Appendix

This appendix contains tables for evaluation on the edit images (i.e. each of the 6 training images). We include these evaluations to check if the edited models perform better than the original model on the image used for the corresponding edit. The evaluations on the unseen images are included in the main paper. In general, we see that the performance of each mapping on the edit images loosely corresponds to its performance on the unseen images.

Image	Precision		Recall		IoU	
	Original	Edited	Original	Edited	Original	Edited
A	0.82 ± 0.19	$\mathbf{0.88 \pm 0.19}$	$\mathbf{0.78 \pm 0.24}$	0.63 ± 0.28	$\mathbf{0.66 \pm 0.25}$	0.57 ± 0.29
B	0.53 ± 0.28	$\mathbf{0.54 \pm 0.27}$	$\mathbf{0.78 \pm 0.24}$	0.67 ± 0.24	$\mathbf{0.66 \pm 0.27}$	0.61 ± 0.26
C	0.83 ± 0.20	$\mathbf{0.95 \pm 0.09}$	$\mathbf{0.77 \pm 0.24}$	0.70 ± 0.23	0.64 ± 0.30	$\mathbf{0.67 \pm 0.24}$
D	0.79 ± 0.23	$\mathbf{0.91 \pm 0.13}$	$\mathbf{0.78 \pm 0.28}$	0.74 ± 0.23	0.64 ± 0.30	$\mathbf{0.66 \pm 0.26}$
E	0.85 ± 0.15	$\mathbf{0.92 \pm 0.10}$	$\mathbf{0.70 \pm 0.26}$	0.63 ± 0.28	0.55 ± 0.32	$\mathbf{0.58 \pm 0.29}$
F	0.83 ± 0.16	$\mathbf{0.90 \pm 0.10}$	$\mathbf{0.82 \pm 0.22}$	0.72 ± 0.21	$\mathbf{0.67 \pm 0.26}$	0.64 ± 0.23

Table 9: **Mapping non-archaeocyathids to red mud: Performance on edit image.** This table contains the mean instance-level metrics $\pm$ one standard deviation run on each image used for mapping non-archaeocyathids to red mud. For example, the first row contains the mean instance-level precision, recall, and IoU across the identified archaeocyathids in image A alone. All the precisions increase, and some of the IoU scores increase as well. The recall scores decrease.

Image	Precision		Recall		IoU	
	Original	Edited	Original	Edited	Original	Edited
A	$\mathbf{0.82 \pm 0.19}$	$\mathbf{0.82 \pm 0.18}$	0.78 ± 0.24	$\mathbf{0.81 \pm 0.22}$	0.66 ± 0.25	$\mathbf{0.67 \pm 0.26}$
B	$\mathbf{0.53 \pm 0.28}$	$\mathbf{0.53 \pm 0.25}$	$\mathbf{0.78 \pm 0.24}$	0.77 ± 0.24	$\mathbf{0.66 \pm 0.27}$	0.62 ± 0.29
C	0.83 ± 0.20	$\mathbf{0.84 \pm 0.15}$	$\mathbf{0.77 \pm 0.24}$	0.75 ± 0.30	$\mathbf{0.64 \pm 0.30}$	0.60 ± 0.34
D	0.79 ± 0.23	$\mathbf{0.85 \pm 0.16}$	0.78 ± 0.28	$\mathbf{0.82 \pm 0.26}$	0.64 ± 0.30	$\mathbf{0.66 \pm 0.28}$
E	$\mathbf{0.85 \pm 0.15}$	0.80 ± 0.16	0.70 ± 0.26	$\mathbf{0.73 \pm 0.30}$	$\mathbf{0.55 \pm 0.32}$	0.48 ± 0.35
F	$\mathbf{0.83 \pm 0.16}$	0.82 ± 0.14	$\mathbf{0.82 \pm 0.22}$	0.81 ± 0.26	$\mathbf{0.67 \pm 0.26}$	0.65 ± 0.28

Table 10: **Mapping all archaeocyathids to pitted texture: Performance on edit image.** This table contains the mean instance-level metrics $\pm$ one standard deviation run on each image used for mapping archaeocyathids to the pitted texture. For example, the first row contains the mean instance-level precision, recall, and IoU across the identified archaeocyathids in image A alone. There does not appear to be a clear trend in any of the metrics. For example, image D produces an improvement across all metrics, while image F does not.

Image	Precision		Recall		IoU	
	Original	Edited	Original	Edited	Original	Edited
A	$\mathbf{0.82 \pm 0.19}$	0.80 ± 0.21	$\mathbf{0.78 \pm 0.24}$	0.60 ± 0.32	$\mathbf{0.66 \pm 0.25}$	0.44 ± 0.35
B	$\mathbf{0.53 \pm 0.28}$	0.51 ± 0.32	$\mathbf{0.78 \pm 0.24}$	0.64 ± 0.28	$\mathbf{0.66 \pm 0.27}$	0.51 ± 0.32
C	$\mathbf{0.83 \pm 0.20}$	$\mathbf{0.83 \pm 0.16}$	$\mathbf{0.77 \pm 0.24}$	0.64 ± 0.32	$\mathbf{0.64 \pm 0.30}$	0.44 ± 0.36
D	$\mathbf{0.79 \pm 0.23}$	0.80 ± 0.21	$\mathbf{0.78 \pm 0.28}$	0.77 ± 0.24	$\mathbf{0.64 \pm 0.30}$	0.53 ± 0.36
E	$\mathbf{0.85 \pm 0.15}$	$\mathbf{0.85 \pm 0.14}$	$\mathbf{0.70 \pm 0.26}$	0.60 ± 0.32	$\mathbf{0.55 \pm 0.32}$	0.46 ± 0.34
F	0.83 ± 0.16	$\mathbf{0.87 \pm 0.11}$	$\mathbf{0.82 \pm 0.22}$	0.74 ± 0.24	$\mathbf{0.67 \pm 0.26}$	0.64 ± 0.26

Table 11: **Mapping simultaneously: Performance on edit image.** This table contains the mean instance-level metrics $\pm$ one standard deviation run on each image used for simultaneously mapping non-archaeocyathids to red mud and archaeocyathids to the pitted texture. For example, the first row contains the mean instance-level precision, recall, and IoU across the identified archaeocyathids in image A alone. Only image F produces an increase in precision. None of the edited models have higher recall or IoU scores than the original model.