problem of finding a normal $n$ of a hyperplane that separates the latent codes based on their codes as Eq. (11):

$$\underset{n}{\operatorname{argmax}} \max\left(0, t\left(n^{\mathrm{T}} z\right)\right) \quad \text{subject to} \quad \|n\| = 1 \tag{11}$$

where

$$t = \begin{cases} 1, & f_{\mathrm{s}}\left(f_{\mathrm{g}}(z)\right) > 0.5 \\ -1, & \text{otherwise} \end{cases} \tag{12}$$

We can then control the output domain of the artwork by moving towards a particular domain. This can be achieved by adjusting a value $\alpha$ along the normal direction to obtain a new latent vector: $z' = z + \lambda n$, as shown in Fig. 10. This approach enables us to generate the artwork without the explicit need for an input style image by setting $z = 0$ and a larger value of $\lambda$.

In particular, we implemented $f_{\mathrm{s}}$ using a VGG classifier [57] trained to predict a domain from 40,000 synthesized images in the proposed dataset.

### 4.4 Objective function

Our loss functions include adversarial, feature matching [32], perceptual [63], and KL divergence losses.

**Adversarial loss**. Let $E$ be the domain-specific variation encoders, $G$ be the dual attentional SSTAN generator, and $D_k$ be the $k$-th discriminator at the different scales. Given an input artwork $A$, a latent code $l$, that represents the style of $A$ and are extracted by $E$, can be defined as $l = E(A)$. Let $M$ be the corresponding segmentation mask for $A$. The adversarial loss is then defined as

$$\begin{aligned} L_{\mathrm{adv}} = {} & \mathbb{E}[\max(0, 1 - D_k(A, M))] \\ & + \mathbb{E}[\max(0, 1 - D_k(G(l, M), M))] \end{aligned} \tag{13}$$

where $G$ attempts to minimize the objective against an adversarial $D_k$ that attempts to maximize it during the training.

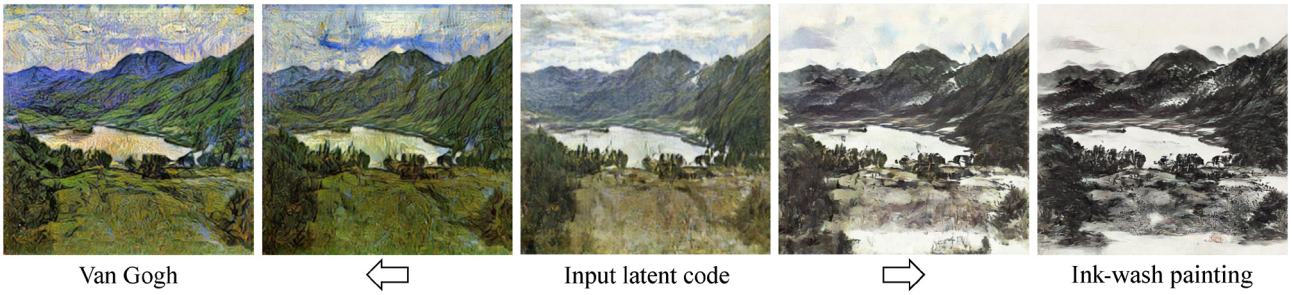



**Fig. 10** Latent space manipulation. For the same semantic map, we interpolate between latent codes of different artwork styles to obtain diverse results.

**Feature matching loss**. Let $T = 2$ be the total number of layers in discriminator $D_k$, and $D_k^{(i)}$ and $N_i$ be the output feature maps and the number of elements in the $i$-th layer of $D_k$, respectively. We formulate the feature matching loss $L_{\text{FM}}$ as

$$L_{\text{FM}} = \mathbb{E}\sum_{i=1}^{T} -\frac{1}{N_i}[\|D_k^{(i)}(A, M) - D_k^{(i)}(G(l, M), M)\|_1] \quad (14)$$

where $G$ attempts to minimize this objective against an adversarial $D_k$ that attempts to maximize it during the training.

**Perceptual loss**. Let $N$ be the total number of layers used to calculate the perceptual loss, $F^{(i)}$ be the output feature maps of the $i$-th layer of the VGG [57] network, and $M_i$ be the number of elements of $F^{(i)}$. We define the perceptual loss $L_{\text{P}}$ as

$$L_{\text{P}} = \mathbb{E}\sum_{i=1}^{N} -\frac{1}{M_i}[\|F^{(i)}(A) - F^{(i)}(G(l, M))\|_1] \quad (15)$$

**KL divergence loss**. Let $p(z)$ be a standard Gaussian distribution. The variational distribution $q$ is fully determined by a mean vector $\mu$ and a variance vector$\sigma$, which are the output of our encoder as shown in Fig. 7. We use the reparameterization trick [64] for backpropagating the gradient from the generator to the encoder.

$$L_{\text{KLD}} = D_{\text{KL}}(q(z|x)) \,\|\, p(z)) \quad (16)$$

**Full objective**. Our full objective is as Eq. (17):

$$L_{\text{full}} = L_{\text{adv}} + \lambda_1 L_{\text{FM}} + \lambda_2 L_{\text{P}} + \lambda_3 L_{\text{KLD}} \quad (17)$$

where $\lambda_1 = 10$, $\lambda_2 = 10$, and $\lambda_3 = 0.01$ control the relative importance of individual objectives in all the experiments.

## 5 Results

We evaluated our approach and performed qualitative and quantitative comparisons with other methods.

### 5.1 Baselines

We chose three existing models as baselines for multimodal semantic image synthesis: SMIS [13], OASIS [12], and Co-ModGAN [11] models. For generation within specific domains, we use the SEAN [14] model as our baseline model. All baseline models have shown an impressive ability to synthesize photorealistic images in different styles. For a fair comparison, all models are trained using our proposed dataset for semantic artwork synthesis. Furthermore, all baseline models were trained using the implementations provided by the authors. We chose three style transfer methods as our baseline models for the reference-based generation comparison, which are neural style transfer [2], AdaIN [31], and STROTSS [29].

### 5.2 Implementation details

We trained all the networks from scratch using our dataset. The image size was set to $512 \times 512$ pixels in our proposed model and the Co-Mod GAN [11], and $256 \times 256$ in other baseline models owing to GPU memory limitations. For further details, please refer to the ESM. A learning rate of 0.0002 was used for the first 40 epochs before decaying linearly to zero over the following 20 epochs. In the inference phase, the SEAN model required the style code as input, which we pre-computed from the mean style codes in each domain for all the training data. For our model, we set the input latent code $z$ to a zero vector and $\alpha = 3$, leading to a positive style appearance for a particular domain in the quantitative comparison. This setting results in the stable generation of synthesized artwork in each domain; therefore, the proposed framework requires only label maps as input.

### 5.3 Quantitative comparison

We used the Fréchet inception distance (FID) [65] as our primary evaluation metric to capture the

perceptual similarity of generated images with real ones. After resizing all the images to the same size of $512 \times 512$ pixels, we calculated FID between the real and generated artwork for each domain. As listed in Table 3, our approach significantly outperforms existing approaches. When input reference images are available, we compare our proposed method with SEAN model, and three style transfer methods, including NST [2], AdaIN [31], STROTSS [29], as shown in Table 4. Specifically, we used a simple two-stage method to enable style transfer models for semantic image synthesis by adding a pre-trained SPADE model [10] to generate synthesized images from semantic layouts. We randomly selected 200 style images and 100 synthesized images as content images for each domain and then calculated FID based on the 20,000 generated images in each domain.

### 5.4 Perceptual user study

We evaluated our method through a perceptual user study with 15 participants. We randomly select 5000 generated images per domain for each approach, and all real artworks from our dataset. In each round, two randomly selected images from different approaches or real images were shown to the user. We asked the participants to choose which image seemed better in terms of realism and style for 500 rounds per user.

**Table 3** Quantitative evaluation. We compare the proposed method with existing approaches for conditional image generation using the FID metric. The baseline models SMIS, OASIS, and Co-Mod can not specify the domain of generated artworks; therefore we only evaluate their results in a mix of 4 domains. The best results are highlighted in bold

| Domain | SMIS | OASIS | Co-Mod GAN | SEAN | CMSAS (ours) |
|---|---|---|---|---|---|
| Ink-wash | — | — | — | 86.47 | **49.33** |
| Monet | — | — | — | 84.47 | **65.97** |
| Van Gogh | — | — | — | 145.74 | **115.30** |
| Watercolor | — | — | — | 106.37 | **68.96** |
| Mixed | 97.38 | 78.04 | 67.38 | 89.52 | **55.03** |

**Table 4** Quantitative evaluation with style transfer algorithms. We compare the proposed method with existing approaches for neural style transfer using the FID metric. The best results are highlighted in bold

| Domain | NST | AdaIN | STROTSS | SEAN | CMSAS (ours) |
|---|---|---|---|---|---|
| Ink-wash | 76.15 | 96.59 | 75.26 | 62.33 | **45.35** |
| Monet | 94.02 | 93.25 | 76.23 | 75.74 | **67.40** |
| Van Gogh | 155.17 | 161.53 | 128.29 | 123.36 | **116.20** |
| Watercolor | 109.13 | 124.53 | 75.25 | 73.66 | **66.00** |

As shown in Table 5, our approach is preferred over existing approaches. Furthermore, compared to real artwork, our approach was considered better than 21.64% of the time. This matches the quantitative comparison of results.

To evaluate the layout preservation, we use the same 20,000 selected images from each approach along with all the paired semantic layouts. We asked the same 15 participants to choose an absolute scale from 0 to 3 of how well they think the generated artwork matches the label map. In each round, a randomly selected image and its paired input layout were shown. The results in Table 6 indicate that our approach is preferred over the baseline models in terms of layout preservation.

### 5.5 Visual quality comparison

Given that some of the baseline models cannot generate artwork in a specific domain, we divide the visual quality comparison into two parts: joint multimodal generation in all domains and generation in a specific domain.

For the multimodal generation scenario, we provide a qualitative comparison of results in Fig. 11 against the baseline approaches. SMIS generates poorer results in terms of style, whereas OASIS and Co-mod GAN models generate higher-quality artwork albeit with few details. The reflections on the water, texture, and style details on the mountains, as well as the clouds of our proposed method, exhibit higher quality compared to baseline models. We hypothesize that the performance increase comes from SSTAN enforcing the influence of style information, leading to better style-aware generation.

**Table 5** Perceptual user study results. The numbers indicate the percentage of users that prefer our method with respect to existing approaches

| | SMIS | OASIS | Co-Mod | SEAN | Real |
|---|---|---|---|---|---|
| CMSAS (ours) | 79.61% | 67.71% | 69.15% | 65.26% | 21.64% |

**Table 6** Perceptual user study on layout preservation. The numbers indicate the average scores of how well the participants think the generated artworks of each method preserve the semantic layout of the input. The scores are on a scale of 0–3, with higher scores corresponding to better layout preservation. The best results are highlighted in bold

| | SMIS | OASIS | Co-ModGAN | SEAN | CMSAS (ours) |
|---|---|---|---|---|---|
| Score | 2.32 | 2.61 | 2.51 | 2.55 | **2.68** |


**Fig. 11** Qualitative comparison of multimodal generations. Our proposed model generates a higher quality of texture details and style intensity compared to the baseline models. The SIMS, OASIS, and Co-ModGAN models cannot generate artwork in a pre-specified domain due to using a random input vector. The domain-specific encoders used in our proposed method provide improved controllability via latent space manipulation, as discussed in Section 4.3.

We qualitatively compare domain-specific generation against the SEAN model in Fig. 12. Compared to our approach, the results using SEAN tend to be fuzzier and have fewer details. Furthermore, the proposed CMSAS model allows for more control with the latent codes of the style instead of a fixed pre-computed style code used in the SEAN model.

We also provided a qualitative comparison when reference images were available. In CMSAS model, domain-specific encoders were used to convert the reference images into latent codes, which are then passed to a generator that synthesizes artwork with a similar appearance. To compare with existing style transfer methods without semantic image synthesis, we added a pre-trained SPADE model to generate synthesized images from the semantic layouts. The synthesized images are then passed on to style transfer models to apply the reference style. As shown in Fig. 13, the proposed method significantly

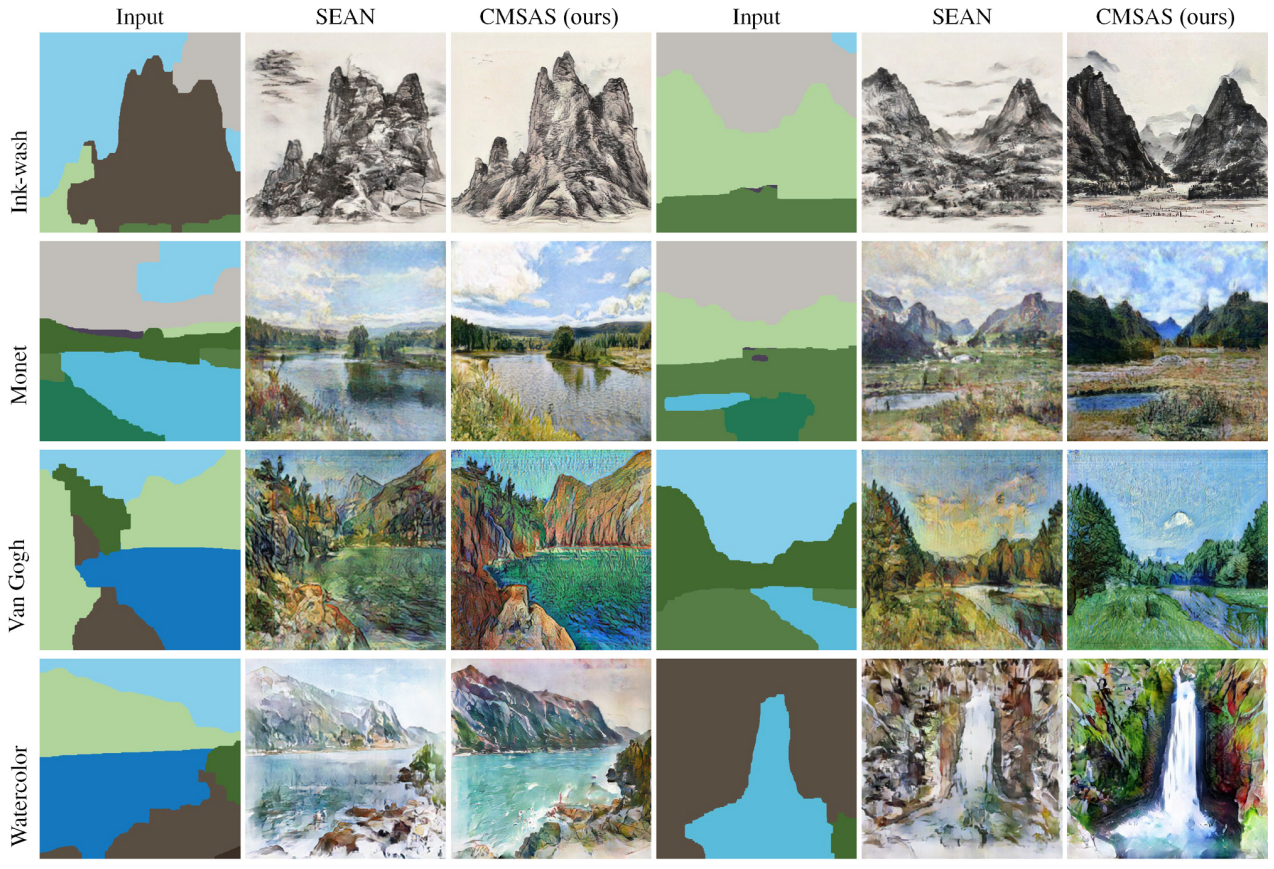



**Fig. 12** Qualitative comparison for domain-specific generations. We compare our approach against existing approaches for artwork generation in each of the specific domains.

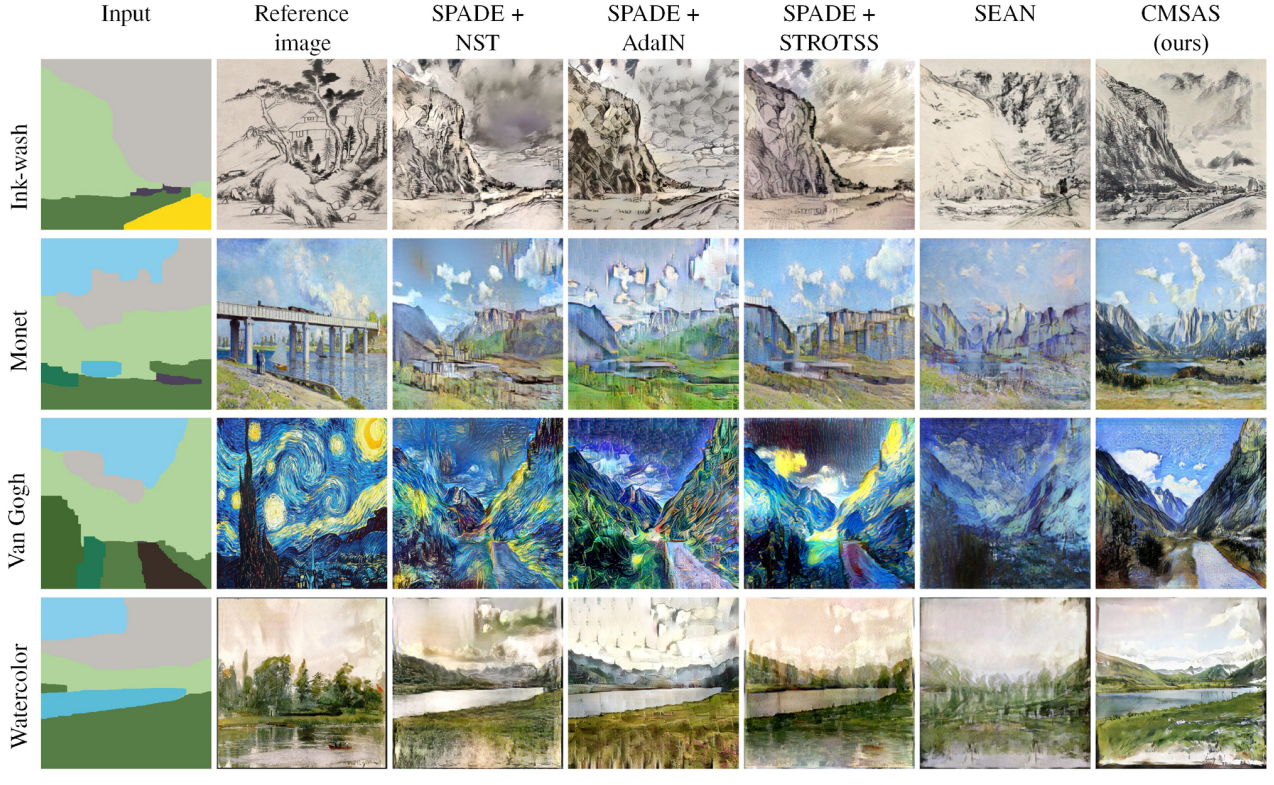



**Fig. 13** Qualitative comparison of reference-based artwork generation with neural style transfer models. All reference images are licensed under the public domain.

outperformed the baseline models. However, when the reference image is extremely stylistic, as in the Starry Night example in Fig. 13, unlike baseline methods that try to extract style representations from the reference image, the proposed method mimics a style similar to that of the learned latent space, which may result in less stylism owing to the limitations of the training dataset. The additional reference-based generation results of the proposed model are shown in Fig. 14. Given the same fixed semantic layouts as input, our proposed method can apply a similar style based on the reference image.

### 5.6 Ablation study

An ablation study was conducted to verify the effectiveness of each element. The results presented in Table 7 and Fig. 15 show that performance is significantly improved by adding each element to the complete CMSAS model. In addition, we provide a comparison with the traditional class-conditional

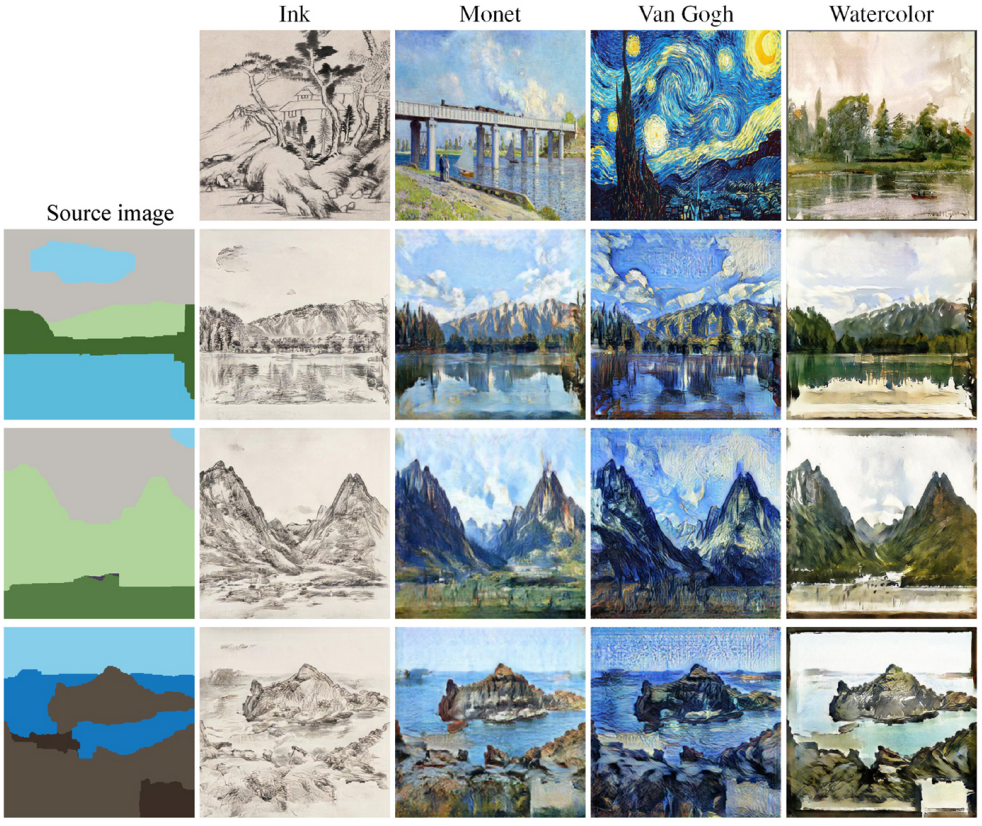



**Fig. 14** Reference-based generation of our proposed model in multiple domains. Our proposed method can apply styles similar to a reference image while maintaining coherency with the input semantic information.

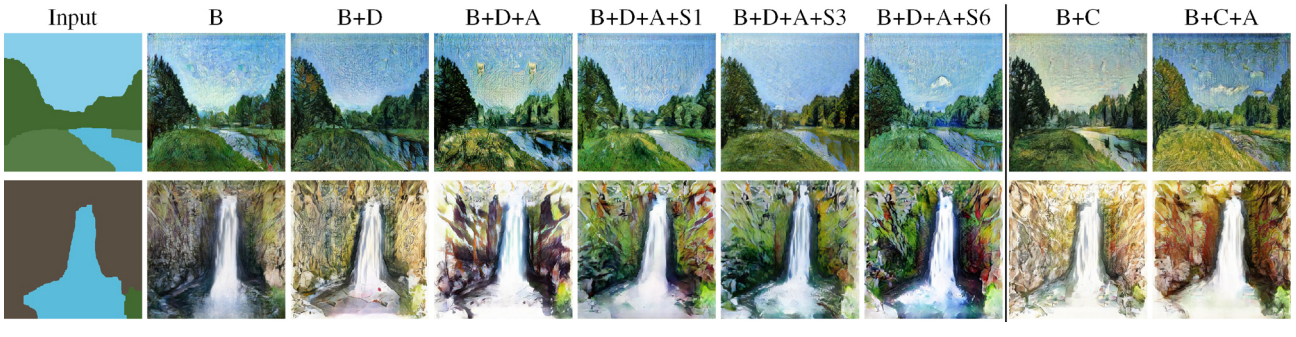



**Fig. 15** Examples from ablation study. We show sets of results for different configurations given two different semantic label maps and latent codes. We show results for the base model (B) with different components, including attention modules (A), separate domain encoders (D), and SSTAN for 1 layer (S1), 3 layer (S3), and 6 layer (S6) configurations, alongside a class-conditional approach (C). The top and bottom rows show the Van Gogh and Watercolor domains, respectively.

**Table 7** Ablation study. We evaluate the effect of different components on the FID scores. Our base model (B) is equivalent to an SPADE model trained with a single image encoder. In particular, we look at the effects of the attention module (A), separate domain encoders (D), and SSTAN for 1 layer (S1), 3 layer (S3), and 6 layer (S6) configurations. We also compare with a class-conditional (C) method in the last two columns, in which we concatenate the latent code with a one-hot class label that indicates a different domain. Thus, we generate artwork in different domains via one-hot class labels in the last two methods, whereas other approaches use latent space manipulation 4.3 to separate domains for more controllability, such as cross-domain style morphing or reference-based generation. The method (B + D + A + S6) corresponds to our full CMSAS model. The best results are highlighted in bold

| Method | Ink-wash | Monet | Van Gogh | Watercolor |
|---|---|---|---|---|
| B | 78.83 | 74.31 | 140.23 | 95.43 |
| B + D | 77.63 | 72.14 | 134.96 | 89.85 |
| B + D + A | 61.25 | 69.34 | 123.21 | 76.15 |
| B + D + A + S1 | 56.36 | 68.51 | 121.13 | 73.01 |
| B + D + A + S3 | 51.17 | 67.32 | 119.05 | 72.24 |
| B + D + A + S6 | **49.33** | **65.97** | **115.3** | **68.96** |
| B + C | 71.74 | 72.97 | 134.14 | 84.45 |
| B + C + A | 60.97 | 68.71 | 124.11 | 75.52 |

approach that concatenates latent code with a one-hot class labels indicating different domains. The results indicate that the proposed method outperformed class-conditional approaches. In addition, our proposed model can perform across-domain style morphing and reference-based generation, leading to higher controllability.

### 5.7 Limitations and discussion

Although our framework can generate high-quality artwork using semantic label maps, it has several limitations owing to its data-driven approach. In particular, its application is limited to known labels and new data must be acquired to extend the model to new labels, such as animals. Furthermore, the model learns a mapping from realistic semantic maps to artwork images and can fail if the input semantic maps diverge significantly from the training data, as shown in Fig. 16. Moreover, the style or tone of the generated artwork was restricted to those similar to real artwork used for the training dataset. In certain scenes, the input semantic can also produce unexpected results owing to dataset bias. Examples are presented in Fig. 17.

## 6 Conclusions

We present a novel paired dataset of semantic maps, landscape images, and artwork from different domains, along with a high-performance method for generating artwork from semantic maps. Our approach uses multi-domain artworks, exploiting the


**Fig. 16** Generation conditioned on unnatural inputs. When the inputs are unnatural, such as clouds under mountains or 90° rotations, the model may struggle to generate results as the input differs significantly from the training images.

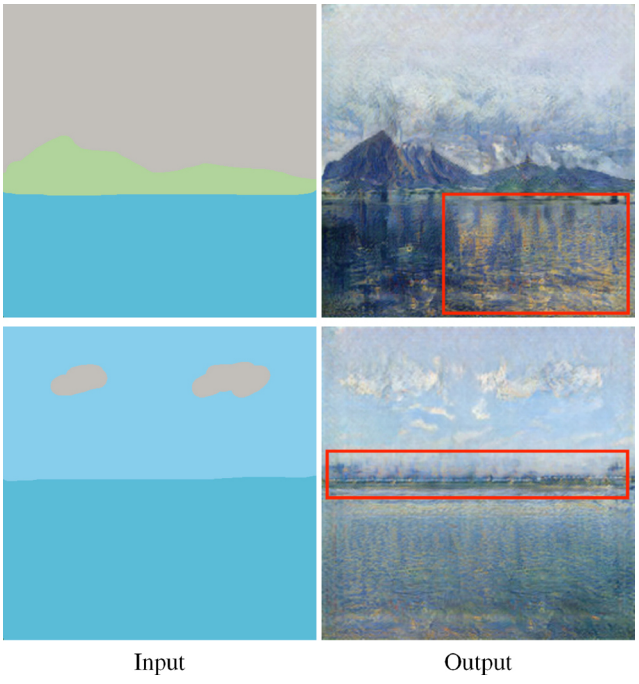



**Fig. 17** Dataset bias. When the input semantic maps do not share characteristics with the training data, the model may fail to accurately convey the intended scene. In the top row, although the user specified a fully cloudy sky, there is a reflection of the sun in the water. In the bottom row, despite the absence of land in the input semantic map, the model attempts to generate a thin piece of land on the horizon.

structure of the latent space to precisely manipulate the resulting artwork. Furthermore, because our approach is generated from semantic maps, it can be used in interactive scenarios. We plan to release our dataset and hope that it will encourage other researchers to further investigate artwork generation tasks.

## Acknowledgements

This study was supported by the Japan Science and Technology Agency Support for Pioneering Research Initiated by the Next Generation (JST SPRING) under Grant No. JPMJSP2124.

## Declaration of competing interest

The authors have no competing interests to declare that are relavent to the content of this article.

## Electronic Supplementary Material

Supplementary material is available in the online version of this article at `https://doi.org/10.1007/s41095-023-0356-2`.

## References

[1] Hertzmann, A. Can computers create art? *Arts* Vol. 7, No. 2, 18, 2018.

[2] Gatys, L. A.; Ecker, A. S.; Bethge, M. Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2414–2423, 2016.

[3] Tan, W. R.; Chan, C. S.; Aguirre, H. E.; Tanaka, K. ArtGAN: Artwork synthesis with conditional categorical GANs. In: Proceedings of the IEEE International Conference on Image Processing, 3760–3764, 2017.

[4] Elgammal, A.; Liu, B.; Elhoseiny, M.; Mazzone, M. CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. *arXiv preprint* arXiv:1706.07068, 2017.

[5] Zhu, J. Y.; Park, T.; Isola, P.; Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, 2242–2251, 2017.

[6] Isola, P.; Zhu, J. Y.; Zhou, T. H.; Efros, A. A. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5967–5976, 2017.

[7] Liu, B. C.; Song, K. P.; Zhu, Y. Z.; Elgammal, A. Sketch-to-art: Synthesizing stylized art images from sketches. In: *Computer Vision – ACCV 2020. Lecture Notes in Computer Science, Vol. 12627*. Ishikawa, H.; Liu, C. L.; Pajdla, T.; Shi, J. Eds. Springer Cham, 207–222, 2021.

[8] Men, Y. F.; Lian, Z. H.; Tang, Y. M.; Xiao, J. G. A common framework for interactive texture transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6353–6362, 2018.

[9] Champandard, A. J. Semantic style transfer and turning two-bit doodles into fine artworks. *arXiv preprint* arXiv:1603.01768, 2016.

[10] Park, T.; Liu, M. Y.; Wang, T. C.; Zhu, J. Y. Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2332–2341, 2019.

[11] Zhao, S. Y.; Cui, J.; Sheng, Y. L.; Dong, Y.; Liang, X.; Chang, E. I.; Xu, Y. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint* arXiv:2103.10428, 2021.

[12] Sushko, V.; Schönfeld, E.; Zhang, D.; Gall, J.; Schiele, B.; Khoreva, A. You only need adversarial supervision for semantic image synthesis. *arXiv preprint* arXiv:2012.04781, 2020.

[13] Zhu, Z.; Xu, Z. L.; You, A. S.; Bai, X. Semantically multi-modal image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5466–5475, 2020.

[14] Zhu, P. H.; Abdal, R.; Qin, Y. P.; Wonka, P. SEAN: Image synthesis with semantic region-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5103–5112, 2020.

[15] Hertzmann, A.; Jacobs, C. E.; Oliver, N.; Curless, B.; Salesin, D. H. Image analogies. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, 327–340, 2001.

[16] Dehlinger, H. On fine art and generative line drawings. *Journal of Mathematics and the Arts* Vol. 1, No. 2, 97–111, 2007.

[17] Phon-Amnuaisuk, S.; Panjapornpon, J. Controlling generative processes of generative art somnuk phon-. *Procedia Computer Science* Vol. 13, 43–52, 2012.

[18] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Communications of the ACM* Vol. 63, No. 11, 139–144, 2020.

[19] Tan, W. R.; Chan, C. S.; Aguirre, H. E.; Tanaka, K. Improved ArtGAN for conditional synthesis of natural image and artwork. *IEEE Transactions on Image Processing* Vol. 28, No. 1, 394–409, 2019.

[20] Xue, A. End-to-end Chinese landscape painting creation using generative adversarial networks. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 3862–3870, 2023.

[21] Dobler, K.; Hübscher, F.; Westphal, J.; Sierra-Múnera, A.; de Melo, G.; Krestel, R. Art creation with multi-conditional StyleGANs. *arXiv preprint* arXiv:2202.11777, 2022.

[22] Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint* arXiv:2204.06125, 2022.

[23] Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10674–10685, 2022.

[24] Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, Vol. 1, 1097–1105, 2012.

[25] Dumoulin, V.; Shlens, J.; Kudlur, M. A learned representation for artistic style. *arXiv preprint* arXiv:1610.07629, 2016.

[26] Li, Y. H.; Wang, N. Y.; Liu, J. Y.; Hou, X. D. Demystifying neural style transfer. *arXiv preprint* arXiv:1701.01036, 2017.

[27] Yin, R. J. Content aware neural style transfer. *arXiv preprint* arXiv:1601.04568, 2016.

[28] Gatys, L. A.; Bethge, M.; Hertzmann, A.; Shechtman, E. Preserving color in neural artistic style transfer. *arXiv preprint* arXiv:1606.05897, 2016.

[29] Kolkin, N.; Salavon, J.; Shakhnarovich, G. Style transfer by relaxed optimal transport and self-similarity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10043–10052, 2019.

[30] Kusner, M. J.; Sun, Y.; Kolkin, N. I.; Weinberger, K. Q. From word embeddings to document distances. In: Proceedings of the 32nd International Conference on Machine Learning, 957–966, 2015.

[31] Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision, 1510–1519, 2017.

[32] Wang, T. C.; Liu, M. Y.; Zhu, J. Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional GANs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8798–8807, 2018.

[33] Yi, Z. L.; Zhang, H.; Tan, P.; Gong, M. L. DualGAN: Unsupervised dual learning for image-to-image translation. In: Proceedings of the IEEE International Conference on Computer Vision, 2868–2876, 2017.

[34] Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J. Y.; Isola, P.; Saenko, K.; Efros, A. A.; Darrell, T. CyCADA: Cycle-consistent adversarial domain adaptation. In: Proceedings of the 35th International Conference on Machine Learning, 1989–1998, 2018.

[35] Zhu, J. Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A. A.; Wang, O.; Shechtman, E. Toward multimodal image-to-image translation. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 465–476, 2017.

[36] Huang, X.; Liu, M. Y.; Belongie, S.; Kautz, J. Multimodal unsupervised image-to-image translation. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11207*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 179–196, 2018.

[37] Lee, H. Y.; Tseng, H. Y.; Huang, J. B.; Singh, M.; Yang, M. H. Diverse image-to-image translation via disentangled representations. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11205*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 36–52, 2018.

[38] Lee, H. Y.; Tseng, H. Y.; Mao, Q.; Huang, J. B.; Lu, Y. D.; Singh, M.; Yang, M. H. DRIT++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision* Vol. 128, Nos. 10–11, 2402–2417, 2020.

[39] Liu, M. Y.; Huang, X.; Mallya, A.; Karras, T.; Aila, T. M.; Lehtinen, J.; Kautz, J. Few-shot unsupervised image-to-image translation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 10550–10559, 2019.

[40] Chen, X. Y.; Xu, C.; Yang, X. K.; Song, L.; Tao, D. C. Gated-GAN: Adversarial gated networks for multi-collection style transfer. *IEEE Transactions on Image Processing* Vol. 28, No. 2, 546–560, 2019.

[41] Chang, H. Y.; Wang, Z. X.; Chuang, Y. Y. Domain-specific mappings for generative adversarial style transfer. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12353*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 573–589, 2020.

[42] Park, T.; Zhu, J.-Y.; Wang, O.; Lu, J.; Shechtman, E.; Efros, A. A.; Zhang, R. Swapping autoencoder for deep image manipulation. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, Article No. 604, 7198–7211, 2020.

[43] He, B.; Gao, F.; Ma, D. Q.; Shi, B. X.; Duan, L. Y. ChipGAN: A generative adversarial network for Chinese ink wash painting style transfer. In: Proceedings of the 26th ACM International Conference on Multimedia, 1172–1180, 2018.

[44] Chen, L. C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 40, No. 4, 834–848, 2018.

[45] Wang, T. C.; Liu, M. Y.; Zhu, J. Y.; Liu, G. L.; Tao, A.; Kautz, J.; Catanzaro, B. Video-to-video synthesis. *arXiv preprint* arXiv:1808.06601, 2018.

[46] Choi, Y.; Choi, M.; Kim, M.; Ha, J. W.; Kim, S.; Choo, J. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8789–8797, 2018.

[47] Qi, X. J.; Chen, Q. F.; Jia, J. Y.; Koltun, V. Semi-parametric image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8808–8816, 2018.

[48] Wang, M.; Yang, G. Y.; Li, R. L.; Liang, R. Z.; Zhang, S. H.; Hall, P. M.; Hu, S. M. Example-guided style-consistent image synthesis from semantic labeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1495–1504, 2019.

[49] Zhang, S. Y.; Liang, R. Z.; Wang, M. ShadowGAN: Shadow synthesis for virtual objects with conditional adversarial networks. *Computational Visual Media* Vol. 5, No. 1, 105–115, 2019.

[50] Zhou, W. Y.; Yang, G. W.; Hu, S. M. Jittor-GAN: A fast-training generative adversarial network model zoo based on Jittor. *Computational Visual Media* Vol. 7, No. 1, 153–157, 2021.

[51] Wang, C.; Tang, F.; Zhang, Y.; Wu, T. R.; Dong, W. M. Towards harmonized regional style transfer and manipulation for facial images. *Computational Visual Media* Vol. 9, No. 2, 351–366, 2023.

[52] Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. M. Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8107–8116, 2020.

[53] Cohen, N.; Newman, Y.; Shamir, A. Semantic segmentation in art paintings. *Computer Graphics Forum* Vol. 41, No. 2, 261–275, 2022.

[54] Lin, T. Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C. L. Microsoft COCO: Common objects in context. In: *Computer Vision – ECCV 2014. Lecture Notes in Computer Science, Vol. 8693*. Fleet, D.; Pajdla, T.; Schiele, B.; Tuytelaars, T. Eds. Springer Cham, 740–755, 2014.

[55] Zhou, B. L.; Zhao, H.; Puig, X.; Xiao, T. T.; Fidler, S.; Barriuso, A.; Torralba, A. Semantic understanding of scenes through the ADE20K dataset. *International Journal of Computer Vision* Vol. 127, No. 3, 302–321, 2019.

[56] Boykov, Y.; Veksler, O.; Zabih, R. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 23, No. 11, 1222–1239, 2001.

[57] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint* arXiv:1409.1556, 2014.

[58] Xie, S. N.; Tu, Z. W. Holistically-nested edge detection. In: Proceedings of the IEEE International Conference on Computer Vision, 1395–1403, 2015.

[59] Gu, S.; Bao, J.; Chen, D.; Wen, F. GIQA: Generated image quality assessment. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12356*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 369–385, 2020.

[60] Fu, J.; Liu, J.; Tian, H. J.; Li, Y.; Bao, Y. J.; Fang, Z. W.; Lu, H. Q. Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3141–3149, 2019.

[61] Shen, Y. J.; Gu, J. J.; Tang, X. O.; Zhou, B. L. Interpreting the latent space of GANs for semantic face editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9240–9249, 2020.

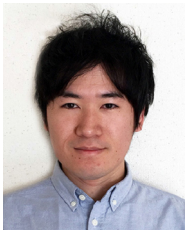
[62] McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint* arXiv:1802.03426, 2018.

[63] Johnson, J.; Alahi, A.; Li, F. F. Perceptual losses for real-time style transfer and super-resolution. In: *Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Vol. 9906*. Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 694–711, 2016.

[64] Kingma, D. P.; Welling, M. Auto-encoding variational Bayes. *arXiv preprint* arXiv:1312.6114, 2013.

[65] Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 6629–6640, 2017.

**Yuantian Huang** is a Ph.D. student at Degree Programs in Systems and Information Engineering, University of Tsukuba. His research interests include image editing and generation, computer graphics, and computer vision.



**Satoshi Iizuka** is an associate professor at University of Tsukuba's Faculty of Engineering, Information and Systems. He received his Ph.D. degree in engineering from the University of Tsukuba. His research interests are in computer graphics and vision, including image processing and editing based on machine learning.

**Edgar Simo-Serra** is an associate professor at Waseda University. He received his Ph.D. degree from BarcelonaTech (UPC). His general research interests are in the intersection of computer vision, computer graphics, and machine learning with applications to large-scale real-world problems.

**Kazuhiro Fukui** joined the Toshiba Corporate R&D Center and served as a senior research scientist at Multimedia Laboratory. He received his Ph.D. degree from Tokyo Institute of Technology in 2003. He is currently a professor in the Department of Computer Science, Graduate School of Systems and Information Engineering at the University of Tsukuba. His research interests include the theory of machine learning, computer vision, pattern recognition, and their applications. He is a member of IEEE and SIAM.