

1 SUPPLEMENT MATERIAL & APPENDIX

1.1 Attributes of Datasets

In choosing the experimental datasets, we primarily considered two factors: firstly, the recognition and value within the community, and secondly, the complexity and refinement of the images in the datasets. Figure 1 displays the distribution and quantity of attributes/traits in images from *Cryptopunks*, *BAYC*, and *Azuki* - three highly popular NFT projects. Consequently, these three datasets aptly represent various complexities of datasets and NFT projects, further substantiating the performance of our model in feature extraction and image restoration.

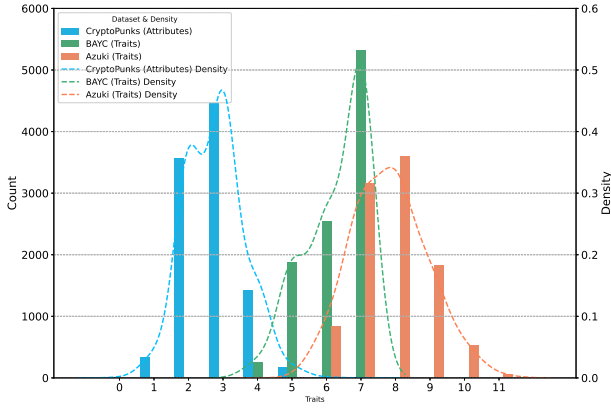


Figure 1: The Distribution of Attributes/Traits for different NFT datasets

1.2 Training Loss Result

By collecting the best MSE loss and the corresponding step numbers during training, we have generated the plot shown in Figure 2. Across all three datasets, there is a consistent trend of decreasing optimal loss as the embedding size increases. However, an exception occurs with the *BAYC* dataset at an embedding size of 70, where the EarlyStopping condition is triggered prematurely, resulting in a higher optimal loss. It's important to note that due to the use of a fixed random seed during training, repeated training does not lead to improved results. We consider this situation as a reproducible outlier in the search for gradients during training, but it does not significantly impact the overall trend. When comparing the three datasets, it is evident that datasets with more complex details require a longer time to converge, necessitating more training steps, and ultimately, they achieve higher optimal loss values.

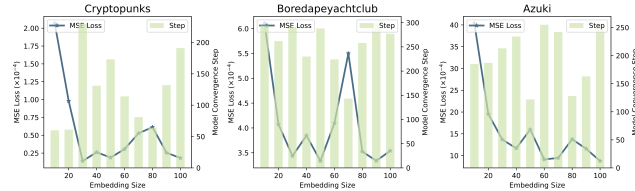


Figure 2: The MSE Loss and Model Convergence Step for training Cryptopunks, BAYC and Azuki.

1.3 Factor of Embedding Size

In Figure 3, we have selected one image from each dataset, along with their corresponding reconstructed images using float precision for embedding sizes of 10 and 100. Each subimage is labeled with its SSIM and PSNR information for both numerical and visual comparisons of the differences. The height of lines in the 3D plot means the difference of RGB pixels with original images.

From the images provided, we can observe that in the *Cryptopunks* dataset, there isn't a significant difference in the quality of two reconstructed images, which means size of 10 is enough for the series. However, in the *BAYC* dataset, the reconstructed images for embedding size 10 exhibit a significant difference around the monkey's eyes. This is due to the limited feature space of embedding size 10. In the *Azuki* dataset, the reconstructed images display higher differences in the character section than the background. In contrast, the pixel difference of reconstructed image in embedding size 100 is obviously lower than the image in embedding size 10.

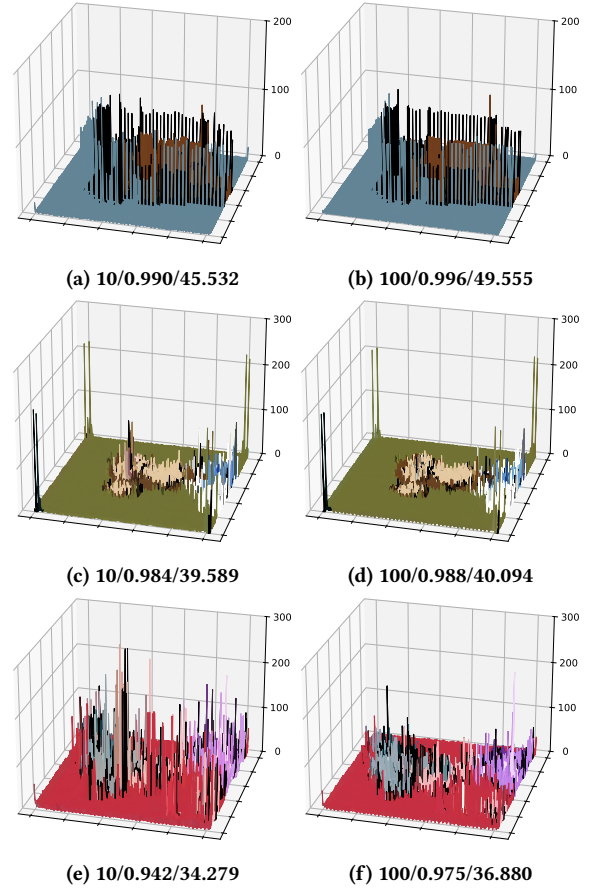


Figure 3: The comparison between original images and reconstructed images with embedding sizes of 10 and 100, along with their SSIM and PSNR values. The datasets, from top to bottom, represent Cryptopunks, BAYC, and Azuki.