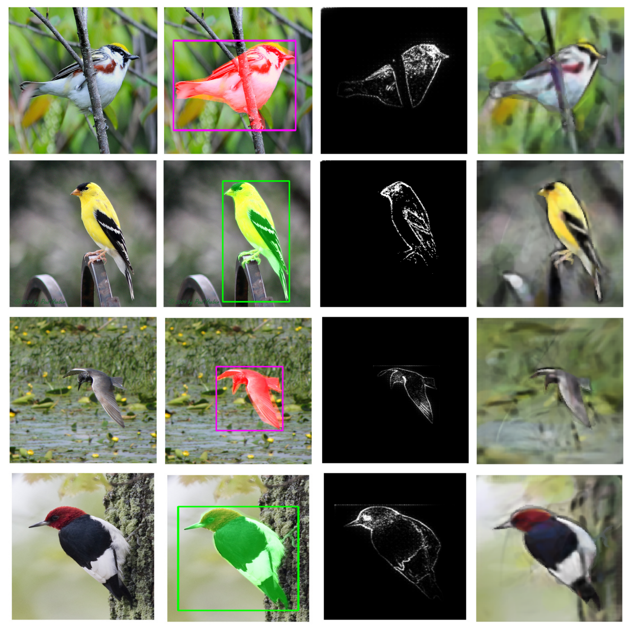
Combining Semantic Guidance and Deep Reinforcement Learning For Generating Human Level Paintings

Jaskirat Singh
Australian National University
jaskirat.singh@anu.edu.au

Liang Zheng
Australian National University
liang.zheng@anu.edu.au

Abstract

Generation of stroke-based non-photorealistic imagery, is an important problem in the computer vision community. As an endeavor in this direction, substantial recent research efforts have been focused on teaching machines "how to paint", in a manner similar to a human painter. However, the applicability of previous methods has been limited to datasets with little variation in position, scale and saliency of the foreground object. As a consequence, we find that these methods struggle to cover the granularity and diversity possessed by real world images. To this end, we propose a Semantic Guidance pipeline with 1) a bi-level painting procedure for learning the distinction between foreground and background brush strokes at training time. 2) We also introduce invariance to the position and scale of the foreground object through a neural alignment model, which combines object localization and spatial transformer networks in an end to end manner, to zoom into a particular semantic instance. 3) The distinguishing features of the in-focus object are then amplified by maximizing a novel guided backpropagation based focus reward. The proposed agent does not require any supervision on human stroke-data and successfully handles variations in foreground object attributes, thus, producing much higher quality canvases for the CUB-200 Birds [29] and Stanford Cars-196 [17] datasets. Finally, we demonstrate the further efficacy of our method on complex datasets with multiple foreground object instances by evaluating an extension of our method on the challenging Virtual-KITTI [2] dataset. Source code and models are available at https://github.com/1jsingh/semantic-guidance.



(a) (b) (c) (d)

Figure 1. **Semantic Guidance**. We propose a semantic guidance pipeline for the *"learning to paint"* problem. The reinforcement learning agent incorporates (b) object localization and semantic segmentation maps for the target image (a), to achieve enhanced foreground saliency (refer Fig. 3) in the final canvas (d). We also introduce expert guidance to amplify the focus on small but distinguishing features of the foreground objects (*e.g.* bird's eye), by proposing (c) a guided backpropagation based focus reward.

1. Introduction

Paintings form a key medium through which humans express their visual conception, creativity and thoughts. Being able to paint constitutes a vital skill in the human learning process and requires long-term planning to efficiently convey the picture within a limited number of brush strokes. Thus, the successful impartation of this challenging skill to machines, would not only have huge applications in computer graphics, but would also form a key component in the development of a general artificial intelligence system.

Recently, a lot of research [6, 11, 15, 22, 31, 34] is being targeted on teaching machines *"how to paint"*, in a manner similar to a human painter. A popular solution to this problem is to use reinforcement learning and model the painting episode as a Markov Decision Process (MDP). Given a target image, the agent learns to predict a sequence of brush strokes which when transferred on to a canvas, result in a painting which is semantically and visually similar to the input image. The reward function for the agent is usually learnt using a generative adversarial network (GAN) [9], which provides a measure of similarity between the final canvas and the original target image.

In this paper, we propose a *semantic guidance* pipeline which addresses the following three challenges faced by the current painting agents. **First**, the current methods [6, 15, 22] are limited to only datasets which depict a single dominant instance per image (*e.g.* cropped faces). Experimental results reveal that this leads to poor performance on varying the position, scale and saliency of the foreground object within the image. We address this limitation by adopting a *bi-level painting procedure*, which incorporates semantic segmentation into the painting process, to learn a distinction between brush stroke patterns for foreground and background image regions. Here, we utilize the intuition that the human painting process is deeply rooted in our semantic understanding of the image components. For instance, an accurate depiction of a bird sitting on a tree would depend highly on the agent's ability to recognize the bird and the tree as separate objects and hence use correspondingly different stroke patterns / plans.

Second, variation in position and scale of the foreground objects within the image, introduces high variance in the input distribution for the generative model. To this end, we propose a *neural alignment model*, which combines object localization and spatial transformer networks to learn an affine mapping between the overall image and the bounding box of the target object. The neural alignment model is end-to-end and preserves the differentiability requirement for our model-based reinforcement learning approach.

Third, accurate depiction of instances belonging to the same semantic class should require the painting agent to give special attention to different distinguishing features. For instance, while the shape of the beak may be a key feature for some birds, it may be of little consequence for other bird types. We thus propose a novel guided backpropagation based *focus reward* to increase the model's attention on these fine-grain features. The use of guided backpropagation also helps in amplifying the importance of small image regions, like a bird's eye which might be otherwise ignored by the reinforcement learning agent.

In summary, the main contributions of this paper are:

- We introduce a semantically guided bi-level painting process to develop a better distinction between foreground and background brush stroke patterns.
- We propose a neural alignment model, which combines object localization and spatial transformer networks in an end to end manner to zoom in on a particular foreground object in the image.
- We finally introduce expert guidance on the relative importance of distinguishing features of the in-focus object (*e.g.* tail, beak *etc.* for a bird) by proposing a novel guided backpropagation based focus reward.

2. Related Work

Stroke based rendering methods. Automatic generation of non-photorealistic imagery has been a problem of keen interest in the computer vision community. Stroke Based Rendering (SBR) is a popular approach in this regard, which focuses on recreating images by placing discrete elements such as paint strokes or stipples [14].

The positioning and selection of appropriate strokes is a key aspect of this approach [33]. Most traditional SBR algorithms address this task through either, greedy search at each step [13, 19], optimization over an energy function using heuristics [28], or require user interaction for supervising brush stroke positions [12, 27].

RNN-based methods. Recent deep learning based solutions adopt the use of recurrent neural networks for stroke decomposition. However, these methods like Sketch-RNN [11] for drawings and Graves *et al.* [10] for handwriting generation, require access to sequential stroke data, which limits their applicability for most real world datasets. StrokeNet [11] addresses this limitation by using a differentiable renderer, however it fails to generalize to color images.

Unsupervised stroke decomposition using RL. Recent methods [6, 15, 22, 31] use RL to learn an efficient stroke decomposition. The adoption of a trial and error approach alleviates the need for stroke supervision, as long as a reliable reward metric is available. SPIRAL [6], SPIRAL++ [22] and Huang *et al.* [15] adopt an adversarial training approach, wherein the reward function is modelled using the WGAN distance [1, 15]. Learning a differentiable renderer model has also been shown to improve the learning speed of the training process [5, 15, 23, 34].

The above methods generalize only for datasets (*e.g.* cropped, aligned faces from CelebA [21]), with limited variation in scale, position and saliency of the foreground object. We note that while Huang *et al.* [15] evaluate their approach on ImageNet [4], we find that competitive results are achieved only after using the division parameter at inference times. In this setting, the agent divides the overall image into a grid with 16 / 256 blocks, and then proceeds to paint each of them in parallel. We argue that such a division does not follow the constraints of the original problem formulation, in which the agent mimics the human painting process. Furthermore, such a division strategy increases the effective number of total strokes and tends towards a pixel-level image regression approach, with the generated images losing the desired artistic / non-photorealistic touch.

Semantic Divide and Conquer. Our work is in part also motivated by semantic division strategies from [20,30], which propose a division of the overall depth estimation task among the constituent semantic classes. However, to the best of our knowledge, our work is the first attempt on incorporating semantic division (with model-based RL) for the "learning to paint" problem.

3. Overview of the Painting Agent

Similar to Huang *et al*. [15], we adopt a model-based reinforcement learning approach for this problem. The painting episode is modelled as a Markov Decision Process (MDP) defined by state space $\mathcal{S}$, transition function $\mathcal{P}(s_{t+1}|s_t, a_t)$ and action space $\mathcal{A}$.

State space. The state $s_t \in \mathcal{S}$ at any time t is defined by the tuple $(C_t, I, \boldsymbol{\mathcal{S}}_I, \boldsymbol{\mathcal{G}}_I, t)$, where C_t is the canvas image at timestep t and I is the target image. $\boldsymbol{\mathcal{S}}_I, \boldsymbol{\mathcal{G}}_I$ represent the semantic instance probability map $\{\in [0,1]^{H \times W}\}$ and the guided backpropagation map for the target image.

Action space. The action a_t at each timestep, depicts the parameters of a quadratic Bézier curve, used to model the brush stroke. The stroke parameters form a 13 dimensional vector as follows,

$$a_t = (x_0, y_0, x_1, y_1, x_2, y_2, z_0, z_2, w_0, w_2, r, g, b), \quad (1)$$

where the first 10 parameters depict stroke position, shape and transparency, while the last 3 parameters (r, g, b) form the RGB representation for the stroke color.

Environment Model. The environment model / transition function $\mathcal{P}(s_{t+1}|s_t, a_t)$ is modelled through a neural renderer network $\boldsymbol{\Phi}$, which facilitates a differentiable mapping from the current canvas C_t and brush stroke parameters a_t to the updated canvas state C_{t+1}. For mathematical convenience alone, we define two distinct stroke map definitions $\boldsymbol{\Phi}, \boldsymbol{\Phi}^c$. $\boldsymbol{\Phi}(a_t)\{\in [0,1]^{H \times W}\}$ represents the stroke density map, whose value at any pixel provides a measure of transparency of the current stroke. $\boldsymbol{\Phi}^c(a_t)$ is the colored rendering of the original stroke density map $\boldsymbol{\Phi}(a_t)$ on an empty canvas.

Action Bundle. We adopt an action bundle approach which has been shown to be an efficient mechanism for enforcing higher emphasis on the planning process [15]. Thus, at each timestep the agent predicts the parameters for the next $K = 5$ brush strokes.

4. Introducing Semantic Guidance

In the following sections, we describe the complete pipeline for our semantic guidance model (refer Fig. 2). We first outline our approach for a two class (foreground, background) painting problem and then later demonstrate its extension to more complex image datasets with multiple foreground instances per image in Section 5.

4.1. The Bi-Level Painting Process

The human painting process is inherently multi-level, wherein the painter would focus on different semantic regions through distinct brush strokes. For instance, brush strokes aimed at painting the general image background would have a different distribution as compared to strokes depicting each of the foreground instances.

Motivated by this, we propose to use semantic segmentation to develop a distinction between the foreground and the background strokes. This distinction is achieved through a bi-level painting procedure which allocates a specialized reward for each stroke type. More specifically, we first modify the action bundle $\mathbf{a}_t$ to separately predict Bézier curve parameters for foreground and background strokes, *i.e.*

$$\mathbf{a}_t = \{\mathbf{a}_b, \mathbf{a}_f\}, \quad (2)$$

where $\mathbf{a}_f, \mathbf{a}_b$ represent the foreground and background stroke parameters, respectively. Next, given a neural renderer network $\boldsymbol{\Phi}$, target image I and semantic class probability map $\boldsymbol{\mathcal{S}}_I$, the canvas state C_t is updated in the following two stages,

$$C^b_{t+1} = [1 - \boldsymbol{\Phi}(\mathbf{a}_b)] \odot C_t + \boldsymbol{\Phi}^c(\mathbf{a}_b) \odot [1 - \boldsymbol{\mathcal{S}}_I], \quad (3)$$

$$C_{t+1} = [1 - \boldsymbol{\Phi}(\mathbf{a}_f)] \odot C^b_{t+1} + \boldsymbol{\Phi}^c(\mathbf{a}_f) \odot \boldsymbol{\mathcal{S}}_I, \quad (4)$$

where $\odot$ indicates element-wise multiplication and $\boldsymbol{\Phi}^c(a)$ is the colored rendering of the stroke density map $\boldsymbol{\Phi}(a)$.

The reward for each stroke type is then defined as,

$$r^b_t = D^{wgan}(I, C_{t+1}) - D^{wgan}(I, C_t), \quad (5)$$

$$\begin{aligned} r^f_t &= D^{wgan}(I \odot \boldsymbol{\mathcal{S}}_I, C_{t+1} \odot \boldsymbol{\mathcal{S}}_I) \\ &\quad - D^{wgan}(I \odot \boldsymbol{\mathcal{S}}_I, C_t \odot \boldsymbol{\mathcal{S}}_I), \end{aligned} \quad (6)$$

where r^f_t, r^b_t represent the foreground and background rewards, respectively, and $D^{wgan}(I, C_t)$ is the joint conditional discriminator score for image I and canvas C_t.

4.2. Neural Alignment Model

The accuracy of the foreground rewards computed using Eq. 6, depends highly on the ability of the discriminator to accurately capture the similarity between the foreground regions in target image I and the current canvas state C_t. However, the input to the discriminator of the WGAN model would have high variance, if the position and scale of the foreground object varies significantly amongst the input images. This high variance poses a direct challenge to the discriminator's performance while training on complex real world datasets. To this end, we propose a differentiable neural alignment model, which combines object localization and spatial transformer networks [16] to zoom into the foreground object, thereby providing a standardized input for the discriminator.

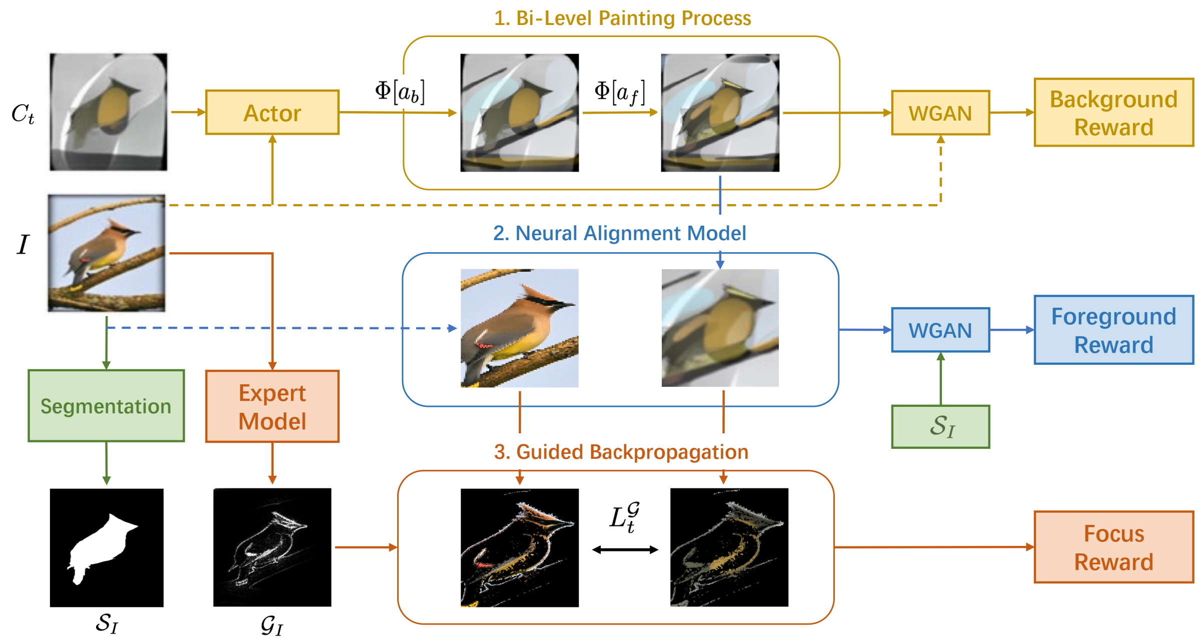


Figure 2. **Overview of Semantic Guidance Pipeline.** Our semantic guidance pipeline consists of three parts. **1)** The bi-level painting process (Section 4.1) develops a distinction between painting foreground and background brush strokes. **2)** The Neural Alignment Model (Section 4.2) provides a differentiable cropping of the foreground object regions for the target image and the updated canvas state. These cropped object images are then used to compute the foreground reward (refer Eq. 12). **3)** Finally, we use guided backpropagation maps from an expert model, to specifically boost the importance of distinguishing object features in the final canvas (Section 4.3).

First, we modify the segmentation model to predict both the foreground object mask $\boldsymbol{\mathcal{S}}_I$ and bounding box coordinates (x_b, y_b, w_b, h_b) of the foreground object in the target image. We then use a spatial transformer network $\boldsymbol{\Omega}$, which uses the predicted bounding box coordinates to compute an affine mapping, from the overall canvas image C_t to the zoomed foreground object image Z_t^C. Mathematically,

$$\boldsymbol{\mathcal{S}}_I, (x_b, y_b, w_b, h_b) = \boldsymbol{\Psi}(I), \tag{7}$$

$$Z_t^C = \boldsymbol{\Omega}(C_t, (x_b, y_b, w_b, h_b)), \tag{8}$$

$$Z^I = \boldsymbol{\Omega}(I, (x_b, y_b, w_b, h_b)), \tag{9}$$

$$Z^{\boldsymbol{\mathcal{S}}} = \boldsymbol{\Omega}(\boldsymbol{\mathcal{S}}_I, (x_b, y_b, w_b, h_b)), \tag{10}$$

where $\boldsymbol{\Psi}$ represents the foreground segmentation and localization network. The 3×2 affine matrix for the spatial transformer network $\boldsymbol{\Omega}$, given bounding box coordinates (x_b, y_b, w_b, h_b) and overall image size (H, W), is defined as,

$$A = \begin{bmatrix} W/w_b & 0 & -Wx_b/w_b \\ 0 & H/h_b & -Hy_b/h_b \end{bmatrix}^T. \tag{11}$$

The modified foreground reward (r_t^f) is then computed using the WGAN discriminator scores for the zoomed-in target and canvas images, as follows,

$$\begin{aligned} r_t^f &= D^{wgan}(Z^I \odot Z^{\boldsymbol{\mathcal{S}}}, Z_{t+1}^C \odot Z^{\boldsymbol{\mathcal{S}}}) \\ &\quad - D^{wgan}(Z^I \odot Z^{\boldsymbol{\mathcal{S}}}, Z_t^C \odot Z^{\boldsymbol{\mathcal{S}}}). \end{aligned} \tag{12}$$

4.3. Guided Backpropagation Based Focus Reward

The semantic importance of an image region is not necessarily proportional to the number of pixels covered by the corresponding region. While using WGAN loss provides some degree of abstraction as compared with the direct pixel-wise l_2 distance, we observe that a painting agent trained with a WGAN score based reward function, does not pay adequate attention to small but distinguishing object features. For instance, as shown in Fig. 3, for the CUB-200-2011 birds dataset, we see that while the baseline agent captures the global object features like shape and color, it either omits or insufficiently depicts important bird features like eyes, wing texture, color marks around the neck *etc*.

In order to address this limitation, we propose to incorporate a novel focus reward in conjuction with the global WGAN reward, to amplify the focus on the distinguishing features of each foreground instance. The focus reward uses guided back propagation maps from an expert task model (*e.g.* classification) to scale the relative importance of different image regions in the painting process. Guided backpropagation (GBP) has been shown to be an efficient mechanism for visualizing key image features [24, 25]. Thus by maximizing the focus reward, we encourage the painting agent to generate canvases with enhanced granularity at key feature locations.

Mathematically, given the normalized guided backpropagation map $\boldsymbol{\mathcal{G}}_I \{\in \{0,1\}^{H \times W}\}$ for the target image, object bounding box coordinates (x_b, y_b, w_b, h_b) and neural alignment model $\mathbf{\Omega}$, we first define the GBP distance $L_t^{\mathcal{G}}$ as,

$$Z^{\mathcal{G}_I} = \mathbf{\Omega}(\mathcal{G}_I, (x_b, y_b, w_b, h_b)), \tag{13}$$

$$L_t^{\mathcal{G}} = \frac{\left\| Z^{\mathcal{G}_I} \odot \left(Z^I - Z_t^C\right)\right\|_F^2}{\|Z^{\mathcal{G}_I}\|_F}, \tag{14}$$

where $\|.\|_F$ represents the Frobenius norm. Here we normalize the weighted difference between neurally aligned target and canvas images, using the total number of non-zero pixels in the guided backpropagation map. Thus, the scale of GBP distance $L_t^{\mathcal{G}}$ is invariant to extent of activations in the zoomed key-point importance map $Z^{\mathcal{G}_I}$.

The focus reward is then defined as the difference between GBP distances at successive timesteps,

$$r_t^{focus} = L_t^{\mathcal{G}} - L_{t+1}^{\mathcal{G}}. \tag{15}$$

5. Handling Multiple Foreground Instances

The semantic guidance pipeline discussed in Section 4, mainly handles images with a single foreground object instance per image. In this section, we show how the proposed approach can be used to "learn how to paint" on datasets depicting multiple foreground objects per image.

At training time, we maintain the bi-level painting procedure described in Section 4.1. The action bundle at each timestep describes the brush stroke parameters for the background and one of the foreground instances. The foreground instance for a particular painting episode is kept fixed and is selected with a probability proportional to the total number of pixels covered by that object.

At inference time however, the agent would need to pay attention to all of the foreground instances. Given N total foreground objects, the agent at any timestep t of the painting episode, would choose to predict brush stroke parameters for the foreground class with the highest l_2 difference in the corresponding areas on the canvas and the target image. Mathematically, the foreground instance (u) at each timestep t is selected as,

$$u = \arg\max_i \|\boldsymbol{\mathcal{S}}_i \odot (I - C_t)\|_F, \tag{16}$$

where $\boldsymbol{\mathcal{S}}_i$ represents the foreground segmentation map for the i^{th} object. We also note that the distinction between foreground and background strokes allows us to perform data augmentation with a specialized dataset to improve the quality of foreground data examples. Thus, in our experiments, we augment the Virtual KITTI dataset with images from Stanford Cars-196 in ratio of 0.8:0.2 while training.

6. Experiments

6.1. Datasets

We use the CUB-200-2011 Birds [29] and Stanford Cars-196 [17] dataset for performing qualitative evaluation of our method. The above datasets mainly feature one foreground instance per image and hence can be trained using the bi-level semantic guidance pipeline described in Section 4. We also use the high-fidelity Virtual-KITTI [2] dataset to demonstrate the extension of the proposed method to multiple foreground instances per image.

CUB-200-2011 Birds [29] is a large-scale birds dataset frequently used for benchmarking fine-grain classification models. It consists of 200 bird species with annotations available for class, foreground mask and bounding box of the bird. The dataset features high variation in object background as well as scale, position and the relative saliency of the foreground bird with respect to its immediate surroundings. These properties make it a challenging benchmark for the "learning to paint" problem.

Stanford Cars-196 [17] is another dataset used for testing fine-grain classification. It consists of 16185 total images depicting cars belonging to 196 distinct categories and having varying 3D orientation. The dataset only provides object category and bounding box annotations. We compute the foreground car masks using the pretrained DeepLabV3-Resnet101 network [3].

Virtual KITTI [2] is a high fidelity dataset containing photo-realistic renderings of urban environments from 5 distinct scene backgrounds. Each scene contains images depicting variation in camera location, weather, time of day and density / location of foreground objects. The high variability of these image attributes, makes it a very challenging dataset for training the painting agent. Nevertheless, we demonstrate that our method helps in improving the semantic quality of the generated canvases despite these obstacles.

6.2. Training Details

Neural Renderer. We closely follow the architecture from Huang *et al.* [15], while designing the differentiable neural renderer $\mathbf{\Phi}$. Given a batch of random brush stroke parameters a_t, the network output $\mathbf{\Phi}[a_t]$ is trained to mimic the rendering of the corresponding Bézier curve on an empty canvas. The training labels are generated using an automated graphics module and the renderer is trained for 4×10^5 iterations with a batch size of 64.

Learning foreground mask and bounding box. A key component of the semantic guidance pipeline is foreground segmentation and bounding box prediction. We use a fully convolutional network, with separate heads to predict a per-pixel foreground probability map and the coordinates of the bounding box. The foreground mask prediction is trained with the standard cross-entropy loss L_{fg}, while the bounding box coordinates are learned using Smooth L1 [8] regression loss L_{bbox}.

Expert model for Guided Backpropagation. We use the pretrained fine-grain classification NTS-Net model [32] as the expert network used for generating guided backpropagation maps on the CUB-200-2011 birds dataset. Note that we use NTS-Net due the easy accessibility of the pretrained model. We expect that using a more state of the art model like [7] would lead to better results with the focus reward.

The expert model for the Standford Cars-196 dataset is trained in conjunction with the reinforcement learning agent, with an EfficientNet-B0 [26] backbone network. The EfficientNet architecture allows us to limit the total number of network parameters while respecting the memory constraints for a NVIDIA GTX 2080 Ti. The expert model is trained for a total of 200 epochs with a batch size of 64. EfficientNet-B7 model pretrained on ImageNet [4] dataset, is used as the expert for the Virtual KITTI dataset.

Overall Training. The reinforcement learning agent follows an actor-critic architecture. The actor predicts the policy function $\pi(a|s)$, while the critic computes the value function $V(s)$. The agent is trained using model-based DDPG [18] with the following policy and value loss,

$$L_{actor} = -\mathbf{E}_{s_t,a_t}\left[r(s_t,a_t) + \gamma V(s_{t+1})\right], \quad (17)$$

$$L_{critic} = \mathbf{E}_{s_t,a_t}\left[\left(r(s_t,a_t) + \gamma V(s_{t+1}) - V(s_t)\right)^2\right] \quad (18)$$

where γ is the discount factor and the final reward function $r(s_t,a_t)$ is computed as the weighted sum of the foreground, background and focus rewards,

$$r(s_t,a_t) = r_t^b + \eta\, r_t^f + \nu r_t^{focus}, \quad (19)$$

where η, ν are hyperparameters. A hyper-parameter selection of $\{\eta = 2, \nu = 10\}$ was seen to give competitive results for our experiments. The model-based RL agent is trained for a total of 2M iterations with a batch size of 96.

6.3. Results

We compare our method with the baseline "learning to paint" pipeline from Huang *et al*. [15] which uses an action bundle containing 5 consecutive brush strokes. In order to provide a fair comparison, we use the same overall bundle size but divide it among foreground and background strokes in the ratio of 3:2. That is, the agent at each timestep predicts 3 foreground and 2 background brush strokes.

Improved foreground saliency. Fig. 3 shows the results[1] for the CUB-200 Birds and Stanford-Cars196 dataset. We clearly see that our method leads to increased saliency of foreground objects, especially when the target object is partly camouflaged by its immediate surroundings (refer Fig. 3a, row-4 and Fig. 3b, row-3). This increased contrast between foreground and background perception, results directly from our semantically guided bi-level painting process and the neural alignment model.

Enhanced feature granularity. We also observe that canvases generated using our method show improved focus on key object features as compared to the baseline. For instance, the red head-feather, which is an important feature of pileated woodpecker (refer Fig. 3a: row-1), is practically ignored by the baseline agent due to its small size. The proposed guided backpropagation based focus reward, helps in amplifying the importance of this key feature in the overall reward function. Similarly, our method also leads to improved depiction of wing patterns and claws in (Fig. 3a: row-2), the small eye region, feather marks in (Fig. 3a: row-3) and car headlights, wheel patterns in (Fig. 3b: row-1,2).

Multiple foreground instances. We use the Virtual-KITTI dataset and the extended training procedure outlined in Section 5, to demonstrate the applicability of our method on images with multiple foreground instances. Note that due to computational limits and the nature of ground-truth data, we stick to vehicular foreground classes like cars, vans, buses *etc*, for our experiments. Results are shown in Fig. 4. We observe that due to the dominant nature of image backgrounds in this dataset, the baseline agent fails to accurately capture the presence / color spectrum of the foreground vehicles. In contrast, our bi-level painting procedure learns a distinction between foreground and background strokes in the training process itself, and thus provides a much better balance between foreground and background depiction for the target image.

7. Analysis

7.1. Ablation Study: Isolating Impact of Focus Loss

In this section, we design a control experiment in order to isolate the impact of focus reward proposed in Section 4.3. To this end, we construct a modified birds dataset from CUB-200-2011 dataset. We do this by first setting the background image pixels to zero, which alleviates the need for the bi-level painting procedure. We next eliminate the need for the neural alignment model by cropping the bounding box for each bird. The resulting dataset is then used to train the baseline [15], and a modified semantic guidance pipeline trained only using a weighted combination of the WGAN reward [15] and the focus reward r_t^{focus},

$$r(s_t,a_t) = r_t^{wgan} + \kappa\, r_t^{focus}, \quad (20)$$

where $\kappa = 0$ represents baseline model without the focus loss. We then analyse the effect on the resulting canvas as the weightage κ of the focus reward is increased. All models are trained for 1M iterations with a batch size of 96.

Fig. 5 describes the modified training results. We clearly see that while the baseline [15] trained with wgan reward captures the overall bird shape and color, it fails to accurately pay attention to finer bird features like texture of the wings (row 1,3,4), density of eyes (row 3,4) and sharp color contrast (red regions near the face for row 1,2). We also observe that the granularity of the above discussed features

[1]Please refer supp. material for further quantitative results.

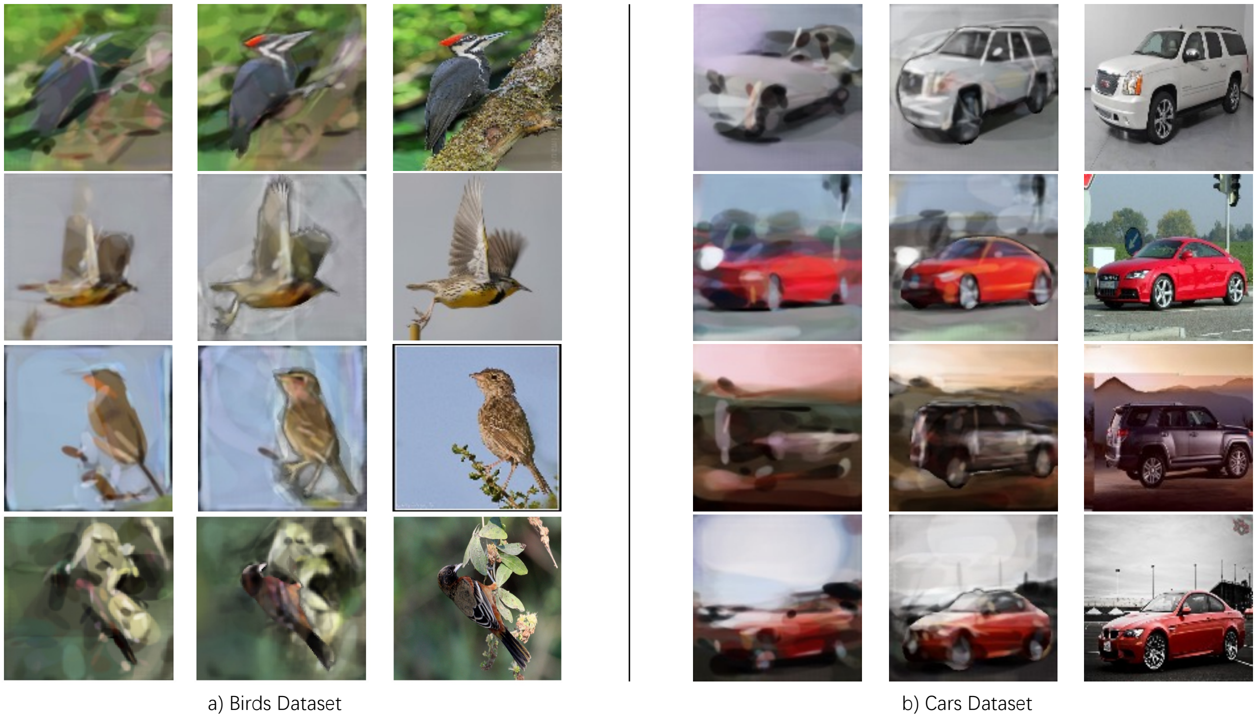


Figure 3. **Results on CUB-200 Birds and Stanford-Cars196 Datasets.** Left: Huang *et al*. [15], Middle: Canvas generated using Semantic Guidance pipeline (Ours), Right: the original target image. We clearly see that our method results in enhanced foreground saliency and achieves better granularity of key object features.

Figure 4. **Results on Virtual KITTI.** Left: Baseline [15], Middle: Canvas generated using Semantic Guidance pipeline (Ours), Right: target image. By developing a distinction between foreground and background strokes, our method better captures the color / saliency of visually small foreground vehicles.

in the painted canvas, improves as the weightage κ of the focus reward is increased.

7.2. Analysing Effect of Semantic Guidance on Painting Sequence

Recall that the main goal of the "learning to paint" problem, is to make the machine paint in a *manner similar to a human painter*. Thus, the performance of a painting agent should be measured, not only by the resemblance between the final canvas and the target image, but also by the similarity of the corresponding painting sequence with that of a human painter. In this section, we demonstrate that unlike previous methods, semantic guidance helps the reinforce-

Figure 5. **Ablation results for focus reward.** (Column 1-3): From left to right, the painted canvases for $\kappa = 0, 5, 10$ respectively, where $\kappa = 0$ represents the baseline [15]. (Column-4): the target image from modified birds dataset (refer Sec. 7.1). We see a clear increase in the amount of finer feature details like wing texture, density of eyes *etc*, as the weightage of focus loss is increased.

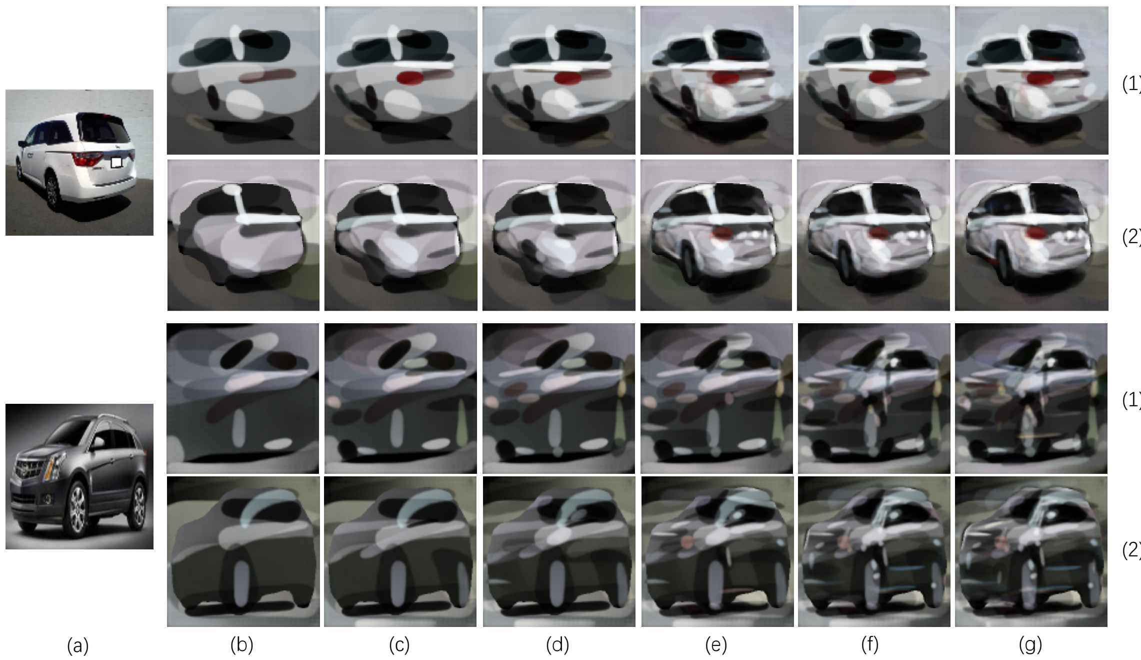


Figure 6. **Effect of Semantic Guidance on Painting Sequence.** (1) Baseline [15], (2) Semantic Guidance (Ours). For each target image in (a), (b-g) represent the canvas state after 10, 20, 30, 50, 100, 200 brush strokes respectively. We observe that there is huge difference between the painting styles of the two agents. In contrast to the baseline agent (which follows a bottom-up approach), the top-down painting style of our method offers better resemblance with a human painter.

ment learning agent adopt a painting trajectory that is highly similar to the human painting process.

In order to do a fair comparison of agent trajectories between our method and the baseline [15], we select test images from the Stanford Cars-196 dataset, such that the final canvases from both methods are equally similar to the target image. That is, the l_2[2] distance between the final canvas and the target image is similar for both methods.

Results are shown in Fig. 6. We can immediately observe a stark difference between the painting styles of the two agents. The standard agent displays bottom-up image understanding, and proceeds to first paint visually distinct car edges / parts like windows, red tail light, black region near the bottom of the car *etc*. In contrast, the semantically guided agent follows a more human-like top-down approach, wherein it first begins with a rough structural outline for the car and only then focuses on other structurally non-relevant parts. For instance, in the first example from Fig. 6, the semantically guided agent adds color to the tail-light only after finishing painting the overall structure of the car. On the other hand, the red brush stroke for the tail-light region is painted quite early by the baseline agent, even before the overall car structure begins to emerge on the canvas.

[2]We note that, in general l_2 distance may not be a reliable measure of semantic similarity between two images. As shown in Fig. 6, two canvases can be qualitatively quite different while having similar l_2 distance with the target image.

We thus note that the semantically guided agent resembles the human painting style on two broad levels. *1) On the canvas level*, the bi-level procedure allows the painting agent to learn different stroke patterns for semantically distinct image regions (as is done by humans). *2) On the object level*, expert guidance and specialization of the foreground strokes to a specific object class (*e.g.* cars) leads to a top-down painting sequence. That is, the model first pays attention to high-level structural features which are shared by several instances of the foreground object class, and only then focuses on finer instance-specific lower-level features.

8. Conclusion

In this paper, we propose a semantic guidance pipeline for the "learning to paint" problem. Our method incorporates semantic segmentation to propose a bi-level painting process, which helps in learning a distinction between foreground and background brush stroke patterns. We also introduce a guided backpropagation based focus reward, to increase the granularity and importance of small but distinguishing object features in the final canvas. The resulting agent successfully handles variations in position, scale and saliency of foreground objects, and develops a top-down painting style which closely resembles a human painter.

Acknowledgements. This work was supported by ARC Discovery Early Career Researcher Award (DE200101283) and ARC Discovery Project (DP210102801).

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 2

[2] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020. 1, 5

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 5

[4] 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. 2, 6

[5] Kevin Frans and Chin-Yi Cheng. Unsupervised image to sequence translation with canvas-drawer networks. *arXiv preprint arXiv:1809.08340*, 2018. 2

[6] Yaroslav Ganin, Tejas Kulkarni, Igor Babuschkin, SM Eslami, and Oriol Vinyals. Synthesizing programs for images using reinforced adversarial learning. *arXiv preprint arXiv:1804.01118*, 2018. 1, 2

[7] Weifeng Ge, Xiangru Lin, and Yizhou Yu. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3034–3043, 2019. 6

[8] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 5

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[10] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013. 2

[11] David Ha and Douglas Eck. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017. 1, 2

[12] Paul Haeberli. Paint by numbers: Abstract image representations. In *Proceedings of the 17th annual conference on Computer graphics and interactive techniques*, pages 207–214, 1990. 2

[13] Aaron Hertzmann. Painterly rendering with curved brush strokes of multiple sizes. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 453–460, 1998. 2

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

[15] Zhewei Huang, Wen Heng, and Shuchang Zhou. Learning to paint with model-based deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8709–8718, 2019. 1, 2, 3, 5, 6, 7, 8

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 3

[17] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 1, 5

[18] 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. 6

[19] Peter Litwinowicz. Processing images and video for an impressionist effect. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 407–414, 1997. 2

[20] Beyang Liu, Stephen Gould, and Daphne Koller. Single image depth estimation from predicted semantic labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1253–1260. IEEE, 2010. 3

[21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 2

[22] John FJ Mellor, Eunbyung Park, Yaroslav Ganin, Igor Babuschkin, Tejas Kulkarni, Dan Rosenbaum, Andy Ballard, Theophane Weber, Oriol Vinyals, and SM Eslami. Unsupervised doodling and painting with improved spiral. *arXiv preprint arXiv:1910.01007*, 2019. 1, 2

[23] Reiichiro Nakano. Neural painters: A learned differentiable constraint for generating brushstroke paintings. *arXiv preprint arXiv:1904.08410*, 2019. 2

[24] Weili Nie, Yang Zhang, and Ankit Patel. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. *arXiv preprint arXiv:1805.07039*, 2018. 4

[25] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 4

[26] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019. 6

[27] Daniel Teece. 3d painting for non-photorealistic rendering. In *ACM SIGGRAPH 98 Conference abstracts and applications*, page 248, 1998. 2

[28] Greg Turk and David Banks. Image-guided streamline placement. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 453–460, 1996. 2

[29] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 1, 5

[30] Lijun Wang, Jianming Zhang, Oliver Wang, Zhe Lin, and Huchuan Lu. Sdc-depth: Semantic divide-and-conquer network for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 541–550, 2020. 3

[31] Ning Xie, Hirotaka Hachiya, and Masashi Sugiyama. 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, 2

[32] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 420–435, 2018. 5

[33] Kun Zeng, Mingtian Zhao, Caiming Xiong, and Song Chun Zhu. From image parsing to painterly rendering. *ACM Trans. Graph.*, 29(1):2–1, 2009. 2

[34] Ningyuan Zheng, Yifan Jiang, and Dingjiang Huang. Strokenet: A neural painting environment. In *International Conference on Learning Representations*, 2018. 1, 2