Supplementary Material: 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

Method	Accuracy	IoU
Huang *et al*. [3]	45.41	27.21
Semantic Guidance (Ours)	**69.26**	**48.15**

Table 1. **Semantic Similarity Results on CUB-200 Birds.** The semantic segmentation maps (refer Appendix A.1) for the canvases generated using our method, result in much better segmentation accuracy and Intersection over Union (IoU) scores.

A. Quantitive Results

A.1. Measuring Semantic Similarity.

The inadequacy of the frequently used pixel-wise $l2$ distance [2, 3] in capturing semantic similarity, poses a major challenge in performing a quantitative evaluation of our method. In order to address this, we present a novel approach to quantitatively evaluate the semantic similarity between the generated canvases and the target image. To this end, we use a pretrained DeeplabV3-ResNet101 model [1] to compute the semantic segmentation maps for the final painted canvases for both Huang *et al*. [3] and the Semantic Guidance (Ours) approach. The detected segmentation maps for both methods are then compared with the ground truth foreground masks for the target image.

Results are shown in Fig. 1. We clearly see that our method learns to paint canvases with semantic segmentation maps having high resemblance with the ground truth foreground masks for the target image. In contrast, the canvases generated using the baseline [3] show low foreground saliency. This sometimes results in the pretrained segmentation model [1] even failing to detect the presence of the foreground object. Note that the semantic guidance pipeline does not directly train the RL agent to mimic the segmentation maps of the original image.

We also provide a more quantitative evaluation of the quality of detected semantic segmentation maps for both methods in Table 2. The accuracy scores are reported on the test set images and represent the percentage of foreground pixels which are correctly detected in the segmentation map of a given canvas. We observe that our method leads to huge improvements in the semantic segmentation accuracy and IoU values for the painted canvases.

The above qualitative and quantitative results conclusively demonstrate that the semantic guidance pipeline leads to huge gains ($\sim 25\%$) in preserving the underlying semantics of a given scene.

A.2. Enhanced Foreground Resemblance

Method	Foreground L2 Distance
Huang *et al*. [3]	8.43
Semantic Guidance (Ours)	**7.81**

Table 2. **Foreground Resemblance Results on CUB-200 Birds.** Our approach leads to a lower average L2 distance between the foreground regions of the target image and the generated canvas.

B. Implementation of Neural Alignment Model

The neural alignment model is implemented by replacing the localization net of a standard spatial transformer network [4] with the bounding box prediction network. We also note that the 3×2 affine matrix defined in Eq. 11 of the main paper, represents the ideal affine mapping operation from input to output image coordinates. However, the affine matrix used for practical implementations may vary based on the conventions of the used deep learning framework. For our implementation (in pytorch), we compute the affine matrix for the spatial transformer network as follows,

$$\tilde{A} = \begin{bmatrix} \tilde{w}_b & 0 & 2\tilde{x}_b + \tilde{w}_b - 1 \\ 0 & \tilde{h}_b & 2\tilde{y}_b + \tilde{h}_b - 1 \end{bmatrix}^T, \tag{1}$$

where $(\tilde{x}_b, \tilde{y}_b, \tilde{w}_b, \tilde{h}_b)$ are the normalized bounding box coordinates of the foreground object.

C. Note on Over-painting Phenomenon

We note that while the proposed semantic guidance pipeline results in huge improvements in enhancing fore-

Figure 1. **Analysing Semantic Similarity**.(a) Huang *et al*. [3], (b) Semantic Guidance (Ours), (c) the target image. The bottom row for each example represents the semantic segmentation maps for the images shown in the top row. We clearly see that the canvases painted using our method generate semantic segmentation maps which are much closer to the ground truth foreground segmentation masks. We also note that, for target images with low foreground background contrast, the segmentation maps for baseline canvases (a) fail to even indicate the presence of the foreground object.

ground object saliency and increasing the granularity of the painted image, we do observe minor background artifacts for images with plain backgrounds. This occurs because as part of the bilevel painting procedure, both foreground and background brush strokes are working simultaneously in an action bundle. Thus for images with high contrast in complexities of foreground and background, the background strokes are forced to *overpaint* while the foreground strokes draw the in-focus object. This *overpainting* phenomenon was seen to cause minor artifacts in plain image backgrounds as can be seen in (Fig. 3a; row-3) of the main paper. The above mentioned artifacts can be reduced by adaptively balancing the number of foreground / background strokes in an action bundle, based on the WGAN distances for the foreground and background image regions. However, the same is out of scope of the current paper and we thus leave it here as a possible future research directive.

References

[1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1

[2] 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

[3] 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

[4] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 1