Intelli-Paint: Towards Developing More Human-Intelligible Painting Agents

Jaskirat Singh[1,2], Cameron Smith[2], Jose Echevarria[2], and Liang Zheng[1]
{jaskirat.singh,liang.zheng}@anu.edu.au, {casmith,echevarr}@adobe.com

[1] Australian National University
[2] Adobe Research

Abstract. Stroke based rendering methods have recently become a popular solution for the generation of stylized paintings. However, the current research in this direction is focused mainly on the improvement of final canvas quality, and thus often fails to consider the intelligibility of the generated painting sequences to actual human users. In this work, we motivate the need to learn more human-intelligible painting sequences in order to facilitate the use of autonomous painting systems in a more interactive context (*e.g.* as a painting assistant tool for human users or for robotic painting applications). To this end, we propose a novel painting approach which learns to generate output canvases while exhibiting a painting style which is more relatable to human users. The proposed painting pipeline Intelli-Paint consists of **1)** a progressive layering strategy which allows the agent to first paint a natural background scene before adding in each of the foreground objects in a progressive fashion. **2)** We also introduce a novel sequential brushstroke guidance strategy which helps the painting agent to shift its attention between different image regions in a semantic-aware manner. **3)** Finally, we propose a brushstroke regularization strategy which allows for ∼60-80% reduction in the total number of required brushstrokes without any perceivable differences in the quality of generated canvases. Through both quantitative and qualitative results, we show that the resulting agents not only show enhanced efficiency in output canvas generation but also exhibit a more natural-looking painting style which would better assist human users express their ideas through digital artwork.

1 Introduction

Paintings form a key medium through which humans express their ideas and emotions. Nevertheless, the creation of finer-quality art is often quite challenging and and requires a considerable amount of time on part of the human painter.

One way to address this problem is to develop autonomous painting agents which can assist human painters to better express their ideas in a quick and concise fashion. To this end, there is a growing research interest [8, 9, 11, 14, 18, 23, 28–30, 35, 44, 48, 49] in teaching machines *"how to paint"*, in a manner similar to a human painter. For instance, Huang *et al.* [18] use deep reinforcement learning

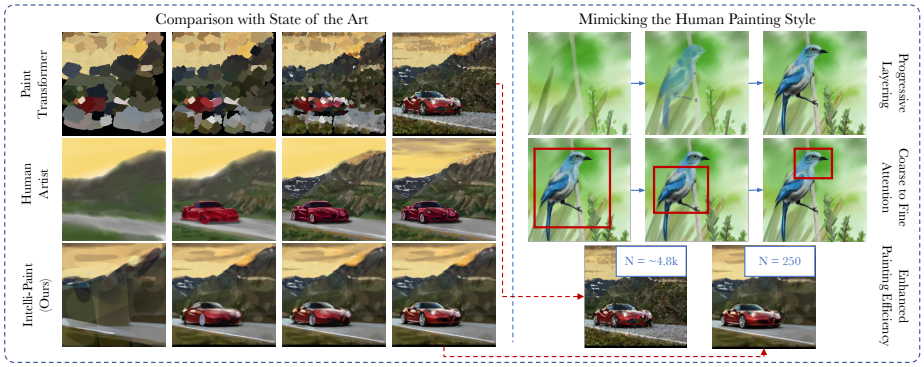

Fig. 1. Developing a more human-relatable painting style. (Left) Painting sequence visualization which demonstrates that our method exhibits higher resemblance with the human painting style as opposed to previous state of the art. (Right) This resemblance is achieved through **1)** a progressive layering strategy which allows for a more human-like evolution of the canvas, **2)** a sequential attention mechanism which focuses on different image regions in a coarse-to-fine fashion and **3)** a brushstroke regularization formulation which allows our method to obtain detailed results while using significantly fewer brushstrokes ($\sim$1/20 as compared to Paint Transformer [28] in above).

to learn an unsupervised brushstroke decomposition for the creation of non-photorealistic imagery. Zou *et al*. [49] use gradient descent to optimize over the brushstroke parameters for the entire painting trajectory. Similarly, Liu *et al*. [28] propose a novel Paint Transformer which formulates the "learning to paint" problem as a feed-forward set prediction problem. Despite their efficacy, existing works often lack semantic understanding of image contents and are invariably reliant on a progressive grid-based division strategy, wherein the painting agent divides the overall image into successively finer grids, and then proceeds to paint each in parallel. This leads to hierarchically bottom-up painting sequences which are quite mechanical and thus not applicable for human users.

In this paper, we propose a novel painting pipeline (*intelli-paint*), which tries to address the need for more human-intelligible painting sequences, by mimicking some commonly found traits of the human painting process. This is achieved in three main ways. **First,** we propose a progressive layering strategy which, much like a human, allows the painting agent to successively draw a given scene in multiple layers. That is, instead of starting to paint the entire scene at once, our method learns to first paint a realistic background scene representation before adding in each of the foreground objects in a progressive layerwise fashion.

Second, the human painting process is often characterized by a localized spatial attention span. For instance, a potential artist would focus on different local image areas while painting distinct parts of the final canvas [47]. This is in sharp contrast with previous works, which either focus on the entire image

or several predefined grid blocks [28, 49]. To better mimic the human style, we introduce a sequential brushstroke guidance approach which allows the painting agent to shift its attention between different image areas through a self-learned sequence of localized attention windows. The spatial dimensions and position of the localized attention window are progressively adjusted during the painting process so as to paint a given scene in a coarse-to-fine fashion.

Third, we note that prior works often use a fixed brushstroke budget irrespective of the complexity of the target image. This not only leads to wasteful (and overlapping) brushstroke patterns (refer Fig. 4) but also imparts an artificial painting style to the final agent. To this end, we propose an inference-time brushstroke regularization formulation which removes brushstroke redundancies by regularizing the total number of brushstrokes required for painting a given canvas. Our experiments reveal that this not only leads to a ∼60-80% enhancement in the brushstroke decomposition efficiency but also results in more natural looking painting sequences which are easily intelligible by a human painter.

To summarize, this paper makes the following contributions.

- We introduce a progressive layering approach, which much like a human, allows the painting agent to draw a given scene in multiple successive layers.
- We propose a sequential brushstroke guidance strategy which enables the painting agent to focus on different image regions through a learned sequence of coarse-to-fine localized attention windows.
- Finally, we introduce an inference time brushstroke regularization procedure which results in a ∼60-80% enhancement in the brushstroke decomposition efficiency and leads to more natural painting sequences. which are better intelligible by a human user.

2 Related Work

Classical stroke based rendering. The problem of teaching machines "how to paint" has been extensively studied in the context of stroke-based rendering (SBR), which focuses on the recreation of non-photorealistic imagery through appropriate positioning and selection of discrete elements such as paint strokes or stipples [15–17, 26, 38, 40, 46]. Classical works for incorporating semantic knowledge into the painting process have also been explored. [2, 5–7] use image saliency to generate a coarse to fine painting sequence in which sailent details (*e.g.* edges) are preserved in increasing amounts. In contrast, our work uses image saliency to learn a progressive layering strategy in which the agent learns to paint a natural background scene (refer Fig. 2) before adding in the foreground objects in a progressive fashion. [34, 39] use local attention over a heuristically determined window for computation of stroke parameters. Our work differs as it provides an unsupervised approach for learning the optimal movement of this attention window for painting in a coarse-to-fine manner (refer Sec. 4.1).

Supervised painting methods. More recent solutions [13, 14] adopt the use of recurrent neural networks for computing optimal brushstroke decomposition. However, these methods require access to dense human brushstroke annotations,

which limits their applicability to most real world problems. In another work, Zhao *et al*. [47] use a conditional variational autoencoder framework for synthesising time-lapse videos depicting the recreation of a given target image. However, this requires access to painting time-lapse videos from real artists for training. Furthermore, the time-lapse outputs are generated at very low-resolution as compared to the high-resolution sequences generated using our approach.

Unsupervised painting methods. In recent years, there has been an increased focus on learning an unsupervised brushstroke decomposition without requiring access to dense human brushstroke annotations. For instance, recent works [11, 18, 19, 29, 35, 44] use deep reinforcement learning and an adversarial training approach for learning an efficient brushstroke decomposition. Optimization based methods [49] directly search for the optimal brushstroke parameters by performing gradient descent over a novel optimal-transport-based loss function. In another recent work, Liu *et al*. [28] propose a Paint Transformer which formulates the painting problem as a feed-forward stroke set prediction problem.

While the above works show high proficiency in painting high-quality output canvases, the generation of the same invariably depends on a progressive grid-based division strategy. In this setting, the agent divides the overall image into successively finer grids, and then proceeds to paint each of them in parallel. Experimental analysis reveals that this not only reduces the efficiency of the final agent, but also leads to mechanical (grid-based) painting sequences which are not directly applicable to actual human users.

3 Need for More Human-Intelligible Painting Agents

Interactive applications. The need for stroke-based-rendering methods (as opposed to pixel-based methods [12, 21]) is often motivated from the need to mimic the human artistic creation process [3, 28, 49], which can then be used for development of painting assistant and teaching tools [18, 28] for human users. The development of more human-intelligible painting sequences is thus important as it will allow for the use of autonomous painting methods in an interactive context.

Robotic painting tasks. Robotic applications for expression of AI creativity are being increasingly explored [22, 31, 42]. Our contribution is significant in this direction, as our method not only learns a painting sequence which is more interpretable to actual human users, but more importantly it provides an *efficient painting plan* which would allow a robotic agent to paint a vivid scene using significantly less number of brushstrokes as compared to previous works.

Approximating the manifold of human paintings. While sketch based methods for photorealistic image generation [4, 29, 43, 45] have been extensively studied, the use of partially drawn human paintings for image synthesis remains unexplored due to lack of large-scale collection of human (or human-like) painting trajectories. In an exciting concurrent work, Singh *et al*. [36] use the improved human-likeliness of *our* painting sequences in order to perform photorealistic image synthesis and editing from rudimentary user paintings and brushstrokes.

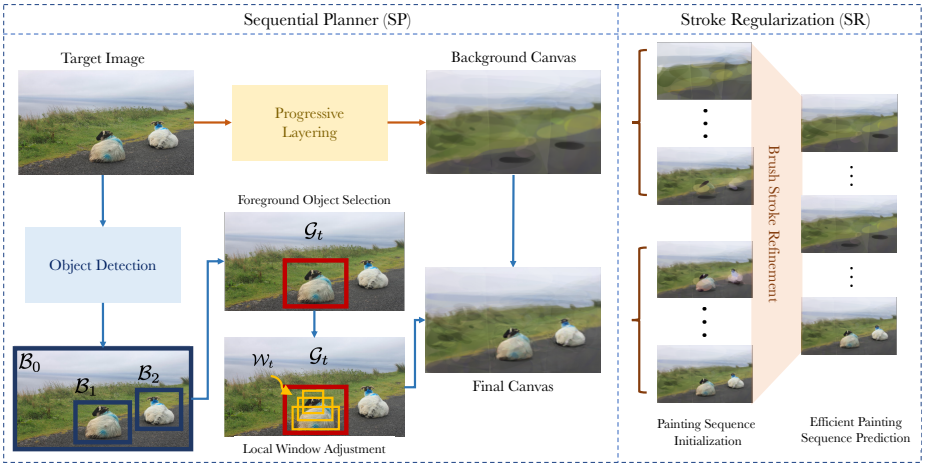

Fig. 2. Method Overview. Given a target image I, the *Intelli-Paint* agent first learns to paint a realistic background scene on the canvas. Once the background scene has been painted, the agent then proceeds to progressively add each of the foreground objects using a sequential brushstroke guidance procedure. To do this, the painting agent first uses the convex combination formulation from Eq. 6 to select the foreground object it would like to paint (indicated by object window $\mathcal{G}_t$). The features within each object region are then painted in a coarse-to-fine fashion through a sequence of localized attention windows $\mathcal{W}_t$. Finally, the brushstroke sequence is fed into a stroke-regularization procedure which removes brushstroke redundancies (and overlaps) to output the most efficient painting sequence for each test image.

4 Our Method

The *intelli-paint* framework (Fig. 2) is based on a two-stage hybrid optimization strategy which consists of two modules: *sequential-planner* (SP) and *stroke-regularizer* (SR). In the first stage, the *sequential-planner* (SP) learns to predict a coarse but more human-like initialization for the brushstroke sequence $\mathbf{s}_{init}$. The coarse brushstroke sequence initializations are then fed into a gradient descent based *stroke regularization* (SR) procedure, which removes redundant brushstroke patterns and refines the original brushstroke parameters to output the most efficient stroke decomposition $\mathbf{s}_{pred}$ for each test image. This two-stage process can be mathematically formulated as,

$$\mathbf{s}_{init} = SP\ (C_{init}, I_{target}) \quad \rightarrow \quad \mathbf{s}_{pred} = SR\ (\mathbf{s}_{init}, I_{target}), \tag{1}$$

where C_{init} signifies the blank canvas initialization and I is the target image. In the following sections, we discuss each of the above modules in full detail.

4.1 Sequential Planner

Reinforcement Learning Formulation. The *sequential-planner* (SP) is modelled as a deep reinforcement learning agent which learns a painting policy π predicting vectorized brushstroke parameters $\mathbf{a}_t$ (modeled as a Bézier curve [18,35]) from the current agent state s_t. The agent state s_t at any timestep t is modeled as the tuple $(C_t, I, t, \mathcal{G}_t, \mathcal{W}_t, \mathcal{S}_I, l)$, where C_t is the canvas state, I is the target image, $\mathcal{S}_I$ signifies the target-image saliency map, l is the current painting layer (refer Sec. 4.1.2), and $(\mathcal{G}_t, \mathcal{W}_t)$ represent the coarse and fine local attention windows for the painting agent respectively (refer Sec. 4.1.3).

The canvas state C_t is updated using a differentiable neural renderer module, which rasterizes the predicted brushstroke parameters $\mathbf{a}_t$ to output a brushstroke alpha map $S_\alpha(\mathbf{a}_t)$ and its colored rendering $S_{color}(\mathbf{a}_t)$. The canvas updates at each timestep t are then computed as follows,

$$C_{t+1} = C_t \odot (1 - S_\alpha(\mathbf{a}_t)) + S_{color}(\mathbf{a}_t). \tag{2}$$

We next discuss further details regarding the above formulation which allows our painting agent to generate output canvases while exhibiting some commonly found traits (*e.g.* layering, sequential attention) of the human painting process.

Progressive Layering: The human painting process is often progressive and multi-layered [33, 47]. That is, instead of painting everything on the canvas at once, humans often first paint a basic background layer before progressively adding each of the foreground objects on top of it (refer Fig. 1). However, such a strategy is hard to learn using previous works which directly minimize the pixel wise distance the generated canvas C_t and the target image I.

To this end, we propose a progressive layering strategy, which much like a human artist, allows the painted canvas to evolve in multiple successive layers. The objective of the painting agent in the first layer, is to paint a realistic background scene by trying to only focus on the non-salient (background) image areas. In doing so, the salient image regions are painted so as to maximize the efficiency of painting the background contents (*e.g.* salient region corresponding to a bird sitting on a tree would be painted while focusing on tree leaves and branches, as in Fig. 4). Once the background layer is drawn, the painting agent in the successive layer then proceeds to add different foreground objects in a decreasing order of saliency. An illustration of a two layer painting process is shown in Fig. 2. The painting agent first draws a realistic background scene (by focusing only on background image contents like ground, grass, *etc.*), before adding in the foreground objects (sheep) in the second layer.

In order to achieve this layering process, we first divide the overall painting episode into multiple layers as follows,

$$C_{out} = \sum_{l=0}^{L-1} \sum_{t=1}^{T/L} C_t^l \odot (1 - S_\alpha(\mathbf{a}_t^l)) + S_{color}(\mathbf{a}_t^l), \tag{3}$$

where $L = 2$ is the number of layers[3], T is the episode length, $C_0^{l=0}$ signifies an empty canvas, and $C_0^{l=1}$ is initialized as the canvas output $C_{T/L}^{l=0}$ from last layer. Given canvas state C_t, input image I and foreground saliency map $\mathcal{S}_I$, the layerwise painting style can then achieved be achieved by optimizing the following layered reward objective for each layer l,

$$r_t^{layer}(l) \;=\; D(I \odot \mathcal{M}_I(l), C_{t+1} \odot \mathcal{M}_I(l)) \;-\; D(I \odot \mathcal{M}_I(l), C_t \odot \mathcal{M}_I(l)), \quad (4)$$

where $D(I, C_t)$ is the joint conditional Wasserstein GAN [1] discriminator score for image I and canvas C_t [18], and the layered-mask $\mathcal{M}_I(l)$ is defined as,

$$\mathcal{M}_I(l) = 1 - \mathcal{S}_I \odot (1 - l). \quad (5)$$

Sequential Brushstroke Guidance: Human painters often exhibit a localized spatial attention span while focusing on distinct image areas [47]. This is in stark contrast with previous works which either compute stroke decomposition globally over the entire canvas or over a set of predefined grid regions [18,28,49]. To this end, we propose a sequential brushstroke guidance strategy, which allows the reinforcement learning agent to shift its attention between different image regions through a sequence of coarse-to-fine attention windows $\{\mathcal{W}_0, \mathcal{W}_1 \dots \mathcal{W}_T\}$. In particular, the computation of the localized attention window $\mathcal{W}_t$ at any timestep t during the painting process is done in the following broad steps,

– *Foreground object selection:* The RL agent first selects the in-focus foreground object by predicting coordinates of a coarse global attention window $\mathcal{G}_t$. Given an input image I with N foreground objects, we model $\mathcal{G}_t = x_t^{\mathcal{G}}, y_t^{\mathcal{G}}, w_t^{\mathcal{G}}, h_t^{\mathcal{G}}$ as a convex combination of each of in-image object bounding box detections $\mathcal{B}_i \in \mathbb{R}^4, i \in [1, N]$.

$$\mathcal{G}_t = \sum_{i=0}^{N} \alpha_i^t \; \mathcal{B}_i, \quad s.t. \;\; \forall t \;\; \sum_i \alpha_i^t = 1, \quad \alpha_i^t \geqslant 0. \quad (6)$$

where $\boldsymbol{\alpha}^t = \{\alpha_0^t, \; \dots \; \alpha_N^t\} \in \mathbb{R}^{N+1}$ are the spatial attention parameters predicted by the RL agent at timestep t. $\mathcal{B}_0$ represents an attention window over the entire canvas and is used to switch focus to background image areas.

– *Local attention window selection*: Within each object window $\mathcal{G}_t$, the agent further learns to sequentially shift its focus on different in-object features through a sequence of coarse-to-fine local attention windows $\mathcal{W}_t$. In particular, given the coarse object window coordinates $\mathcal{G}_t = x_t^{\mathcal{G}}, y_t^{\mathcal{G}}, w_t^{\mathcal{G}}, h_t^{\mathcal{G}}$, the coordinates $\mathcal{W}_t = x_t^{\mathcal{L}}, y_t^{\mathcal{L}}, w_t^{\mathcal{L}}, h_t^{\mathcal{L}}$ for the finer localized attention windows

[3] For simplicity, we primarily use $L = 2$ in the main paper. Further details on extending progressive layering to $L > 2$ are provided in Appendix A.2.

are computed in a Markovian fashion as,

$$x^{\mathcal{L}}_{t+1} = x^{\mathcal{G}}_{t+1} + (x^{\mathcal{L}}_{t} + \Delta x_t)\ w^{\mathcal{G}}_{t+1}, \tag{7}$$

$$y^{\mathcal{L}}_{t+1} = y^{\mathcal{G}}_{t+1} + (y^{\mathcal{L}}_{t} + \Delta y_t)\ h^{\mathcal{G}}_{t+1}, \tag{8}$$

$$w^{\mathcal{L}}_{t+1} = (max(1-\tilde{t}, w_{min}) + \Delta w_t)\ w^{\mathcal{G}}_{t+1}, \tag{9}$$

$$h^{\mathcal{L}}_{t+1} = (max(1-\tilde{t}, h_{min}) + \Delta h_t)\ h^{\mathcal{G}}_{t+1}, \tag{10}$$

where $\tilde{t} \in [0, 1]$ is the normalized episode timestep, (w_{min}, h_{min}) are the minimum attention window dimensions and $(\Delta\mathcal{W}_t = \Delta x_t, \Delta y_t, \Delta w_t, \Delta h_t) \in \mathbb{R}^4$ are successive Markovian [10] updates predicted by the RL agent. The above Markovian update formulation helps ensure spatial closeness of two consecutive local attention windows (Eq. 7,8), while facilitating a coarse-to-fine adjustment of the spatial attention window dimensions (Eq. 9,10).

– *Brushstroke parameter adjustment*: Finally, the coordinates of attention window $\mathcal{W}_t$ are used to modify the predicted brushstroke parameters $\mathbf{a}^l_t$ (modeled as Bézier curve), so as to constrain the painting agent to only draw within the local attention window. This procedure can be expressed as,

$$\mathbf{a}^l_t \leftarrow ParamAdjustment(\mathbf{a}^l_t, \mathcal{W}_t). \tag{11}$$

Please refer Appendix C for detailed implementation notes and instructions.

Human-Consistency Penalties: Human artists inherently try to focus on spatially close image areas and try to avoid unnecessary spatial oscillations when painting a given image [47]. In this regard, while the Markovian adjustment procedure introduced in Sec. 4.1.3 ensures the spatial closeness of two consecutive local attention windows $\mathcal{W}_t$, unnecessary movements may still arise due to oscillations between different coarse attention windows $\mathcal{G}_t$. To prevent learning such stroke decompositions we introduce the following spatial penalty,

$$r^{spatial}_t = -\|\mathcal{G}_{t+1} - \mathcal{G}_t\|_F, \tag{12}$$

where $\|.\|_F$ represents the Frobenius norm.

Similarly, human painting sequences are also characterized by the use of same (or similar) color patterns at consecutive timesteps [47]. Thus, in order to mimic this behaviour we propose the following color transition penalty r^{color}_t,

$$r^{color}_t = -\|(R, G, B)_{t+1} - (R, G, B)_t\|_F, \tag{13}$$

where $(R, G, B)_t$ represents the brushstroke color prediction at timestep t.

4.2 Brushstroke Regularization

Existing works on autonomous painting systems are often limited to using (an almost) fixed brush stroke budget irrespective of the complexity of the target

image. Experiments reveal that this not only reduces the efficiency of the generated painting sequence but also results in redundant (overlapping) brushstroke patterns (Fig. 4) which impart an unnatural painting style to the final agent.

To address this, we propose an inference-time brushstroke regularization strategy which refines and removes redundancies from the initial brushstroke sequence predictions $\mathbf{s}_{init}$ to output the most efficient stroke decomposition $\mathbf{s}_{pred}$ for each test image. To do this, we first associate each brushstroke with an importance vector $\beta_t^l \in [0, 1]$ by modifying the stroke rendering process as,

$$C_{out} = \sum_{l=0}^{L-1} \sum_{t=1}^{T/L} C_t^l \odot (1 - \beta_t^l \; S_\alpha(\mathbf{a}_t^l)) + \beta_t^l \; S_{color}(\mathbf{a}_t^l),$$

where $\beta_t^l = sign(x_t^l)$ and $x_t^l \sim \mathcal{N}(0, 10^{-3})$ is randomly initialized from a normal distribution. We then use gradient descent to optimize the following loss function over both brushstroke parameters $\mathbf{a}_t^l$ and importance vectors β_t^l (through x_t^l)

$$\mathcal{L}_{total}(\mathbf{a}_t^l, x_t^l) = \mathcal{L}_2(I, C_{out}) + \gamma \sum_{l=0}^{L-1} \sum_{t=1}^{T/L} \|\beta_t^l\|_1, \tag{14}$$

where the backpropagation gradients $\partial\beta_t^l / \partial x_t^l$ are computed as $\sigma(x_t^l)(1 - \sigma(x_t^l))$, $\sigma(.)$ is the sigmoid function and γ balances the weightage between brushstroke refinement and the need to use as few brushstrokes as possible.

5 Implementation Details

Neural renderer. In this paper, we primarily adopt the *PixelShuffleNet* architecture from Huang *et al.* [18] while designing the neural differentiable renderer. While our approach is not limited to a particular rendering mechanism, we find that as opposed to the opaque brushstroke models used in [28, 49], the use of a more naturally blending brushstroke representation from [18], allows our method to mimic the human painting style in a more closer fashion.

Layered training. The use of progressive layering module requires conditionally training the painting agent policy at each layer while initializing the canvas state with the output from the last layer. In order to save computation time during training, we train the successive layer policies in consecutive batches while using the canvas output from the last layer. Furthermore, we only use $L = 2$ layers at the training time. At inference time, the trained progressive layering policy can then be applied for $L > 2$ layers by appropriately modifying the target image saliency maps. Please refer Appendix A.2 for further details.

Saliency and bounding box predictions. A key component of the Intelli-Paint pipeline is the sequential brushstroke guidance strategy which relies on the computation of object saliency and bounding box predictions. In this work, we use a pretrained U-2-Net model [32] model in order to compute foreground saliency predictions. The bounding box predictions are then computed as the

(a) Target (b) Ours (c) Transformer (d) Optim (e) RL (f) Semantic-RL

Fig. 3. Qualitative method comparison w.r.t painting efficiency. Comparing final canvas outputs while using $\sim$ 300 brushstrokes for (b) Ours, (c) Paint Transformer [28], (d) Optim [49], (e) RL [18] and (f) Semantic-RL [35]. We observe that our approach results in more accurate depiction of the fine-grain features in the target image while using a low brushstroke count. Please zoom in for better comparison.

union over bounding box outputs from pretrained Yolo-v5 [20] and the overall bounding box for the saliency prediction output.

Overall training. The RL-based *sequential-planner* (SP) agent is trained using the model-based DDPG algorithm [18] with the following overall reward function for each layer l,

$$r_t^{overall}(l) = r_t^{layer}(l) + \mu\, r_t^{gbp} + \eta\, r_t^{spatial} + \lambda\, r_t^{color}, \tag{15}$$

where r_t^{gbp} is the guided-backpropagation based reward from [35]. The final RL agent is trained for a total of 5M iterations with a batch size of 128.

6 Comparison with State of the Art

In this section, we provide extensive qualitative and quantitative results comparing our method with recent state-of-the-art neural painting methods [18, 28, 35, 49]. First, in Sec. 6.1, we demonstrate the improved painting efficiency of our method in generating detailed paintings when using limited number of brushstrokes. Second, we show that our method leads to painting sequences with

increased resemblance with the human painting style (refer Sec. 6.2). Finally, in Sec. 6.3, we provide a discussion of some limitations of our approach in order to aid a more holistic understanding of the proposed method and future directions.

6.1 Painting Efficiency

As discussed in Sec. 3, the ability to learn an *efficient painting plan*, in order to paint detailed output canvases using as few brushstrokes as possible, is essential for most interactive and robotic painting applications [22,31,42]. In this section, we compare the painting efficiency of our approach with previous works while painting under a limited brushstroke budget.

Qualitative Comparison. Fig. 3 shows a qualitative comparison between the generated canvases using a low budget of 300 brushstrokes per canvas. Note that due to grid-wise formulation for Paint Transformer [28] and Optim [49], the corresponding results are reported after ∼360 and 330 brushstrokes respectively. We observe that our method results in more accurate depictions of target image (*e.g.* fine-grain features for car, hut, and birds in row 1-3 from Fig. 3) when using a limited number of brushstrokes. In contrast, previous methods often lack an intelligent mechanism for efficient brushstroke distribution across the canvas which leads to poor performance when using a limited brushstroke budget. Surprisingly, we also find that Paint Transformer [28] performs worse than previous methods like Optim [49] when using a small number of brushstrokes.

Quantitative Comparison. Table 1 shows quantitative results on the quality of the finally generated canvases while using $\sim$ 300 brushstrokes per canvas. The final results are reported in terms of both pixel wise l_2 distance $\mathcal{L}_{pixel}$ and perceptual similarity loss $\mathcal{L}_{pcpt}$ [21] between the final canvas and the target image. The quantitative values show that our method helps in significantly lowering the distance metrics between the painted canvas and the target image as compared to previous works. In particular, we note that for the CUB-Birds dataset [41], our approach leads to a reduction of 30.1%, 25.6% 24.9% and 38.2% in the $\mathcal{L}_{pixel}$ distance metric as compared to RL [18], Semantic-RL [35], Optim [49] and Paint Transformer [27], respectively.

6.2 Resemblance with Human Painting Style

Qualitative Comparison. We demonstrate the practical applicability of our method to actual human users by qualitatively comparing the painting sequences generated by our method with those drawn by actual human artists (refer Fig. 4). We observe that our method bears high resemblance with the human painting style in terms of both layerwise painting evolution and localized attention. In contrast, previous state-of-the-art methods often try to directly minimize the pixel-wise distance between painted canvas and the target image, thereby leading to intermediate canvas states which are less intelligible for a human user.

For instance, consider the first example from Fig. 4. Much like a human painter, our method first paints a realistic background representation (consisting of the sky, mountains, river and the ground) before drawing in the foreground

Fig. 4. Qualitative method comparison w.r.t resemblance with the human painting style. We compare different methods (b-f). All painting sequences are generated using a different brushstroke count (indicated in the boxes), so as to ensure similar pixel-wise reconstruction loss with the target image. The corresponding frames for each sequence are computed after $\sim$ 10%, 40%, 60% and 100% of the overall painting episode. We observe that our method offers higher resemblance with the human painting style (shown in column-a) as compared to previous works.

Method	Stanford Cars [24]		CUB-Birds [41]		Intelli-Paint Preference	
	$\mathcal{L}_{pixel}$	$\mathcal{L}_{pcpt}$	$\mathcal{L}_{pixel}$	$\mathcal{L}_{pcpt}$	Study A	Study B
RL [18]	78.06	0.54	72.93	0.56	87.17 %	83.11 %
Semantic [35]	79.98	0.55	68.46	0.55	84.46 %	69.09 %
Optim [49]	76.52	0.54	67.90	0.53	76.95 %	75.41 %
Transformer [28]	87.78	0.57	82.43	0.56	91.11 %	86.50 %
Ours	**56.92**	**0.44**	**50.94**	**0.45**	N/A	N/A

Table 1. Quantitative Evaluations. (Left) Method comparison w.r.t painting efficiency using a limited brushstroke budget. (Right) User-study results, showing % of painting samples for which human users prefer intelli-paint sequences over previous works.

car in a coarse-to-fine fashion. This results in a more human-like evolution of the painted canvas which can be easily relatable to actual human artists. In contrast, methods like Paint Transformer [28], Optim [49] and RL [18] directly make brushstrokes based on low-level image features (*e.g.* red brushstrokes for the car in row-1 and head of the bird in row-5). This leads to more bottom-up painting sequences which are different from the human style. Meanwhile, Semantic-RL [35] tries to paint both foreground and background regions in parallel, thereby lacking the semantic painting evolution exhibited by human users.

Quantitative Comparison. We also report quantitative results demonstrating the human-like resemblance of our approach as compared to previous works. To this end, we devise a human user study wherein each human participant is shown a series of paired painting sequences comparing our method with previous works. For each pair, the subject is then asked to select the painting sequence which best resembles the human painting style. The user study is performed in two different variations: **1)** User-Study A, where subjects are provided with a human painting sequence to act as reference in their decision-masking, and **2)** User-Study B, where participants are only shown a pair of artificial painting sequences (ours vs competing method) and are thus asked to make the decision based on their own subjective understanding of the human painting style. User-Study A was conducted across 10 different full-length painting sequences procured from real human artists, while User-Study B uses a set of randomly chosen 100 painting sequences from the CelebA [25] and CUB-Birds [41] datasets. A total of 50 unique Amazon Mechanical Turk subjects were used for both studies.

Results for both user-studies are shown in Table 1. User-Study A reveals that human subjects consider our painting sequences to be closer to those of a *particular human artist* (used as reference). However, as noted in Sec. 6.3, since each person has its own subjective understanding of what a human-like painting style constitutes, it does not answer the broader question on the relatability of these painting sequences from the context of a generic human user. User-study B tries to address this question. While we observe that the corresponding preference scores are lower than User-study A, it provides evidence that our approach is considered more relatable by a majority of human subjects.

6.3 Discussion and Limitations

In this section, we provide a discussion of some limitations of our method in order to facilitate a more holistic understanding of our approach.

Limited variation in painting style. We note that our method only mimics some commonly found traits (progressive layering, coarse to fine localized attention) of the human painting process, and, thus does not claim to be calibrated to the fine-grain variations in the painting styles of each human artist. Nevertheless, as demonstrated in Table 1, we find that our painting style is considered more relatable (over previous works) by majority of human users.

Human-like vs human-intelligible. We note that while our work provides a step towards improving the *human-intelligibility* of the painting sequences over previous works, it does not claim to be *truly human-like*. Several factors *e.g.* limited variation in painting style (discussed above), use of primitive brushstrokes (Bézier curves) *etc.* contribute to this limitation. This leaves much room for improvement in the development of truly human-like painting agents, which could motivate future work in this area (*e.g.* using advanced stroke representation [37]).

Reliance on pretrained image saliency models. Our method relies on the computation of image saliency masks for allowing a human-like evolution of the painted canvas. Thus limitations of the pretrained U2-Net [32] model become our limitations. Nevertheless, we note that failure to detect a particular salient object would simply lead to painting the corresponding region in the background layer, and thus does not affect the quality of the final canvas.

Training requirements. In order to learn a human-relatable style, our method requires self-supervised training on a dataset of *real* images. This is in contrast with Paint transformer [28] which performs self-training on an *artificial* dataset, and Optim [49] which does not require any training. That said, once trained we find that our method is able to generalize across a range of domains at inference time. For instance, we note that all results in Fig. 3, 4 were generated using an Intelli-paint model trained only on the CUB-Birds [41] dataset.

7 Conclusion

In this paper, we emphasize that the practical merits of an autonomous painting system should be evaluated not only by the quality of generated canvas but also by the interpretability of the corresponding painting sequence by actual human artists. To this end, we propose a novel *Intelli-Paint* pipeline, which uses progressive layering to allow for a more human-like evolution of the painted canvas. The painting agent focuses on different image areas through a sequence of coarse-to-fine localized attention windows and is able to paint detailed scenes while using a limited number of brushstrokes. Experiments reveal that in comparison with previous state-of-the-art methods, our approach not only shows improved painting efficiency but also exhibits a painting style which is much more relatable to actual human users. We hope our work opens new avenues for the further development of interactive and robotic painting applications in the real world.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. pp. 214–223. PMLR (2017) 7
2. Bangham, J.A., Gibson, S.E., Harvey, R.W.: The art of scale-space. In: BMVC. pp. 1–10. Citeseer (2003) 3
3. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017) 4
4. Chen, W., Hays, J.: Sketchygan: Towards diverse and realistic sketch to image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9416–9425 (2018) 4
5. Collomosse, J.P., Hall, P.M.: Genetic paint: A search for salient paintings. In: Workshops on Applications of Evolutionary Computation. pp. 437–447. Springer (2005) 3
6. Collomosse, J.P., Hall, P.M.: Salience-adaptive painterly rendering using genetic search. International Journal on Artificial Intelligence Tools **15**(04), 551–575 (2006) 3
7. Collomosse, J., Hall, P.: Painterly rendering using image salience. In: Proceedings 20th Eurographics UK Conference. pp. 122–128. IEEE (2002) 3
8. Frans, K., Cheng, C.Y.: Unsupervised image to sequence translation with canvas-drawer networks. arXiv preprint arXiv:1809.08340 (2018) 1
9. Frans, K., Soros, L., Witkowski, O.: Clipdraw: exploring text-to-drawing synthesis through language-image encoders. arXiv preprint arXiv:2106.14843 (2021) 1
10. Gagniuc, P.A.: Markov chains: from theory to implementation and experimentation. John Wiley & Sons (2017) 8
11. Ganin, Y., Kulkarni, T., Babuschkin, I., Eslami, S., Vinyals, O.: Synthesizing programs for images using reinforced adversarial learning. arXiv preprint arXiv:1804.01118 (2018) 1, 4
12. 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. pp. 2414–2423 (2016) 4
13. Graves, A.: Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850 (2013) 3
14. Ha, D., Eck, D.: A neural representation of sketch drawings. arXiv preprint arXiv:1704.03477 (2017) 1, 3
15. Haeberli, P.: Paint by numbers: Abstract image representations. In: Proceedings of the 17th annual conference on Computer graphics and interactive techniques. pp. 207–214 (1990) 3
16. Hertzmann, A.: Painterly rendering with curved brush strokes of multiple sizes. In: Proceedings of the 25th annual conference on Computer graphics and interactive techniques. pp. 453–460 (1998) 3
17. Hertzmann, A.: A survey of stroke-based rendering. Institute of Electrical and Electronics Engineers (2003) 3
18. Huang, Z., Heng, W., Zhou, S.: Learning to paint with model-based deep reinforcement learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8709–8718 (2019) 1, 4, 6, 7, 9, 10, 11, 13
19. Jia, B., Brandt, J., Mech, R., Kim, B., Manocha, D.: Lpaintb: Learning to paint from self-supervision. arXiv preprint arXiv:1906.06841 (2019) 4
20. Jocher, G., Stoken, A., Chaurasia, A., Borovec, J., NanoCode012, TaoXie, Kwon, Y., Michael, K., Changyu, L., Fang, J., V, A., Laughing, tkianai, yxNONG,

Skalski, P., Hogan, A., Nadar, J., imyhxy, Mammana, L., AlexWang1900, Fati, C., Montes, D., Hajek, J., Diaconu, L., Minh, M.T., Marc, albinxavi, fatih, oleg, wanghaoyang0106: ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support (Oct 2021). https://doi.org/10.5281/zenodo.5563715, https://doi.org/10.5281/zenodo.5563715 10

21. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016) 4, 11
22. Kite-Powell, J.: This ai robot will paint a canvas at sxsw 2021 (Mar 2021), https://www.forbes.com/sites/jenniferhicks/2021/03/10/this-ai-robot-will-paint-a-canvas-at-sxsw-2021/?sh=5b1f0d1ab449 4, 11
23. Kotovenko, D., Wright, M., Heimbrecht, A., Ommer, B.: Rethinking style transfer: From pixels to parameterized brushstrokes. arXiv preprint arXiv:2103.17185 (2021) 1
24. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13). Sydney, Australia (2013) 13
25. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 13
26. Litwinowicz, P.: Processing images and video for an impressionist effect. In: Proceedings of the 24th annual conference on Computer graphics and interactive techniques. pp. 407–414 (1997) 3
27. Liu, B., Gould, S., Koller, D.: Single image depth estimation from predicted semantic labels. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 1253–1260. IEEE (2010) 11
28. Liu, S., Lin, T., He, D., Li, F., Deng, R., Li, X., Ding, E., Wang, H.: Paint transformer: Feed forward neural painting with stroke prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6598–6607 (2021) 1, 2, 3, 4, 7, 9, 10, 11, 13, 14
29. Mellor, J.F., Park, E., Ganin, Y., Babuschkin, I., Kulkarni, T., Rosenbaum, D., Ballard, A., Weber, T., Vinyals, O., Eslami, S.: Unsupervised doodling and painting with improved spiral. arXiv preprint arXiv:1910.01007 (2019) 1, 4
30. Nakano, R.: Neural painters: A learned differentiable constraint for generating brushstroke paintings. arXiv preprint arXiv:1904.08410 (2019) 1
31. Nemire, B.: Ai painting robot (May 2017), https://developer.nvidia.com/blog/ai-painting-robot/ 4, 11
32. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O., Jagersand, M.: U2-net: Going deeper with nested u-structure for salient object detection. vol. 106, p. 107404 (2020) 9, 14
33. Reyner, N.: How to paint with layers - in acrylic & oil (Dec 2017), https://nancyreyner.com/2017/12/25/what-is-layering-for-painting/ 6
34. Shiraishi, M., Yamaguchi, Y.: An algorithm for automatic painterly rendering based on local source image approximation. In: Proceedings of the 1st international symposium on Non-photorealistic animation and rendering. pp. 53–58 (2000) 3
35. Singh, J., Zheng, L.: Combining semantic guidance and deep reinforcement learning for generating human level paintings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021) 1, 4, 6, 10, 11, 13

36. Singh, J., Zheng, L., Smith, C., Echevarria, J.: Paint2pix: Interactive painting based progressive image synthesis and editing. In: European conference on computer vision. Springer (2022) 4
37. Sochorová, Š., Jamriška, O.: Practical pigment mixing for digital painting. ACM Transactions on Graphics (TOG) **40**(6), 1–11 (2021) 14
38. Teece, D.: 3d painting for non-photorealistic rendering. In: ACM SIGGRAPH 98 Conference abstracts and applications. p. 248 (1998) 3
39. Treavett, S., Chen, M.: Statistical techniques for the automated synthesis of non-photorealistic images. In: Proc. 15th Eurographics UK Conference. pp. 201–210 (1997) 3
40. Turk, G., Banks, D.: Image-guided streamline placement. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. pp. 453–460 (1996) 3
41. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011) 11, 13, 14
42. Wikipedia contributors: Ai-da (robot) — Wikipedia, the free encyclopedia (2022), https://en.wikipedia.org/w/index.php?title=Ai-Da_(robot)&oldid=1070639724, [Online; accessed 7-March-2022] 4, 11
43. Xiang, X., Liu, D., Yang, X., Zhu, Y., Shen, X., Allebach, J.P.: Adversarial open domain adaptation for sketch-to-photo synthesis. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1434–1444 (2022) 4
44. Xie, N., Hachiya, H., Sugiyama, M.: Artist agent: A reinforcement learning approach to automatic stroke generation in oriental ink painting. IEICE TRANSACTIONS on Information and Systems **96**(5), 1134–1144 (2013) 1, 4
45. Yang, S., Wang, Z., Liu, J., Guo, Z.: Controllable sketch-to-image translation for robust face synthesis. IEEE Transactions on Image Processing **30**, 8797–8810 (2021) 4
46. Zeng, K., Zhao, M., Xiong, C., Zhu, S.C.: From image parsing to painterly rendering. ACM Trans. Graph. **29**(1), 2–1 (2009) 3
47. Zhao, A., Balakrishnan, G., Lewis, K.M., Durand, F., Guttag, J.V., Dalca, A.V.: Painting many pasts: Synthesizing time lapse videos of paintings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8435–8445 (2020) 2, 4, 6, 7, 8
48. Zheng, N., Jiang, Y., Huang, D.: Strokenet: A neural painting environment. In: International Conference on Learning Representations (2018) 1
49. Zou, Z., Shi, T., Qiu, S., Yuan, Y., Shi, Z.: Stylized neural painting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15689–15698 (2021) 1, 2, 3, 4, 7, 9, 10, 11, 13, 14