Neural Collage Transfer: Artistic Reconstruction via Material Manipulation

Ganghun Lee, Minji Kim, Yunsu Lee, Minsu Lee,* Byoung-Tak Zhang*

Seoul National University

{khlee, mjkim, yslee, mslee, btzhang}@bi.snu.ac.kr

Figure 1: The results of neural collage transfer. Given an image and materials, each output collage was generated based on the proposed complexity-aware multi-scale collage method. The sequence of grapes shows our collage generation process.

Abstract

Collage is a creative art form that uses diverse material scraps as a base unit to compose a single image. Although pixel-wise generation techniques can reproduce a target image in collage style, it is not a suitable method due to the solid stroke-by-stroke nature of the collage form. While some previous works for stroke-based rendering produced decent sketches and paintings, collages have received much less attention in research despite their popularity as a style. In this paper, we propose a method for learning to make collages via reinforcement learning without the need for demonstrations or collage artwork data. We design the collage Markov Decision Process (MDP), which allows the agent to handle various materials and propose a model-based soft actor-critic to mitigate the agent's training burden derived from the sophisticated dynamics of collage. Moreover, we devise additional techniques such as active material selection and complexity-based multi-scale collage to handle target images at any size and enhance the results' aesthetics by placing relatively more scraps in areas of high complexity. Experimental results show that the

trained agent appropriately selected and pasted materials to regenerate the target image into a collage and obtained a higher evaluation score on content and style than pixel-wise generation methods. Code is available at https://github.com/northadventure/CollageRL.

1. Introduction

Collage, derived from the French word *coller* meaning *to glue*, is a fundamental art form where disparate scraps of images are assembled and arranged geometrically to create a complete scene. The intriguing and alluring impression of collage was influenced by the art movement of *Cubism*, particularly the work of *Pablo Picasso*, and has now become a prevalent art style. However, creating high-quality collage artworks in the style of artists such as *Derek Gores* requires professional-level skills, much like other art forms such as painting. Collage may seem like a simple variation of painting, but it is much more challenging when attempted by artificial agent for several reasons. Unlike predefined brush strokes in paintings, the strokes of a collage depend on selecting materials from a diverse range of candidates, which involves uncertainty. After the material selection, the complicated manipulation process of cutting and past-

---

*Corresponding authors

ing follows. Moreover, detailed representations are strictly limited when the content of material pieces is unmodifiable. Even for a human, a large amount of experience is required to handle these difficulties.

Fortunately, recent advances in AI-based art creation have made art more accessible to ordinary people. Pixel-based generation approaches evolved from Deep Convolutional Neural Networks (DCNNs) [48, 14], Variational Auto-Encoders (VAEs) [23], and Generative Adversarial Networks (GANs) [9] have fostered the creation of novel artworks [39, 6] and transferred images into various styles [8, 5, 20, 57, 54, 16, 34, 25] under Neural Style Transfer (NST) [21]. However, pixel-based approaches directly discover the final look in the pixel space; the result can be seen as unnatural and cannot obtain a natural creation sequence.

Since most art formulations enjoy stroke-by-stroke, a more natural and aesthetic collage can be obtained if a machine learns to create collages in a stroke-by-stroke fashion. There have been some previous learning approaches in stroke-based rendering (SBR) [15] to imitate stroke-by-stroke image generation, such as supervised learning [11, 33, 29, 1], reinforcement learning (RL) [7, 17], and optimization [59, 24, 36]. Although they produced attractive sketches and paintings, they are partly incompatible with the concept of collages. For supervised learning, public collage process data is hardly available, and generating sufficient data is almost infeasible. Existing RL and optimization methods assuming the fixed material structure for every stroke do not align well with collages incorporating dynamic materials. Overall, collage has received comparatively less attention in the artistic creation domain.

In this paper, we introduce a novel RL-based method for learning how to generate a collage artwork using a given target image and materials. RL offers experience-based learning to handle various possible situations in the collage procedure without data cost, leading to the development of distinct creation styles. To effectively train the agent, we build a novel environment, the collage Markov Decision Process (MDP), which allows the agent to explore a wide range of material spaces to learn appropriate material selection and manipulation skills. We propose the model-based Soft Actor-Critic (MB-SAC) to mitigate the burden of predicting complicated dynamics in collage. At each training step, the agent observes the canvas, the given material, and the target image to decide on cutout shapes for the material and the pasting locations on the canvas. The reward increases as the canvas becomes more similar to the target image, encouraging the agent to learn more elaborate behaviors to reproduce the original content in a collage form. Moreover, we extend the trained collage agent to make a multi-scale collage for any size of the target images, considering the image complexity for more aesthetic appeal.

Our extensive experimental results show that our pro-

posed method enables an agent to learn the nature of collage autonomously without expensive collage datasets or collage demonstrations. Moreover, the proposed method can handle target images of any size, and the generated collage produces sophisticated and aesthetic output. This is because the agent can select appropriate scraps from materials and place more scraps in areas of high complexity. Furthermore, through quantitative analysis, we have shown that our approach outperforms other competitive methods in terms of image similarity from the target image and semantic consistency using CLIP [41]-based measures and user study. These results demonstrate that our method successfully transfers the target image to collage style while preserving its content.

2. Related Work

2.1. Neural Style Transfer

For the past couple of years, Neural Style Transfer (NST) [21] has been the standout tool in artistic style transfer. NST aims to convert a target image to the given target style preserving the content of the target image. General approaches of NST rely on pixel-wise gradient descent which is directly applied onto the converted image [8], or on using models trained to approximate the target style distribution [5, 16, 20, 57, 54, 56]. Some recent works support style transfer from text prompts [25, 45]. Although advanced NST models cover a wide range of styles, their utility for collage style is limited since the original purpose of NST may not match with collage. The purpose of their pixel-wise style extraction is to look for patterns commonly seen for each style type. However, as collages are composed of separate images, common patterns are rarely detected except for the edges between image pieces, thus not aligning with the purpose of NST. This unique feature of collage demands another approach for collage style transfer.

2.2. Stroke-based Rendering

Stroke-based Rendering (SBR) is a non-pixel-based automated approach for generating non-photorealistic imagery by arranging discrete elements such as paint strokes or stipples [15]. Deep learning-based SBR was first proposed in Sketch-RNN [11] where human sketch demonstration is provided to train the Recurrent Neural Networks (RNNs). However, supervised methods suffer from severe data collection and refinement costs. Instead, methods using RL [7, 17] train the painting agent without supervision. The gradient-based procedure optimization methods [24, 59] leverages the design of a fully-differentiable painting. Studies using transformers [49, 1, 29, 33] or robotics applications [27, 18] were also conducted.

Unfortunately, existing SBR-based research has mainly focused on the environments in which stroke structure is
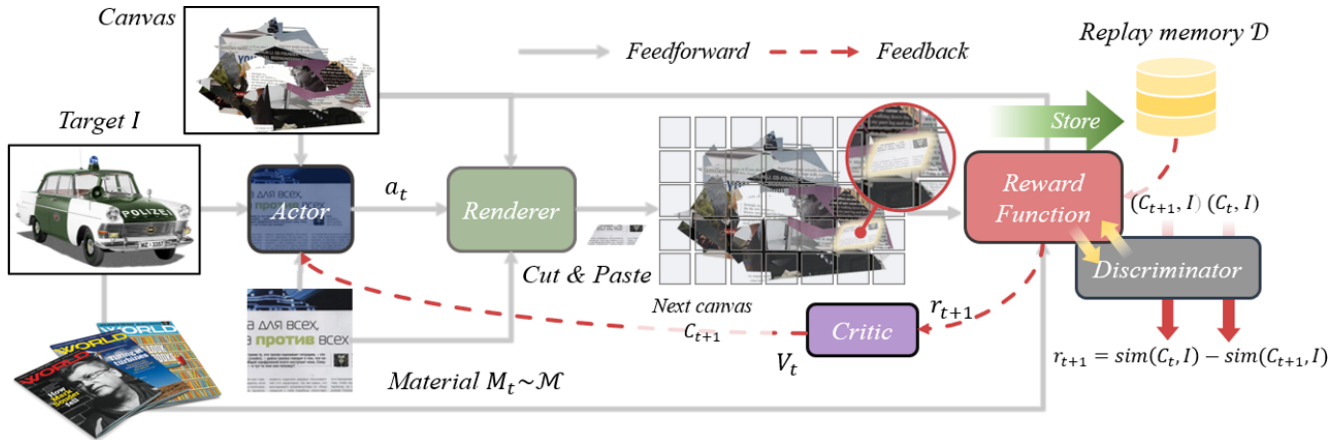
Figure 2: Overview of collage MDP and its associated training process. The agent first observes an image triplet consisting of a canvas, a target image, and a material image and then makes a scrap of the material that will make the canvas fit the target image as closely as possible. Once the scrap has been pasted onto the canvas, the transition process is saved in replay memory for training later. During the update stage, the discriminator in the reward function is updated using samples from the replay memory. The actor and critic are then updated using the rewards obtained from these samples.

pre-modeled (*e.g.*, sketches and paintings). These differ from collage environments where the available strokes (materials) are only known during the process. In this paper, we extend [17] to the collage domain, proposing a novel training environment and an alternative training algorithm.

2.3. Collage Generation

Several studies have explored collage generation, although not specifically collage transfer. Picture collage [51, 35, 31] aims to assemble a set of given pictures into a complete image, while CLIP-CLOP [36] generates collage artworks from text prompts using predefined strokes with modifiable properties. In contrast, our method uses non-predefined materials without unrealistic transformations for image transfer. Additionally, while their goal is text-based generation, our approach focuses on reproducing target images as natural collages. To the best of our knowledge, no research has been conducted on achieving this specific goal.

3. Proposed Methods

3.1. Collage MDP

Preliminary. The problem of RL is formulated as a Markov Decision Process (MDP) consisting of a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, where $\mathcal{S}$ is state space, $\mathcal{A}$ is action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ is the transition function, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function, and $\gamma \in [0, 1]$ is the discount factor. The MDP has a state $s_t \in \mathcal{S}$ at every time t. When the agent takes action $a_t \in \mathcal{A}$, the transition function $\mathcal{P}(\cdot)$ of the environment decides the next state $s_{t+1} = \mathcal{P}(s_t, a_t)$. Then the environment gives the reward $r_{t+1} = \mathcal{R}(s_t, a_t)$ to the agent. At every time t, the agent observes s_t and decides

the action $a_t \sim \pi(s_t)$ according to the policy $\pi(\cdot)$. In the finite horizon MDP, value V_π of the policy π is defined as $V_\pi(s_t) = r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^{T-t-1} r_T$, *i.e.*, the sum of discounted cumulative rewards until the end of the episode $t = T$. The objective of RL is to find the optimal policy $\pi^* = \underset{\pi}{\arg\max} \, V_\pi(s)$.

State and Transition. In constructing the collage MDP for our RL agents, we highlight the three key components: canvas C_t, target image I, and material M_t as illustrated in Fig. 2. The collage MDP is initialized with white canvas C_0, target image $I \sim \mathcal{I}$, and the initial material image $M_0 \sim \mathcal{M}$, where $\mathcal{I}$, and $\mathcal{M}$ are respectively the sets of target images and material images. If the agent takes the action a_t observing the canvas C_t, the material M_t is cut and pasted onto the canvas, resulting in the next canvas $C_{t+1} = \delta(C_t, M_t, a_t)$. The function δ is responsible for cut-and-paste, as an essential mechanism of collage state transition. The next material M_{t+1} is randomly sampled from $\mathcal{M}$. Two additional components are included in the state. One is CoordConv [32] c that is a constant vector known to help DCNNs to boost its performance in coordinating tasks. Another one is the remaining time $l_t = (T_M - t_M)/T_M$ about the number of material pieces that need to be pasted, where the constant T_M is the total number of pieces to paste, and t_M is the number of pieces already pasted before t. In summary, the observable state for the agent is $s_t = (C_t, I, M_t, l_t, c)$ at time t. The overall transition function can be represented using $\mathcal{P}$ as $s_{t+1} = \mathcal{P}(s_t, a_t) = (\delta(C_t, M_t, a_t), I, M_{t+1}, (T_M - t_M)/T_M, c)$ where $M_{t+1} \sim \mathcal{M}$.

Action Design. Each parameter of the agent's action determines the strategy for cutting and pasting the given material. We define the action as a twelve-dimensional vector $a = \langle x_{cut}, y_{cut}, w, h, p_1, p_2, p_3, p_4, x_{glue}, y_{glue}, \theta, v \rangle$. When cutting the material, the location, size, and shape are determined for a rectangle by center coordinates x_{cut}, y_{cut}, width w, and height h. Then the rectangle is divided into a quadrilateral scrap, following four points on each side based on the point ratio $p_{1:4}$. When pasting the piece, the pasting location x_{glue}, y_{glue} and rotation θ are determined. The material acceptor $v \in [0, 1]$ determines whether to use the given material. The agent can deny a poor given material and request another one if $v < 0.5$, and t_M increases only when the agent accepts the material. If the agent denies the material at t, a new material is given, but the canvas and remaining time do not change as $C_{t+1} = C_t$, $l_{t+1} = l_t$. The MDP normally terminates when $t_M = T_M$, but if the agent has denied too many given materials by $t = T_{max}$, the MDP terminates even if $t_M < T_M$.

Differentiable Collage. Since our training algorithm requires a differentiable $\mathcal{P}(s, a)$ about the action a, $\delta(C, M, a)$ should be differentiable. To ensure the function δ is differentiable, all operations involved in the process, from selecting an action to generating the next canvas, must also be differentiable. As in Fig. 3, we generate a $mask$ that specifies where the location of the cutting region and shape of the material, but the process of generating $mask$ is typically non-differentiable. To overcome this challenge, we use a pre-trained shaping network ψ that is capable of generating differentiable $mask$ images. Additionally, the material acceptor v is a discrete factor that is also non-differentiable. We address this issue by including v in the shaping network ψ, which has been pre-trained to output an all-zero mask when the material is denied. To perform the material pasting, we use a differentiable transformation operation from Kornia [44] that allows us to translate and rotate the material piece as needed. The new canvas is generated by adding the material piece to the current canvas, while excluding the $mask$ region. Through these modifications, we have achieved a fully differentiable series of connections that allows for the implementation of δ.

Reward Function. The main goal of collage transfer is to create a final canvas C_T that resembles the target image I as a collage artwork. As the style emerges naturally during the collage creation process, we designed a reward system that increases as the content of C_T becomes more similar to I. To measure the similarity between C_T and I, we used a reward scheme proposed in [17], which calculates the amount of similarity change as $r_t = sim(C_{t-1}, I) - sim(C_t, I)$. In this study, the discriminator in the Wasserstein GAN with gradient penalty
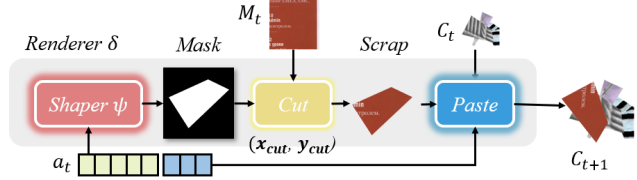

Figure 3: The differentiable rendering process in our collage MDP. The actions determining the cutting shapes are input into the pre-trained shaper network ψ. The resulting mask is then used to cut the material, generating a scrap to be pasted onto the canvas.

(WGAN-GP) [10] was used for sim. The sim discriminator is trained simultaneously with the collage agent to discriminate the canvases-in-progress and the target images. This dynamic reward system takes into account the agent's learning progress, resulting in better performance than using a constant reward, such as mean-squared error (MSE) [17]. In addition, we add step penalty -1 to the reward at every step to prevent the agent from neglecting materials.

3.2. Training

Model-based SAC. We propose a modified version of model-free SAC [12], called model-based Soft Actor-Critic (MB-SAC), to relieve the burden of learning collage dynamics in model-free RL. SAC is based on the maximum entropy RL framework [58] and adopts a stochastic policy approximating the optimal action distribution. Unlike traditional RL, the objective of the SAC policy update is not only to maximize the rewards but also to maximize the entropy of the policy.

The difference between the value functions of traditional RL and SAC is described as

$$V_\pi^{RL}(s_t) = \mathbb{E}_{a_t \sim \pi} \left[Q(s_t, a_t) \right], \qquad (1)$$
$$V_\pi^{SAC}(s_t) = \mathbb{E}_{a_t \sim \pi} \left[Q(s_t, a_t) - \log \pi(a_t|s_t) \right]. \qquad (2)$$

This ensures that the policy being trained behaves as randomly as possible while maximizing exploration, improving learning efficiency, and effectively finding improved behaviors. As an off-policy policy gradient method with an actor-critic structure, Model-free SAC utilizes past experience for training. When the past experience is stored in $\mathcal{D}$, the objectives of SAC policy evaluation and policy improvement are as follows [13].

$$
\begin{aligned}
J_Q = \mathbb{E}_{(s_t, a_t) \sim \mathcal{D}} [\tfrac{1}{2}(Q(s_t, a_t) \\
- (r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim \mathcal{P}} [V(s_{t+1})]))^2] \quad (3)
\end{aligned}
$$

$$J_\pi = \mathbb{E}_{s_t \sim \mathcal{D}} \left[\mathbb{E}_{a_t \sim \pi} \left[\alpha \log(\pi(a_t|s_t)) - Q(s_t, a_t) \right] \right] \quad (4)$$

Model-based RL (MBRL) [38] is an RL extension that can be used when the transition $\mathcal{P}$ and the reward function $\mathcal{R}$ are known or approximated. If the model is correct,

the training agent can leverage this knowledge about transitions and rewards to alleviate the training burden. In collage MDP, the $\mathcal{P}$ is known as a differentiable operator δ. The reward function is also known as described in 3.1. Thus, we can benefit from the model-based architecture. Zhou *et al.* [17] modified Deep Deterministic Policy Gradient (DDPG) [28] to a model-based version using the objective (1), but we use the alternative objective (2) to enhance exploration and performance of the agent. For our MB-SAC formulation, we use the equation (2) and its interchangeable relationship

$$Q(s_t, a_t) = r(s_t, a_t) + \mathbb{E}_{s_{t+1} \sim \mathcal{P}}\left[V(s_{t+1})\right], \quad (5)$$

to reformulate the objectives (3) and (4) as the following objectives:

$$J_V = \mathbb{E}_{(s_t, a_t) \sim \mathcal{D}}[\tfrac{1}{2}(V(s_{t+1}) - \mathbb{E}_{a_{t+1} \sim \pi}[r(s_{t+1}, a_{t+1}) + \gamma \mathbb{E}_{s_{t+2} \sim \mathcal{P}}[V(s_{t+2})] - \alpha \log(\pi(a_{t+1}|s_{t+1}))])^2], \quad (6)$$

$$J_\pi = \mathbb{E}_{s_t \sim \mathcal{D}}[\mathbb{E}_{a_t \sim \pi}[\alpha \log(\pi(a_t|s_t)) - (r(s_t, a_t) + \mathbb{E}_{s_{t+1} \sim \mathcal{P}}[V(s_{t+1})])]]. \quad (7)$$

Since $\mathcal{P}$ and $\mathcal{R}$ are differentiable, the above objectives allow policy gradients on the known transition and reward dynamics.

It should be noted that the transition model is not completely accurate due to a small error in ψ, and the agent is unaware of the upcoming material. However, for uncertain future events, the expectations in equations (6) and (7) combine the potential outcomes of $\mathcal{P}$, and therefore do not necessarily require knowledge of the upcoming material. Hence, the agent simply samples a new random material using the transition model to predict the next state.

Training Scheme. The agent learns to make a collage on the defined MDP, over a time range from $t = 0$ to at least $t = T_M$ (up to $t = T_{max}$) as one episode following the objectives (6) and (7). As the agent iterates through episodes, it gains experience in various state combinations (C, M, I, t). At the end of each episode, the RL and sim are updated together. This creates a virtuous cycle as sim gets tighter and tighter and the policy gets more and more accurate.

3.3. Advanced Techniques

Active Material Selection. During training, the agent makes binary decisions to accept or deny randomly given materials. However, humans actively select the best choice among multiple materials. We utilize the trained action-value function Q denoted in (5) without needing an additional material selector to achieve active material selection

like a human. If Q is accurate, then the optimal material for a given state is the one that maximizes Q, since Q represents the value of a state-action pair. The agent trains V instead of Q, but Q can be recovered using V and the reward function $\mathcal{R}$, as shown in (5). Therefore, the optimal material m_t^* from the set $\mathcal{M}$ can be selected using the following equation:

$$m_t^* = \underset{m}{\operatorname{argmax}}\left(r(s_t, a_t) + \gamma V(s_{t+1})\right), m \in \mathcal{M},$$
$$a_t = \mathbb{E}_{s_t \sim \mathcal{P}}\left[\pi(s_t)\right], s_{t+1} = \mathcal{P}(s_t, a_t). \quad (8)$$

It is worth noting that the agent can still deny the material even if it is optimal, especially if the canvas is already mature enough. In practice, the agent finds m_t^* from a subset of $\mathcal{M}$ to avoid high computational costs.

Multi-Scale Collage. We employ a multi-scale collage similar to the underpainting process to achieve a coarse-to-fine strategy for any size of an image. We define the multi-scale sequence $\mathcal{U} = (u_1, u_2, \ldots, u_n)$, where $u_1 > u_2 > \cdots > u_n$ and $u, n \in \mathbb{N}$ for the target image I with size $W \times H$. Then I is divided into smaller images using a sliding window with a window size of $u \times u$ and stride of $\lceil \rho \cdot u \rceil$, where ρ is the stride ratio. The number of divided target image regions on scale u is calculated as $k(u) = (\lceil (W - u)/\rho \rceil + 1)(\lceil (H - u)/\rho \rceil + 1)$. Finally, we acquire the set of divided target images $\mathbf{I}(u)$ and the set of corresponding initial divided canvases $\mathbf{C}(u)_{t=0}$ for every scale u.

$$\mathbf{I}(u) = (I(u)^1, I(u)^2, \ldots, I(u)^{k(u)}) \quad (9)$$
$$\mathbf{C}(u)_{t=0} = (C(u)_{t=0}^1, C(u)_{t=0}^2, \ldots, C(u)_{t=0}^{k(u)}) \quad (10)$$

Each divided target image is resized uniformly to the trained network's fixed input size during multi-scale collage. Therefore, the agent at earlier scales observes more broad area than at later scales but with higher information loss due to the resizing gap. We define "1-cycle" as one step of progress at each pair $(I(u)^i, C(u)_t^i)$ in a given scale u. A multi-scale collage takes K-cycles for each scale u, in the order $u_1, u_2, \ldots, u_n$, to create the final collage. The initial canvases on the first scale are all white, but the initial canvases of scales after the second inherit the final canvases of the previous scales.

Complexity-Aware Multi-Scale Collage. When creating a multi-scale collage, we utilize the insight that areas of a drawing with higher complexity require more intricate touches. We measure image complexity $Co(I)$ of the image I using an image gradient-based complexity measure introduced in [43]. We use the Sobel filter for the image gradient. The complexity is high when high color changes like edges are prevalent. Using this complexity measure, a
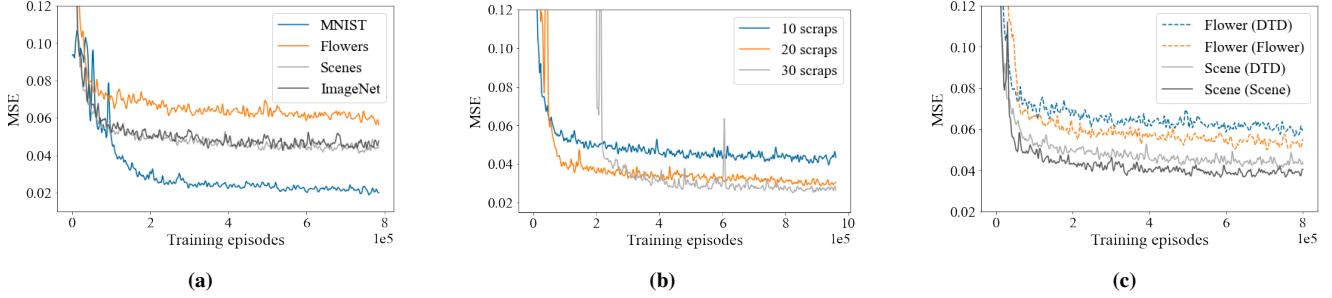
Figure 4: The learning curves of the model-based collage RL: (a) the learning curves of different target image domains using DTD, (b) the learning curves of agents trained to paste 10, 20, and 30 scraps using DTD, and (c) the learning curves of each agent using DTD and using materials from the same target domain.

set of complexities $\mathbf{Co}(I)_u$ of divided target images in (9) can be calculated for every u.

$$\mathbf{Co}(I)_u = (Co(I(u)^1), Co(I(u)^2), \ldots, Co(I(u)^{k(u)})) \tag{11}$$

Assuming the divided image complexities follow a normal distribution, we have the standardized complexities $\mathbf{Co}_Z(I)_u$ by the mean of complexity μ and the standard deviation σ from $\mathbf{Co}(I)_u$.

Referring to cumulative probability $Co_p(I(u)^i)$ for each z-score in $\mathbf{Co}_Z(I)_u$, the number of scraps for each divided canvas can be assigned proportional to Co_p. The scraps can be adjusted if the number of cycles K for each divided canvas are adjusted. Therefore, we applied monotonically increasing K with respect to the Co_p by the following equation

$$K(Co_p) = K_{max} \cdot Co_p^\tau, \tag{12}$$

where K_{max} is the maximum K for each divided canvas and $\tau \in (0, \infty)$ determines the sensitivity.

4. Experiments

To validate the proposed neural collage transfer method, this section presents implementation details of the training process, a comprehensive analysis of experimental results, and a comparison of the outcomes with existing methods.

4.1. Training

Implementation Details. For the backbones of actor and critic networks of the RL agents, ResNet-18 [14] with weight normalization [46] was adopted. Batch normalization [19] was applied to the actor. The discriminator used vanilla CNN with weight normalization, followed by global average pooling [30] at the end of the network. Translated ReLU [53] served as the activation function for all networks. Prior to training, ψ was pre-trained with pixel shuffle [47] for 100k epochs, using pairs of random actions and corresponding masks with batch size 64. During training, the differentiable δ was only used for MB-SAC

updates, while a non-differentiable transition function was utilized for the real MDP step, which excluded any errors from ψ. The replay memory size was set to 20K, and the agents were trained in 16 parallel environments. At the end of each episode, 5 RL updates and one discriminator update were performed, with a batch size of 64.

Datasets. For our experiments, we utilized various domain images, including MNIST [26], Flower [40], Scene[1], and ImageNet [3] as target image domains. For material images, we utilized Describable Textures Dataset (DTD) [2], Times[2], and Newspaper [50]. We adopted DTD as the primary material images for training due to its diverse textures and colors. Newspaper and Times were used as materials for multi-scale collage, as they are common types used in general collage. The data was split into training and evaluation images, ensuring that the agent encountered new images during evaluations.

Single-Scale Collage. To investigate the pure behavior of our agents, we demonstrated the agent's RL-based learning process and results on the fixed image size 128×128 without considering image division or complexities. Fig. 4a shows a learning curve of each agent trained using DTD on four target image domains. The MSE between the target image and the collage generated by the agent decreased as the training episode progressed. However, owing to image structure and domain variations, the range of MSE differed across the target images. Fig. 5 illustrates each agent's collage results using DTD, with a different number of scraps applied to each domain, as indicated in parentheses. For MNIST, the numbers in results are clearly recognizable, being placed with appropriate scrap lengths and sizes. For Flower and Scene, the scraps are correctly pasted considering their shapes and colors, even though the materials have
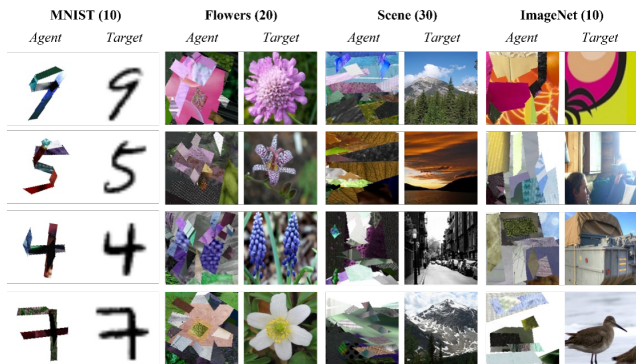
---

[1]https://www.kaggle.com/datasets/puneet6060/intel-image-classification

[2]https://www.kaggle.com/datasets/thegupta/time-magazine-part-1-1923-to-1930

Figure 5: Single-scale collage results using DTD.

Figure 6: Single-scale collage results using materials from the same target domain.

different properties than the target image domain. Notably, for the result in the third column in Flower, the agent could even consider textures if the materials permitted. As ImageNet contains an extensive variety of images, the agent focused on reproducing general colors and shapes, rather than domain-specific collages. Overall, the agent selected and arranged the materials at the appropriate level, resulting in collages that closely resemble the target images at a glance, while each pasted scrap retains its independence when viewed up close.

Fig. 4b illustrates the learning curve of each agent trained to paste 10, 20, and 30 scraps using DTD for the target images from Scene. As the number of scraps increases, the MSE decreases because more room to refine representation is available when more scraps are given. However, it is worth noting that using many scraps does not always lead to high-quality collage; as too many scraps may harm the nature of the collage. It depends on the user's preferred style. Fig. 4c displays the differences in MSE between using DTD and the same materials as the corresponding target images. Because images in the same domain have similar shapes and colors, a lower MSE was achieved when the material image domain was the same as the target image domain. Resulting images of these cases are depicted in Fig. 6. Although the agents learned without labels, they used similar types of scraps with a target area and sometimes placed scraps with similar shapes and colors in different classes.

Complexity-aware Multi-scale Collage. To validate our proposed multi-scale collage approach with complexity comprehension, we trained the agent on ImageNet as the target image domain and DTD. The agent was trained to paste 10 scraps, and the input image size was 64×64. In the single-scale collage, the agent tended to select smaller scraps as t_m increased. To encourage the agent to use more detailed scraps, we fixed $t_m = 9$ (not always good). For the hyper-parameters, we used $\mathcal{U} = \{512, 256, 128, 64, 32\}$, $\rho = 0.5$, $K_{max} = 8$ and $\tau = 1$. Newspapers and Times materials are used with active material selection techniques.
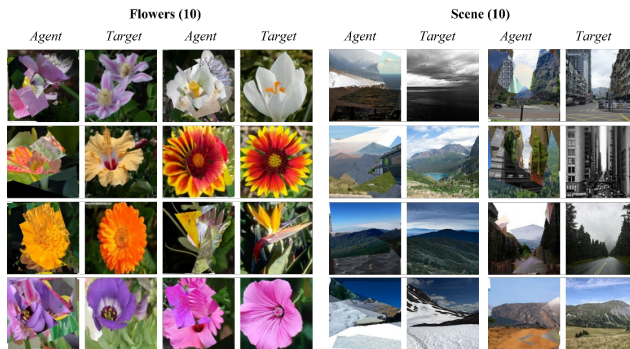
The final results and the results for each scale are shown in Fig. 1. As illustrated below, We found that as the scale increased, the collage became more abstract and unique in style but with higher distortion. On the other hand, smaller scales focused more on precise representation but lost some unique collage style. Overall, the number of scraps pasted on the canvas differed depending on the target image's partial complexity.

4.2. Comparison with NST

To evaluate our method's content maintenance and collage style agreement comparable to NST, quantitative and qualitative comparisons with the five NST methods [8, 16, 34, 22, 56] were conducted. For the quantitative comparison, metric-based evaluation and user study were taken.

For the metric-based evaluation, we prepared 30 target images for each method. Using these 30 target images, we collected 30 pairs of (target image, generated image) for our multi-scale method. The 30 pairs for each NST method were collected using style images consisting of five human-made collage images and five cluttered paper images. We adopted CLIP score [42] for both the content maintenance metric and style agreement metric and adopted LPIPS [55] only for the content maintenance metric. Here, the CLIP score is the cosine similarity between the CLIP representation of the image and text. We considered CLIP as a precise and reproducible protocol and even objective since CLIP was trained with about 400M human-labeled data.

Table 1 demonstrates the metric-based evaluation. Each value in the table is the means and standard deviations for the 30 pairs. In *CLIP score* experiment, the methods are evaluated using the following texts: (*content*) "a *content name*", (*human*) "a human-made collage image", and (*collage*) "a collage artwork", where *content name* is the name of each target image (*e.g.*, a bird). We assumed the content score as the CLIP score of the text-image pair (*content*, generated image) and the style score as the CLIP score of the pair (*human/collage*, generated image). The rows un-
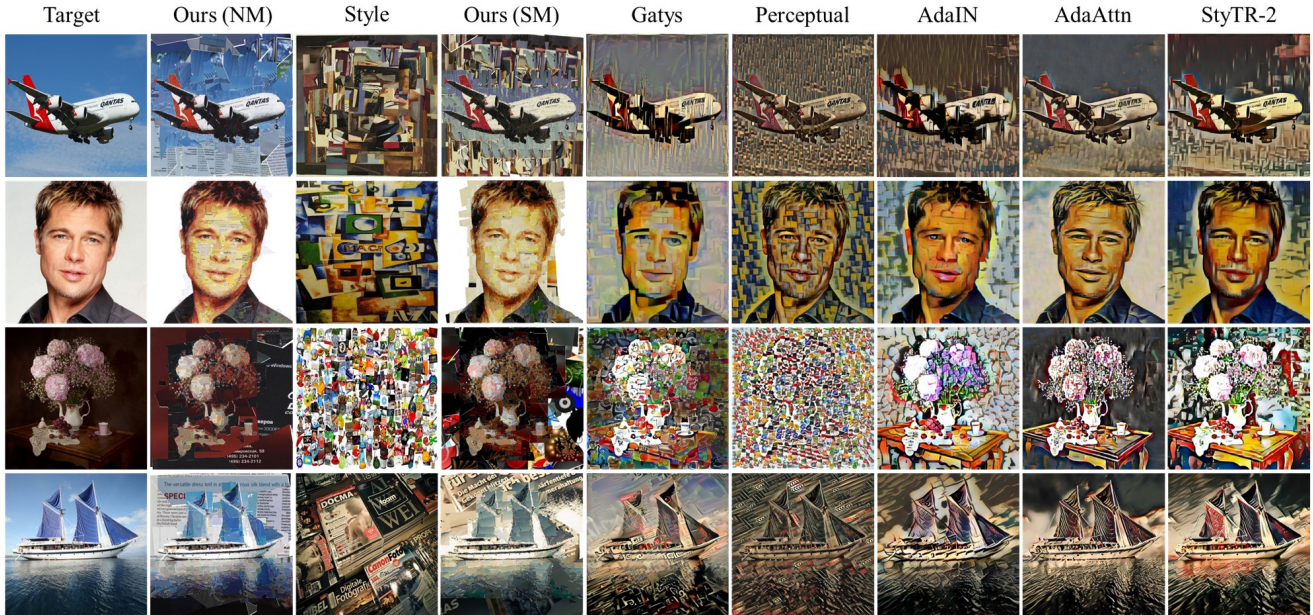
Target   Ours (NM)   Style   Ours (SM)   Gatys   Perceptual   AdaIN   AdaAttn   StyTR-2

Figure 7: Comparison of NST results using a given collage style images and the results from the proposed method. Ours (NM) used numerous other materials, while Ours (SM) solely used the given collage-style image as materials.

| Methods | CLIP score ↑ [42] | | | CLIP vote ↑ | LPIPS [55] ↓ |
	content	human	collage	collage	VGG
Target	0.276 ± 0.027	0.213 ± 0.018	0.200 ± 0.017	0.633	-
AdaAttn [34]	0.278 ± 0.021	0.247 ± 0.018	0.241 ± 0.010	0.027	0.597 ± 0.103
Adain [16]	0.251 ± 0.019	0.239 ± 0.010	0.236 ± 0.008	0.017	0.662 ± 0.103
Gatys [8]	0.226 ± 0.013	0.260 ± 0.006	0.250 ± 0.006	0.290	0.708 ± 0.098
Perceptual [22]	0.239 ± 0.019	0.246 ± 0.006	0.234 ± 0.007	0.307	0.722 ± 0.117
StyTR-2 [4]	0.261 ± 0.023	0.238 ± 0.010	0.235 ± 0.009	0.027	0.613 ± 0.115
Ours (32)	**0.280 ± 0.026**	0.262 ± 0.017	0.281 ± 0.020	0.100	**0.510 ± 0.111**
Ours (64)	0.262 ± 0.028	0.272 ± 0.020	0.259 ± 0.015	0.667	0.565 ± 0.112
Ours (128)	0.225 ± 0.023	**0.288 ± 0.015**	**0.272 ± 0.016**	**1.000**	0.610 ± 0.115

Table 1: Quantitative results in terms of CLIP score and LPIPS for the generated images from different NST methods and ours.

der *CLIP score* show that our generated images achieved a relatively high content and style score. However, the content score significantly decreased as the scale grew, indicating that larger scales emphasized collage style over content. The style scores of our generated images were always the highest. In *CLIP vote* experiment, CLIP predicted the probabilities of how much the generated image matches with the following two texts: (*collage*) "human-made collage", (*nst*) and "neural style transferred artwork." This was intended to investigate if CLIP can distinguish between the NST results and our results based on the semantics of "collage" and "neural style transferred." The rows under *CLIP vote* show that our generated images at larger scales gained the highest probability on *collage* compared to *nst*, but it decreased at scale 32, showing the scale-dependent style-content trade-off of our method.

Further content-focused evaluation using LPIPS with

	style ↑	content ↑	aesthetic ↑	ai ↓	human ↑
AdaAttn	18.3 ± 9.2	38.4 ± 14.8	30.0 ± 10.3	34.2 ± 6.5	31.6 ± 11.2
AdaIN	24.9 ± 8.4	21.1 ± 8.2	26.7 ± 1.5	35.8 ± 5.1	24.3 ± 5.8
Gatys	35.5 ± 5.4	17.4 ± 6.5	26.6 ± 10.2	35.6 ± 6.2	27.4 ± 9.4
Perceptual	26.2 ± 8.2	6.7 ± 3.4	25.4 ± 5.0	37.0 ± 7.9	25.1 ± 7.6
StyTR-2	30.5 ± 3.2	44.1 ± 20.3	41.4 ± 8.5	30.4 ± 8.1	31.5 ± 9.1
Ours	**64.6 ± 13.2**	**72.2 ± 13.3**	**49.9 ± 8.9**	**27.0 ± 6.0**	**60.2 ± 13.4**
Ours (32)	51.8 ± 4.5	**92.8 ± 4.7**	**73.4 ± 5.3**	-	-
Ours (64)	72.5 ± 0.6	67.0 ± 0.8	70.4 ± 2.8	-	-
Ours (128)	**75.7 ± 5.1**	40.2 ± 3.8	56.2 ± 8.1	-	-

Table 2: Quantitative analysis from user study.

VGG [48] was conducted. According to the values in rows under *LPIPS*, our approach achieved a relatively low LPIPS distance, confirming that the proposed method can stably maintain the content of the target image. Since NST treats the target image's color as style, LPIPS distance was relatively low for our approach, which treats the color as content. These results make sense since NST does not use various materials and does not actually create a collage.

We also conducted a user study on 111 non-conflicting respondents, as in Table 2. Each user responded to the three to four question sets. Each set firstly displayed the user a target image and the generated images from NSTs and our method. Then it required the user to rank the top three images for the following five questions: "Which one is the most ..." (*style*) "collage-style?" (*content*) "content-preserving?" (*aesthetic*) "aesthetically appealing?" (*ai*) "AI-generated?" (*human*) "human-created?". The scores were averaged, with weights of 3, 2, and 1 for each rank. Surprisingly, the result aligns well with the metric-based evaluation. Each user also responded to the two additional ques-

tion sets to rank our generated images at scales 128, 64, and 32 for the questions *style*, *aesthetic*, and *human*. Once again, the result supports findings from the metric-based evaluation.

Fig. 7 illustrates the transfer examples of our and NST methods. Ours (NM) refers to *numerous materials*, which means that other images were used as materials, regardless of the style image. Ours (SM) refers to *single material*, meaning only the given style image was used as material. Although NST produced aesthetically appealing transferred results, it did not achieve the intended collage style. Ours (SM) constructed the collage artwork from white canvas, even though only the single style image was given as a material. While NST refers to a single style image and follows the colors in the style image, our method can select materials with proper colors from other image sources to construct content-color-maintaining collages if materials are abundant, as illustrated in the results of Ours (NM).

4.3. Ablation Studies

Ablation studies were conducted to investigate the effects of the main components of training. In this study, agents were trained using Scene as target images and DTD as materials. The agents were trained with 800K episodes for five seeds using input image size 64×64.

RL Algorithm. To confirm if the proposed MB-SAC outperforms existing algorithms, we evaluated four RL algorithms, model-free DDPG [28], model-free SAC [12], MB-DDPG [17], and MB-SAC. Three evaluation metrics, MSE, PSNR, and SSIM, are used as image distance measure. The resulting values are described in table 3. The model-free DDPG almost failed to learn, while the model-free SAC succeeded. DDPG's failure was likely due to the high complexity and uncertainty of the collage procedure, which SAC was able to withstand. Model-based approaches show better performance than model-free as the model-based approach successfully reduced the learning burden of the transition and reward of the MDP. As expected, our MB-SAC achieved the best performance by leveraging the strengths of both SAC and model-based approach.

Reward Function. Since the reward is a vital signal for RL, we investigated alternative discriminators [37, 52] proposed after WGAN-GP as demonstrated in table 3. Among them, WGAN-GP is still the most potent signal for our RL agent, though SNGAN [37] and GNGAN [52] are reported to perform better in general GAN training. Both methods stabilize general GAN training by limiting the output of the discriminator, but it seems that they distort the distance metric between images, making them less informative when used as rewards for RL. Generally, a dense reward with a precise scale is considered helpful than a sparse reward; in that regard, WGAN-GP, which mostly maintains

	MSE ↓	PSNR ↑	SSIM ↑
RL algorithms			
DDPG [28]	0.305 ± 0.144	6.092 ± 3.919	0.213 ± 0.034
SAC [12]	0.042 ± 0.002	13.78 ± 0.285	0.191 ± 0.013
MB-DDPG [17]	0.039 ± 0.004	14.06 ± 0.514	0.204 ± 0.016
MB-SAC (ours)	$\mathbf{0.031} \pm \mathbf{0.001}$	$\mathbf{15.09} \pm \mathbf{0.219}$	$\mathbf{0.220} \pm \mathbf{0.004}$
Rewards			
MSE	0.122 ± 0.025	9.191 ± 0.893	0.101 ± 0.018
WGAN-GP [10]	$\mathbf{0.031} \pm \mathbf{0.001}$	$\mathbf{15.09} \pm \mathbf{0.219}$	$\mathbf{0.220} \pm \mathbf{0.004}$
SNGAN [37]	0.040 ± 0.002	13.96 ± 0.256	0.205 ± 0.014
GNGAN [52]	0.084 ± 0.035	11.06 ± 1.933	0.144 ± 0.038

Table 3: The results of ablations under the different RL algorithms and reward components.

the *Wasserestein* distance was the most informative discriminator for reward signals. The MSE reward signal showed poor performance, even in the MSE evaluation. It indicates that using the reward signal relative to the performance (*e.g.* GANs) of the learning agent was critical in training.

5. Conclusion and Future Works

This paper presented a novel RL-based training architecture, MB-SAC, and complexity-aware multi-scale collage techniques for stroke-based collage transfer. The experimental results demonstrated that our method enables the agents to autonomously learn to create collage with its distinct style without demonstration data. The agents could handle target images of any size and generate sophisticated and aesthetically pleasing output while maintaining collage style and content.

Since this study represents the first exploration of neural collage transfer using materials as-is, several limitations provide ample opportunity for future improvements. Following the proposed method, the stroke shape is limited to quadrilateral, so extending it to more unconstrained shapes is a promising future direction. The stroke choice only considers color and shape, disregarding semantics. Incorporating semantics or contexts of materials and targets would also be an interesting extension. Adding custom reward factors to reflect intentional distortions or style variations is also possible. For practicality, adding interactive commands to cater to the users' tastes would be a powerful use case. Additionally, the proposed methods can be extended beyond collage to other art forms that may not align with traditional pixel-based or fixed-stroke-based approaches.

References

[1] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan, Jorma Laaksonen, and Michael Felsberg. Doodleformer: Creative sketch drawing with transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 338–355. Springer, 2022. 2

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 6

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[4] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11326–11336, 2022. 8

[5] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016. 2

[6] Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068*, 2017. 2

[7] Yaroslav Ganin, Tejas Kulkarni, Igor Babuschkin, SM Ali Eslami, and Oriol Vinyals. Synthesizing programs for images using reinforced adversarial learning. In *International Conference on Machine Learning*, pages 1666–1675. PMLR, 2018. 2

[8] Leon Gatys, Alexander Ecker, and Matthias Bethge. A neural algorithm of artistic style. *Journal of Vision*, 16(12):326–326, 2016. 2, 7, 8

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[10] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017. 4, 9

[11] David Ha and Douglas Eck. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017. 2

[12] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018. 4, 9

[13] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018. 4

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 6

[15] Aaron Hertzmann. A survey of stroke-based rendering. Institute of Electrical and Electronics Engineers, 2003. 2

[16] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 2, 7, 8

[17] Zhewei Huang, Wen Heng, and Shuchang Zhou. Learning to paint with model-based deep reinforcement learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8709–8718, 2019. 2, 3, 4, 5, 9

[18] Ivaylo Ilinkin, Daeun Song, and Young J Kim. Stroke-based rendering and planning for robotic performance of artistic drawing. *arXiv preprint arXiv:2210.07590*, 2022. 2

[19] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 6

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[21] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 26(11):3365–3385, 2019. 2

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 7, 8

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[24] Dmytro Kotovenko, Matthias Wright, Arthur Heimbrecht, and Bjorn Ommer. Rethinking style transfer: From pixels to parameterized brushstrokes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12196–12205, 2021. 2

[25] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022. 2

[26] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 6

[27] Ganghun Lee, Minji Kim, Minsu Lee, and Byoung-Tak Zhang. From scratch to sketch: Deep decoupled hierarchical reinforcement learning for robotic sketching agent. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 5553–5559. IEEE, 2022. 2

[28] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and

Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015. 5, 9

[29] Hangyu Lin, Yanwei Fu, Xiangyang Xue, and Yu-Gang Jiang. Sketch-bert: Learning sketch bidirectional encoder representation from transformers by self-supervised learning of sketch gestalt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6758–6767, 2020. 2

[30] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 6

[31] Lingjie Liu, Hongjie Zhang, Guangmei Jing, Yanwen Guo, Zhonggui Chen, and Wenping Wang. Correlation-preserving photo collage. *IEEE transactions on visualization and computer graphics*, 24(6):1956–1968, 2017. 3

[32] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. *Advances in neural information processing systems*, 31, 2018. 3

[33] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Ruifeng Deng, Xin Li, Errui Ding, and Hao Wang. Paint transformer: Feed forward neural painting with stroke prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6598–6607, 2021. 2

[34] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6649–6658, 2021. 2, 7, 8

[35] Tie Liu, Jingdong Wang, Jian Sun, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Picture collage. *IEEE Transactions on Multimedia*, 11(7):1225–1239, 2009. 3

[36] Piotr Mirowski, Dylan Banarse, Mateusz Malinowski, Simon Osindero, and Chrisantha Fernando. Clip-clop: Clip-guided collage and photomontage. *arXiv preprint arXiv:2205.03146*, 2022. 2, 3

[37] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 9

[38] Thomas M Moerland, Joost Broekens, Aske Plaat, Catholijn M Jonker, et al. Model-based reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 16(1):1–118, 2023. 4

[39] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. 2015. 2

[40] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 6

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7, 8

[43] Christoph Redies, Seyed Ali Amirshahi, Michael Koch, and Joachim Denzler. Phog-derived aesthetic measures applied to color photographs of artworks, natural scenes and objects. In *Computer Vision–ECCV 2012. Workshops and Demonstrations: Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 522–531. Springer, 2012. 5

[44] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3674–3683, 2020. 4

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[46] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29, 2016. 6

[47] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 6

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 8

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[50] AM Vil'kin, IV Safonov, and MA Egorova. Algorithm for segmentation of documents based on texture features. *Pattern recognition and image analysis*, 23:153–159, 2013. 6

[51] Jingdong Wang, Long Quan, Jian Sun, Xiaoou Tang, and Heung-Yeung Shum. Picture collage. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 347–354. IEEE, 2006. 3

[52] Yi-Lun Wu, Hong-Han Shuai, Zhi-Rui Tam, and Hong-Yu Chiu. Gradient normalization for generative adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6373–6382, 2021. 9

[53] Sitao Xiang and Hao Li. On the effects of batch and weight normalization in generative adversarial networks. *arXiv preprint arXiv:1704.03971*, 2017. 6

[54] Wenju Xu, Chengjiang Long, Ruisheng Wang, and Guanghui Wang. Drb-gan: A dynamic resblock generative adversarial network for artistic style transfer. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 6383–6392, 2021. 2

[55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7, 8

[56] Zhentan Zheng, Jianyi Liu, and Nanning Zheng. P^{2}-gan: Efficient stroke style transfer using single style image. *IEEE Transactions on Multimedia*, 2022. 2, 7

[57] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2

[58] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008. 4

[59] Zhengxia Zou, Tianyang Shi, Shuang Qiu, Yi Yuan, and Zhenwei Shi. Stylized neural painting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15689–15698, 2021. 2