

Supplemental Material for “Example-based Authoring of Procedural Modeling Programs with Structural and Continuous Variability”

Daniel Ritchie¹ Sarah Jobalia² Anna Thomas²

¹Brown University

²Stanford University

1. Perceptual Study

To evaluate whether samples generated by learned programs exhibit visual quality and style consistent with the examples, we conducted a perceptual study on Amazon Mechanical Turk (AMT). Our study design extends that of prior work on evaluating automatic image colorization methods [ZIE]. Study participants were given a series of forced-choice image comparison tasks. Each task displayed a collection of 9 images and an additional pair of images. The 9 images depicted examples from one of our 5 example sets. One image in the pair depicted the remaining 10th example from the set, and the other depicted a random sample generated by a program learned from the example set. The order of images in the collection and within the pair was randomized. The task instructed participants to choose the image from the pair which was a better addition to the collection, in that it ‘most belonged’ to the collection. Each participant performed 11 comparison tasks for each of our 5 example sets (55 tasks total). Each task held out a different example for comparison against a randomly-generated sample. One out of each 11 tasks was a ‘vigilance test’: a comparison with an obvious wrong answer (specifically, one image in the pair was drawn from a different example set). Participants were shown a blank screen for 2 seconds between each task. We collected responses from 20 AMT Master workers (workers AMT identifies as consistently high-performing).

We discarded responses from workers who did not achieve 100% accuracy on the vigilance tests (6 workers), leaving a total of 140 comparisons per example set. We hypothesized that participants would show no significant preference between held-out examples and samples generated by our learned programs. Table 1 shows, for each example set, the percentage of time that workers selected the hand-created example from the pair. None of these percentages are different from chance (i.e. 50%) at a 95% confidence level (estimated by bootstrap [ET86]). This result suggests that samples generated by a learned program are, on average, of comparable style and visual quality as the examples from which the program was learned.

References

- [ET86] EFRON B., TIBSHIRANI R.: Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 1, 1 (02 1986), 54–75. 1
- [ZIE] ZHANG R., ISOLA P., EFROS A. A.: Colorful Image Colorization. In *ECCV* 2016. 1

Example set	% Examples Selected
<i>StarSparrow</i>	46.42 ± 4.18
<i>StriderOx</i>	55.75 ± 4.20
<i>Castle</i>	42.81 ± 4.22
<i>SpaceStation</i>	43.57 ± 4.21
<i>Ornament</i>	48.54 ± 4.24

Table 1: Percentage of examples chosen in a Mechanical Turk forced-choice comparison between hand-created examples and program-generated outputs (mean and standard error reported, estimated by bootstrap [ET86]). None of these results are significantly different from chance (50%), indicating that participants had no preference between examples and program outputs.